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

CONSTRUCCIÓN Y VALIDACIÓN DE UNA METODOLOGÍA DE SEGUIMIENTO PARA MODELOS DE REGRESIÓN LOGÍSTICA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

GABRIELA ESMERALDA COVARRUBIAS MONDACA

PROFESOR GUÍA
JOSÉ MIGUEL CRUZ GONZÁLEZ.

MIEMBROS DE LA COMISIÓN
CRISTIÁN BRAVO ROMÁN.
SERGIO LEHMANN BERESI.

SANTIAGO DE CHILE
ABRIL 2012

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: GABRIELA E. COVARRUBIAS MONDACA
FECHA: ABRIL DE 2012
PROFESOR GUÍA: JOSÉ MIGUEL CRUZ G.

CONSTRUCCIÓN Y VALIDACIÓN DE UNA METODOLOGÍA DE SEGUIMIENTO PARA MODELOS DE REGRESIÓN LOGÍSTICA

Este trabajo de título tiene por objetivo general construir, implementar y validar una metodología de seguimiento no supervisada, para detectar cambios significativos en la distribución de las variables, en modelos de regresión logística.

El problema de seguimiento corresponde a detectar cambios en un modelo de minería de datos cuando éste es construido usando bases de datos no-estacionarias, es decir, conjuntos de información a los cuales constantemente se les están agregando nuevas observaciones. La consecuencia de estos cambios es que progresivamente el modelo perderá validez, por lo que deberá ser recalibrado en algún momento.

El enfoque desarrollado para las metodologías propuestas es que cada parámetro de un modelo de regresión logística asociado a la variable x_i del modelo posee un intervalo de confianza donde se presume que se encuentra su valor real. El supuesto es que si la población cambia de tal manera que el nuevo parámetro estimado está fuera de este intervalo, entonces el modelo no es válido para esa nueva muestra.

Se considera que el cuociente entre las medias es una buena medida del cambio entre dos muestras, solamente que no considera el efecto de la forma de la distribución. Es por ello que se corrige la media dividiéndola por la varianza muestral, obteniendo un coeficiente llamado ICV. Se plantean dos modelos a contrastar con el intervalo de cambio máximo: ICV-1 que corresponde al módulo de la diferencia del ICV de cada muestra e ICV-2 que corresponde al cuociente de dichos valores.

Se construyó un modelo de regresión logística utilizando una base datos de comportamiento crediticio, cuyo error de predicción total fue de 23,3%. Con los parámetros de este modelo se construyeron los intervalos de cambio máximo para cada variable y se las perturbó de tres maneras distintas para proceder a la validación.

Al aplicar los modelos propuestos, junto con otras metodologías de seguimiento, se concluye que ICV-1 presenta problemas debido a la forma en que se ha definido el intervalo de cambio máximo e ICV-2 tiene un buen rendimiento, comparable con el de Stability Index y la Distancia de Hellinger y considerablemente mejor que el test K-S y el test Chi-cuadrado.

A mis padres, por estar siempre a mi lado

A mis amigos, por soportar todas mis mañas

A mi equipo, por todos los triunfos y fracasos que hemos vivido. Gracias a ustedes soy mejor persona.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

CAPÍTULO 1

INTRODUCCIÓN

Para las empresas en el escenario actual, es sumamente relevante realizar un continuo análisis de los datos que, con el paso del tiempo y producto del ejercicio de sus actividades, han ido recolectando y almacenando. Estos datos por sí solos no tienen mayor valor, pero pueden ser utilizados para identificar patrones o modelos no triviales, que posteriormente serán usados como apoyo en la toma de decisiones.

Existen metodologías para llevar a cabo la tarea antes mencionada, la más popular de ellas se conoce como proceso de descubrimiento de conocimiento en bases de datos, o proceso KDD por sus siglas en inglés, que comprende desde la obtención de los datos hasta la interpretación y evaluación del conocimiento extraído. Es una metodología estructurada cuya etapa central es la minería de datos. Ésta se define como el proceso iterativo de extraer patrones ocultos de información que se encuentran en las bases de datos, utilizando para ello técnicas estadísticas y de inteligencia artificial (Norving, Russel, 2010).

Las herramientas de la minería de datos pueden ser clasificadas en cinco categorías, de acuerdo al objetivo final que se persigue, existiendo modelos de: agrupamiento, asociación, clasificación, regresión y secuenciación (Fayyad et al., 1996). La idea es ser capaces de identificar cuál de ellos se adapta a las características del problema que se esté modelando.

La regresión logística es una técnica estadística perteneciente al conjunto de herramientas de minería de datos, que se utiliza para explicar y predecir el comportamiento de una variable dicotómica, en función de un set de otras variables observables. Este tipo de modelamiento es comúnmente empleado en muchas áreas, incluyendo: investigación biomédica, finanzas, criminología, ecología, ingeniería, entre otras disciplinas (Hosmer, Lemeshow, 2000).

Este tipo de modelos son construidos y calibrados en base a un conjunto de datos recolectados en un período de tiempo determinado, producto de la observación del fenómeno en estudio. Al utilizar una base de datos fija se asume implícitamente que el comportamiento de las variables futuras será íntegramente representado por el comportamiento de las variables con las que se realizó el modelo. En otras palabras, se asume que las variables que conforman el modelo se relacionan entre sí con las mismas leyes y sus distribuciones de probabilidad se mantendrán inalteradas, con respecto a las variables originales del modelo.

El supuesto anterior es poco realista, pues los cambios en el entorno y la constante generación de nuevos datos pueden afectar el comportamiento de las variables, llegando incluso a invalidar un modelo. Por ello es necesario contar con herramientas estadísticas que permitan detectar estos cambios y determinar el mejor momento en que el modelo debe ser recalibrado. Para estudiar estos cambios existe una rama de la

estadística, econometría y minería de datos que se conoce como “Detección del Cambio” (Bravo et al., 2009).

Dentro de las metodologías de seguimiento para la detección del cambio en la distribución de las variables existen dos tipos: los modelos supervisados y los no supervisados. Los primeros requieren conocer la salida real (valor de la función) para detectar cambios, mientras que los segundos no requieren conocer este dato. En muchas aplicaciones es deseable poder detectar cambios en los modelos antes de obtener las salidas reales, es decir, no tener que esperar a que el modelo deje de ser válido para saber que es necesario recalibrarlo. Por ello, las metodologías de seguimiento no supervisadas toman real importancia en este tipo de industrias.

Un caso particular son los modelos de riesgo de crédito, que en su mayoría son desarrollados como una regresión logística. Estos modelos no escapan al fenómeno anteriormente descrito, dado que la información que se utiliza para su construcción se basa principalmente en características socio-económicas del solicitante y de patrones sobre la evolución de la deuda. A medida que el entorno socio-económico del país evoluciona, el comportamiento de las personas también cambia. Una experiencia empírica demuestra que aproximadamente cada dos años se observan cambios significativos en los hábitos de la población, suficientes para impactar en la capacidad predictiva del modelo (Thomas, 2000).

La principal consecuencia del fenómeno descrito en el párrafo anterior es que la capacidad discriminante de los modelos de riesgo de crédito sufrirá disminuciones progresivas, quedando eventualmente inválidos. Esto significa que habrá cada vez más clientes riesgosos que el modelo clasificará como buenos clientes y viceversa, lo que finalmente se traducirá en pérdidas para la entidad financiera. En este sentido, contar con herramientas sencillas que permitan prevenir este problema es muy necesario.

En este contexto, este trabajo se centra en el desarrollo, implementación y validación de una metodología de seguimiento no supervisada para detectar cambios en la distribución de las variables para el caso particular de modelos de regresión logística (modelo de clasificación).

Para llevar a cabo este objetivo se utilizará una base de datos de comportamiento crediticio, que permitirá estimar un modelo de clasificación en base a regresión logística, al cual será posible aplicar la metodología propuesta y contrastarla con aquellas que ya existen para este fin (benchmarking).

La metodología de seguimiento obtenida, si bien será validada para un caso particular, podrá ser potencialmente aplicable a cualquier modelo construido en base a regresión logística, que trabaje con bases de datos no estacionarias. Los modelos de credit scoring son el foco en este trabajo, pero perfectamente podrá ser aplicable a otros fenómenos similares.

1.1 ALCANCES

El objetivo principal de este trabajo es validar metodología de seguimiento en base a regresión logística. En este contexto, y al momento de aplicar las técnicas de minería de datos del proceso KDD, solo se utilizará la regresión logística como herramienta para generar el modelo de riesgo de crédito. Las otras técnicas de minería de datos que existen y que podrían utilizarse para construir un modelo de credit scoring, serán obviadas, pues el objetivo no es encontrar el mejor modelo que se ajuste a los datos disponibles, sino que plantear un modelo adecuado para la metodología en estudio.

Cabe desatacar que la base de datos disponible corresponde a créditos de consumo, por lo que cualquier extensión a otros tipos de créditos no se abordará en este trabajo.
La metodología de seguimiento obtenida, si bien se restringe a un caso particular, podrá ser potencialmente aplicable a cualquier modelo construido en base a regresión logística que trabaje con bases de datos no estacionarias. Los modelos de credit scoring son el foco en este trabajo, pero perfectamente puede ser aplicable a otros fenómenos similares.

CAPÍTULO 2

MARCO CONCEPTUAL

2.1 EL PROCESO KDD

La metodología KDD es el proceso no trivial de identificar patrones en los datos que sean válidos, potencialmente útiles y comprensibles (Fayyad et al., 1996). Es una metodología que estipula los pasos críticos para cumplir dicho objetivo, desde la selección de los datos que se van a utilizar hasta la interpretación de los modelos obtenidos. Se puede dividir en las siguientes etapas:

I. **Selección de Datos**: Consiste en seleccionar los datos desde las distintas fuentes de las que se dispone. Existen tres tipos de fuentes: internas, que corresponden a bases de datos propias de la entidad producto del ejercicio de sus actividades; externas, que corresponden a datos del entorno o datos que provienen de otras empresas; y variables generadas, que corresponden a indicadores definidos por la entidad en función de los datos disponibles de las dos fuentes anteriores. Esta etapa es fundamental, pues para que el modelo sea representativo del fenómeno en estudio se debe garantizar que la información que se utiliza para construirlo sea de calidad.

II. **Pre- procesamiento de datos*:*** Usualmente los datos contienen mucho ruido e inconsistencias que deben ser tratadas para eliminar cualquier efecto nocivo sobre el modelo final. En esta etapa se analizan las variables con las que se cuenta a fin de determinar cuáles de ellas son útiles para explicar el fenómeno en estudio. Aquí se hace el tratamiento de los datos nulos (se completan mediante algún método o se elimina el registro), el tratamiento de los valores fuera de rango y la eliminación de variables que no son relevantes para explicar el fenómeno.

III. **Transformación de Datos**: En esta etapa se realizan todas las transformaciones a las variables para que se cumplan los requerimientos que el modelo escogido exige. Los métodos más comunes son: reducción de la dimensión del problema, normalización de variables, creación de variables binarias para representar variables categóricas y discretización de variables continuas.

IV. **Minería de Datos**: Esta etapa consiste en ajustar un modelo estadístico al conjunto de variables que se analizan, generalmente estimando los parámetros que compondrán dicho modelo. El proceso comienza con la selección de atributos, dejando fuera aquellas variables que no discriminan bien desde un punto de vista estadístico. El paso siguiente es dividir la base de datos en dos subconjuntos; el primero de ellos se conoce como base de entrenamiento y se utiliza para calibrar el modelo y el segundo se conoce como base de testeo y se utiliza para validar el modelo y probar su capacidad.

En general, existe más de una técnica para abordar el fenómeno que se desea modelar y dependiendo del objetivo que se persigue se han clasificado los métodos de minería de datos en las siguientes categorías (Fayyad, 1996):

a. **Predicción**: El objetivo es predecir el valor de una variable específica en función de un conjunto de variables observables. Si la variable a predecir es categórica se habla de clasificación, si es continua, se habla de regresión.

b. **Segmentación**: El objetivo es agrupar observaciones similares mediante alguna medida de similitud en categorías heterogéneas entre sí.

c. **Asociación**: El objetivo es identificar relaciones entre los valores de un set de variables. Existen relaciones horizontales, es decir, entre atributos de una misma observación y relaciones verticales, es decir, entre valores de un mismo atributo par diferentes observaciones.

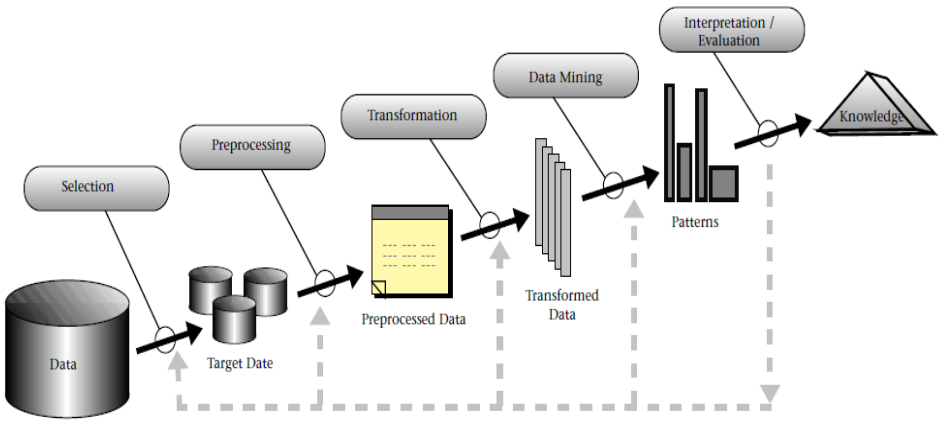
d. **Resumen de datos**: El objetivo es comprender de mejor manera la información disponible, por lo que este tipo de algoritmos presenta los datos de manera compacta, resaltando las relaciones básicas entre las variables.

e. **Detección de cambios y desviaciones**: El objetivo de estos métodos es predecir cambios en los patrones históricos de las variables, que usualmente se presentan de manera secuenciada, como las series de tiempo por ejemplo.

V. **Interpretación y evaluación de resultados**: Es la etapa final del proceso, que corresponde a analizar las salidas obtenidas, con el objetivo de determinar si el modelo utilizado logra explicar de manera satisfactoria el fenómeno en estudio.

A continuación se presenta un esquema del proceso KDD con cada una de las etapas que lo componen:

Ilustración 1: Proceso KDD



Fuente: "From Data Mining to Knowledge Discovery in Databases" Fayyad et al., 1996

2.2 EL MERCADO DE CRÉDITOS A CONSUMIDORES

Uno de los primeros indicios de una revolución financiera mediante la masificación de los préstamos a consumidores fue la fundación del Banco de Inglaterra en 1694. Por los siguientes 150 años la banca comenzó a entregar créditos a la nobleza y a la aristocracia. Cerca del año 1850, las empresas manufactureras comenzaron a vender sus productos en cuotas, lo cual significó un aumento aun más grande del mercado de créditos de consumo. Sin embargo, los préstamos sin garantía realmente comenzaron en los años 1920, cuando Henry Ford y A. P. Sloan se dieron cuenta que para vender autos de manera masiva debían contar con formas de financiamiento que permitieran a los consumidores costear su compra, y así desarrollaron las casas financieras.

La introducción al mercado de las tarjetas de crédito en las décadas del 50 y 60 ha permitido que los consumidores adquieran mediante crédito casi todas sus compras, desde comida rápida hasta pasajes de avión. Actualmente, los créditos de consumo y las tarjetas son la fuerza motriz detrás de las economías de los países líderes, sin estos instrumentos, el enorme crecimiento que ha experimentado el gasto de los consumidores en los últimos 50 años no sería posible (Thomas, 2009)

Un mercado de créditos para consumidores se compone, generalmente, de los siguientes instrumentos financieros:

- Préstamos con garantía que no sean primeras hipotecas.
- Tarjetas de crédito
- Préstamos
- Tarjetas de tiendas
- Compras a plazo o en cuotas
- Cooperativas que entregan créditos.

Estados Unidos es el país que tiene la mayor cantidad de dinero invertido en créditos de consumo. En el año 2007, el total de deuda acumulada por los consumidores fue de 13 trillones de dólares, de los cuales 9,8 trillones correspondían a hipotecas y 2,4 trillones correspondían a créditos de consumo (préstamos bancarios, tarjetas de crédito, préstamos de casas comerciales, etc.). (Thomas, 2009).

La magnitud de transacciones que se realizan día a día es una cifra que va en aumento y el mercado de crédito a los consumidores crece cada vez más. Prestar dinero es una actividad que conlleva riesgos, es por ello que las empresas financieras deben contar con mecanismos que les permitan, por un lado minimizar el riesgo asociado a los clientes y por otro aprovechar las oportunidades captando a aquellas personas que tengan un comportamiento más confiable.

Para ello existe un conjunto de herramientas de análisis estadístico que permite clasificar a los distintos clientes de acuerdo a ciertas características, y predecir su comportamiento a través del tiempo. Es importante contar con estos modelos, que ayudan a las empresas financieras a tomar mejores decisiones, en ese sentido, desarrollar metodologías que den soporten a dichos modelos y les den más

confiabilidad, como las metodologías de seguimiento para modelos de scoring, son siempre un aporte bien recibido.

2.3 MODELOS DE CREDIT SCORING

Credit Scoring se define como el set de modelos de decisión y sus técnicas subyacentes que ayuda a las instituciones financieras en la concesión de créditos a sus clientes. Estas técnicas permiten decidir a qué clientes se les entregará crédito, cuál será el monto asignado y qué estrategias operativas mejorarán la rentabilidad de los prestatarios (Thomas et al., 2002).

A la empresa financiera le interesa conocer la probabilidad de incumplimiento de la deuda asociada a cada solicitante, por lo tanto, los modelos de credit scoring se encargan de estimar dicha probabilidad en base a las características personales del individuo (edad, remuneraciones, trabajo, etc.) y las características del crédito que solicita (monto, plazo). Para calcular este valor se utiliza información del comportamiento de otros individuos con similares características, que ya han recibido un crédito, y que por ende se conoce su comportamiento.

El planteamiento teórico que sigue cualquier modelo de credit scoring se resume en los siguientes puntos (Altman et al., 1981):

I. Es un problema dinámico, es decir, incorpora varios periodos dentro de un horizonte temporal

II. El método consiste en estimar el valor presente descontado de los posibles beneficios o pérdidas derivados de la concesión de un crédito a un individuo en particular, en todos los periodos del horizonte temporal considerado.

III. Al individuo se le concede el crédito solicitado si la esperanza de este valor presente es positiva.

IV. La información inicial que se utiliza es el comportamiento crediticio de otros individuos, con características similares, a los que se les ha concedido un préstamo en el pasado.

Existen varios procedimientos que permiten estimar la probabilidad de incumplimiento para cada uno de los solicitantes, los más utilizados son los métodos estadísticos entre los cuales destacan la regresión lineal, la regresión logística, los árboles de clasificación, entre otros (Thomas et al., 2002).

La probabilidad de incumplimiento, si bien es una muy buena medida cuantitativa del riesgo de crédito de cada cliente, no es el indicador utilizado para decidir si un crédito se debe otorgar o no. Esto se realiza mediante las *scorecards* (Siddiqui, 2005). La importancia del buen diseño de scorecards dentro de una empresa financiera es

fundamental, pero su desarrollo no será abordado en este trabajo pues no aporta al objetivo fundamental de esta investigación.

2.4 REGRESIÓN LOGÍSTICA

Los modelos de regresión son herramientas estadísticas utilizadas para describir la relación entre una variable objetivo y un conjunto de variables explicativas. La regresión logística se utiliza cuando la variable que se desea modelar es dicotómica, es decir, del tipo Sí/No, Bueno/Malo, Presente/Ausente, etc. y busca modelar la influencia de la aparición de las variables explicativas en la ocurrencia del fenómeno dicotómico (Hosmer & Lemeshow, 2000). En la práctica, para aplicar este modelo se crea una variable binaria ficticia cuya estructura es:

$$y_i = \begin{cases} 1 & Cuando\ el\ fenómeno\ ocurre \\ 0 & Cuando\ el\ fenómeno\ no\ ocurre \end{cases}$$

Donde i representa cada observación que se posee.

Por lo tanto, la regresión logística es un modelo estadístico de clasificación binaria que entrega la probabilidad de pertenencia a uno de los dos grupos definidos, utilizando para ello un conjunto de regresores (variables) $x_i \in \mathbb{R}^n$ con $i = \{1 \dots N\}$ y *N* el número de observaciones.

La probabilidad de pertenencia se obtiene mediante:

$$p(x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1' x_i)}}$$

Donde $\beta_1 \in \mathbb{R}^n$

Para la estimación de los parámetros, es decir, la calibración del modelo se utiliza el método de máxima verosimilitud, en el cual se busca maximizar la probabilidad estimada de obtener los resultados categorizados según y_i (Hosmer & Lemeshow, 2000). La función de verosimilitud es la siguiente:

$$\mathcal{L}(\beta) = \prod_i f(x_i, \beta)$$

Donde $f(x_i, \beta)$ corresponde a la función de densidad de probabilidad de x_i que en este caso correspondería al modelo de regresión logística. En otras palabras, la función de verosimilitud puede ser expresada como:

$$\mathcal{L}(\beta) = \prod_i p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i}$$

Usualmente se trabaja con el logaritmo de la función de verosimilitud, pues es más simple de abordar matemáticamente, la expresión es la siguiente:

$$L(\beta) = \ln[\mathcal{L}(\beta)] = \sum_i \{y_i \ln[p(x_i)] + (1 - y_i)\ln[1 - p(x_i)]\}$$

Los estimadores de máxima verosimilitud se calculan al aplicar condiciones de primer orden a la función de verosimilitud. Las ecuaciones obtenidas, conocidas como ecuaciones de verosimilitud son:

$$\sum [y_i - p(x_i)] = 0$$

$$\sum x_i[y_i - p(x_i)] = 0$$

Al resolver este problema de optimización se obtiene como resultados un conjunto de estimadores asintóticamente eficientes, insesgados y distribuidos normalmente. $\hat{\beta}_0$ corresponde al punto de corte en el eje de las ordenadas y los demás estimadores $\hat{\beta}_i, i \neq 0$ corresponden a los coeficientes asociados a las variables explicativas. Es importante señalar que los parámetros de la regresión logística se distribuyen asintóticamente siguiendo una distribución normal de parámetros $(\beta_i, \sigma_{\beta_i})$. (Hosmer & Lemeshow, 2000)

En la práctica, el uso que se le da a este modelo consiste en seleccionar un punto de corte para el valor de la probabilidad, tal que para valores mayores a ese punto de corte se determine que el valor esperado para la variable en estudio sea 1 y en caso contrario se asigna valor 0. De esta manera se logra la clasificación. En términos matemáticos:

$$p(x_i) \geq punto\ de\ corte => \ y_i = 1$$
$$p(x_i) < punto\ de\ corte => \ y_i = 0$$

2.5 METODOLOGÍAS DE SEGUIMIENTO

Cuando se trabaja en un proyecto de minería de datos, la construcción y validación del modelo se hace a partir de un conjunto de datos, que corresponden a aquellos que estaban disponibles al momento de modelar el fenómeno en estudio. Este modelo, si es correctamente validado comienza a ser aplicado para predecir el comportamiento futuro

de la variable de interés, asumiendo para ello que las condiciones en las que el modelo fue construido y las condiciones en las que será aplicado son exactamente las mismas (Cieslak et al., 2007). Este supuesto es bastante débil y rara vez se cumple, dado que las condiciones en las que un fenómeno está inmerso tienden a cambiar constantemente.

Entonces, el objetivo de realizar seguimiento es detectar estos posibles cambios en las variables que afectan la capacidad discriminante, para así re-calibrar el modelo cuando éste deje de ser efectivo, asegurando su buen funcionamiento en todo momento.

Existen tres posibles cambios que pueden ocurrir en el modelo, que inducen un cambio en los parámetros de las variables (Zeira et al., 2005), enunciados a continuación:

I. Distribución de las variables

II. Capacidad discriminante de las variables

III. Capacidad discriminante del modelo en su conjunto

A continuación se describe el planteamiento teórico del problema de seguimiento de modelos (Zeira et al., 2005). Sea D una base de datos compuesta por X registros, que cuenta con N atributos (variables explicativas) y una variable objetivo Y. Sea además M un modelo de minería de datos que describe la relación entre N e Y.

En la mayoría de los algoritmos D se divide en dos partes, un subconjunto de entrenamiento para calibrar el modelo de manera que sea estadísticamente significativo y un subconjunto de validación, el cual permite evaluar el modelo utilizando observaciones que hasta ahora son desconocidas para él. La idea en este último caso es evaluar que tan bien se ajusta el modelo a este nuevo conjunto de datos.
Cuando la base de datos D es no-estacionaria, producto de un fenómeno que evoluciona en el tiempo, en cada periodo k un nuevo set de observaciones d_k es agregado a la muestra, por lo tanto la base de datos acumulada hasta el período K es:

$$D_K = \bigcup_{k=1}^{K} d_k$$

Si M_K es el modelo para el período k, lo que interesa averiguar es si se cumple:

$$M_k = M_{k+1} \qquad \forall K = 1 \dots \infty$$

En otras palabras, lo que interesa saber es si el modelo que representa el fenómeno en el momento k sigue siendo válido en el momento $k+1$, o no. Esto se traduce en las siguientes interrogantes:

I. ¿Cambió significativamente en modelo durante en nuevo período?

II. ¿Cuál fue la naturaleza de ese cambio?

III. ¿Los períodos pasados son redundantes para estimar un nuevo modelo?

La forma de determinar si estos cambios son significativos es mediante un test de hipótesis, donde se define la hipótesis nula como la no-ocurrencia de cambios significativos en el modelo. El objetivo es entonces construir un estadístico que permita rechazar la hipótesis nula cuando esta no es cierta.

Las metodologías de seguimiento propuestas en este trabajo buscan determinar cambios significativos en la distribución de las variables sin la necesidad de conocer las salidas reales del fenómeno en estudio (modelo no supervisado). Se entiende por cambio significativo aquella variación en la distribución de una variable que compone un modelo predictivo, que impacta en el funcionamiento del modelo, afectando la capacidad con la que éste discrimina entre los distintos casos. Es natural que las variables que describen un fenómeno en estudio vayan cambiando su comportamiento conforme pasa el tiempo, los pequeños cambios en la distribución de las variables son bastante probables y en general no impactan de manera significativa el modelo, por lo que identificarlos no es el interés de este trabajo. Sin embargo hay otro tipo de cambios, que son menos probables, y que usualmente son provocados por cambios estructurales del proceso en estudio o por la suma de los pequeños cambios no detectados que llegado cierto punto comienzan a ser significativos (con respecto a la variable original) que impactan de manera negativa la capacidad del modelo.

Para testear el rendimiento de las metodologías propuestas se escogieron cinco trabajos de similares características que permitirán comparar la eficacia cada metodología, los cuales son presentados a continuación.

2.5.1 Test Beta 1

Este enfoque fue propuesto en (Bravo et al., 2009) y corresponde a un test empírico no supervisado capaz de identificar cambios en la distribución de las variables. Este test permite capturar la variación máxima que puede experimentar un parámetro en función de la media empírica de las variables y del intervalo de confianza calculado para cada parámetro. Formalmente:

Sea $[\beta_i^{inf}, \beta_i^{sup}]$ el intervalo de confianza para el parámetro β_i, con i = 1,…,n (β_i es el parámetro asociado a la variable x_i del modelo de regresión) donde se cree que está su valor poblacional. El supuesto que se realiza es que si la población cambia de tal manera que el nuevo parámetro estimado está fuera de este intervalo de confianza, entonces el modelo no es válido para esa nueva muestra.

La variación máxima permitida para el parámetro β_i corresponde a:

$$C_j \in \left[\frac{\beta_i^{inf}}{\widehat{\beta}_\iota}, \frac{\beta_i^{sup}}{\widehat{\beta}_\iota}\right]$$

Entonces una posible medida empírica del cambio máximo de la variable puede expresarse en términos de la media de las variables nuevas observadas. Si la media inicial de la variable x_i es $\overline{x_\iota}$ y la media de la misma variable en la nueva muestra es $\overline{x_\iota'}$, entonces:

$$\frac{\overline{x_\iota}}{\overline{x_\iota'}} \in \left[\frac{\beta_i^{inf}}{\widehat{\beta}_\iota}, \frac{\beta_i^{sup}}{\widehat{\beta}_\iota}\right]$$

La expresión anterior determina un umbral para determinar si la variable aun no cambia lo suficiente. Si el cociente entre las medias cae fuera del intervalo, significa que la distribución que genera la variable ha experimentado un cambio significativo, en caso contrario se considera que, de haber cambio, este no es significativo como para impactar de manera negativa el modelo.

2.5.2 Stability Index

Stability Index es un indicador de la similitud entre la muestra observada en el periodo t, y la muestra utilizada para construir el modelo (Baesens et al., 2010). Se calcula como:

$$SI = \sum_{i=1}^{m} (E_i - P_i) ln \frac{E_i}{P_i}$$

Donde m es el número de clases consideradas, E_i representa el porcentaje de observaciones de la clase m en el conjunto de entrenamiento del modelo (muestra original) y P_i corresponde al porcentaje de observaciones de la clase m del conjunto de prueba del modelo (nueva muestra).

Los criterios para determinar si existe un cambio significativo en la distribución de las variables son:

I. $SI < 0{,}1$ Ha ocurrido un cambio no significativo

II. $0{,}1 < SI < 0{,}25$ Ha ocurrido un cambio menor

III. $SI > 0{,}25$ Ha ocurrido un cambio significativo

La idea es que cuando se identifique una divergencia significativa entre las distribuciones de ciertas variables, se proceda a estudiarlas con mayor detención con el fin de determinar estadísticamente esta diferencia con contraste de hipótesis.

2.5.3 Test Kolmogorov – Smirnov

El test Kolmogorov-Smirnov (KS) para variables continuas, es un test de hipótesis que se utiliza para determinar si hay divergencia entre las distribuciones de probabilidad de dos muestras independientes y para contrastar la distribución empírica de una muestra contra una distribución teórica (Cieslak et al., 2007).

En este caso interesa comparar la distribución de una variable entre dos muestras: aquella con la que se calibró el modelo y una nueva muestra de datos, con el objetivo de determinar si la variable en estudio ha sufrido cambios significativos en su distribución, que pudiesen afectar de manera negativa un modelo construido en base a la primera muestra de datos. Por lo tanto, las hipótesis del test pueden ser planteadas de la siguiente forma:

H_0: La distribución de la variable entre las muestras no es divergente
H_1: La distribución de la variable entre las muestras es divergente

El contraste de este test se base en las diferencias entre las frecuencias relativas acumuladas para los mismos puntos de corte en cada muestra.

$$D_i = F_1(x_i) - F_2(x_i)$$

Donde $F_1(x_i)$ es la frecuencia acumulada de la variable x_i, para un punto de corte determinado, en la muestra original y $F_2(x_i)$ es la frecuencia acumulada de la variable x_i, para el mismo punto de corte, en la nueva muestra.

Con estas diferencias, y para varios puntos de corte, se construye el siguiente estadístico cuya distribución es conocida y puede ser consultada en una tabla.

$$Z_{KS} = max_i\{|D_i|\}$$

2.5.4 Distancia de Hellinger

La distancia de Hellinger es un indicador de la divergencia entre la distribución de dos variables categóricas (Cieslak et al., 2007). Sean X e Y dos poblaciones con p categorías, la distancia de Hellinger entre ellas se calcula como:

$$Hellinger\ (X,Y) = \sqrt{\sum_{j=1}^{p}\left(\sqrt{\frac{X_j}{n_X}} - \sqrt{\frac{Y_j}{n_Y}}\right)^2}$$

Donde X_j corresponde al número de observaciones de la clase j en va variable X, Y_j corresponde al número de observaciones de la clase j en va variable Y, n_X y n_Y son el número total de observaciones de la variable X e Y, respectivamente.

Este indicador alcanza un valor mínimo 0, lo que significa que las distribuciones son idénticas y un valor máximo de $\sqrt{2}$, en cuyo caso las distribuciones son totalmente divergentes.

2.5.5 Test χ^2

El test χ^2 puede ser utilizado para determinar si una distribución empírica de una variable coincide con una distribución de probabilidad conocida para esa variable (Zeira et al., 2005)**.** Si el conjunto de entrenamiento es lo suficientemente grande, entonces es correcto asumir que la verdadera distribución de la variable será bien estimada por éste.

Por lo tanto, este test permitirá comparar la distribución de la variable en el conjunto de prueba (distribución empírica) con la del conjunto de entrenamiento (verdadera distribución de la variable).

El test de hipótesis que se plantea es el siguiente:

H_0: La variable X es estacionaria entre períodos
H_1: La variable X no es estacionaria entre períodos

El estadístico de prueba se calcula como:

$$\chi_p^2 = n_K \sum_{i=1}^{j} \frac{\left(\frac{x_{i,K}}{n_K} - \frac{x_{i,K-1}}{n_{K-1}}\right)^2}{\frac{x_{i,K-1}}{n_{K-1}}}$$

Este estadístico sigue una distribución χ^2 de $j-1$ grados de libertad, donde j es la cantidad de categorías para cada variable en estudio. Por lo tanto, si el valor del estadístico χ_{j-1}^2 es mayor que el valor crítico de la distribución χ^2, entonces se rechazará la hipótesis nula, es decir, la variable no es estacionaria entre períodos y si presenta cambios en su distribución.

CAPÍTULO 3

PLANTEAMIENTO DEL MODELO

3.1 PROBLEMA A RESOLVER

El estudio y desarrollo de nuevas metodologías de seguimiento que sean más robustas desde el punto de vista estadístico y que respondan de manera eficiente al problema de detección de cambios, es un aspecto de la estadística y la econometría que está en constante investigación (Hosmer, Lemeshow, 2000).

Existen tres posibles cambios que pueden ocurrir en el modelo que inducen un cambio en los parámetros de las variables: el cambio en la distribución de las variables, el cambio en la capacidad discriminante de las variables y el cambio en la capacidad discriminante del modelo en su conjunto. La metodología desarrollada se encarga del primero de ellos.

Por lo tanto, el problema a resolver puede enunciarse de la siguiente manera. Se tiene una base de datos con i atributos, cada variable x_i tiene una distribución de probabilidad conocida que depende de la muestra de datos en cuestión. Con esta base de datos se ha calibrado un modelo de regresión logística, obteniendo los coeficientes $\hat{\beta}_i$ asociados a cada variable.

Se tiene además una nueva muestra de datos de la misma variable con su propia distribución de probabilidad. Lo que interesa saber es si ambas distribuciones son lo suficientemente parecidas como para que el modelo calibrado con la primera base de datos siga siendo válido para la segunda. En otras palabras, que el cambio C_i experimentado por la variable x_i de una muestra a otra no es lo suficientemente significativo como para afectar el rendimiento del modelo en la nueva muestra.

Por lo tanto es necesario definir el cambio máximo permitido en la distribución de una variable y una forma de medir dicho cambio en función de parámetros de la muestra conocidos o calculables.

3.2 DEFINICIÓN DE CAMBIO MÁXIMO

El primer paso es definir qué significa un cambio significativo tal que el modelo calibrado con las variables originales deja de ser válido para la nueva muestra de datos, para ello se utiliza el coeficiente β_i del modelo de regresión logística, que se distribuye asintóticamente siguiendo una distribución normal de parámetros $(\hat{\beta}_i, \sigma_{\beta_i})$, lo que permite aprovechar todas las propiedades de dicha distribución. Cada parámetro β_i asociado a la variable x_i del modelo posee un intervalo de confianza donde se presume

que se encuentra el valor real de dicho parámetro. Este intervalo puede ser escrito como:

$$\beta_i \in \left[\beta_i^{inf}, \beta_i^{sup}\right]$$

El supuesto es que si la muestra ha experimentado un cambio de una magnitud tal que el nuevo parámetro estimado está fuera de este intervalo de confianza, entonces el modelo no será válido para esta muestra. Entonces, la variación máxima permitida será:

$$C_i \in \left[\frac{\beta_i^{inf}}{\hat{\beta}_i}, \frac{\beta_i^{sup}}{\hat{\beta}_i}\right]$$

Dado que la distribución que siguen los parámetros β_i es normal, el intervalo donde se encuentra el valor real, al 95% de confianza, puede reescribirse como:

$$\beta_i \in \left[\hat{\beta}_i - 2\sigma_{\beta_i}, \hat{\beta}_i + 2\sigma_{\beta_i}\right]$$

Y por lo tanto la variación máxima permitida para el parámetro corresponde a:

$$C_i \in \left[\frac{\hat{\beta}_i - 2\sigma_{\beta_i}}{\hat{\beta}_i}, \frac{\hat{\beta}_i + 2\sigma_{\beta_i}}{\hat{\beta}_i}\right]$$

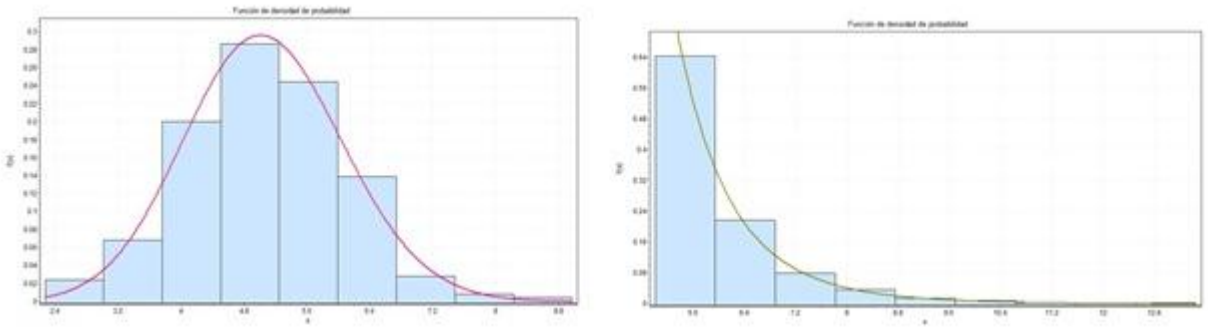
3.3 DEFINICIÓN DE UNA MEDIDA DE CAMBIO

El siguiente paso es encontrar una manera de medir el cambio que experimenta la distribución de las variables de una muestra a otra. Para ello es necesario comparar algún indicador representativo de cada una de ellas. El test Beta 1, que presenta una metodología similar a la de este trabajo, propone como medida de cambio la razón entre las medias, asumiendo que si dicha razón está en el intervalo de cambio permitido, entonces la distribución de las variables en cada muestra no ha variado significativamente como para tener que recalibrar el modelo.

El problema de medir el cambio entre la distribución de las variables de esa manera es que se está ignorando la “forma” que tiene la distribución. Sean dos muestras de variables aleatorias con una media similar, pero una desviación estándar muy distinta, como se muestra en la figura 2, si se utiliza solo la razón entre las medias como indicador del cambio experimentado, se llegaría a que las muestras son muy similares

ya que las medias lo son, pero claramente las muestras son muy distintas (por inspección se puede ver). Es por ello que se propone corregir la media de cada muestra dividiéndola por la desviación estándar muestral de los datos, de esta forma se introduce el efecto de la forma de la distribución.

Ilustración 2: Distribuciones con la misma media y distinta forma



Por otra parte, la razón entre la media y la varianza es muy similar al inverso de un indicador estadístico conocido como coeficiente de variación, que sirve para medir el grado de dispersión de los datos en torno a la media, lo que implica que su inverso indicaría el grado de proximidad de los datos a la media, lo cual permite tener una idea intuitiva de la forma de la distribución. Se define entonces la razón entre la media y la varianza de la siguiente forma:

$$ICV_i = \frac{\bar{x}_i}{S_i}$$

Por lo tanto, una muestra muy dispersa, cuya varianza es muy grande tendrá un ICV pequeño, mientras que una muestra con una varianza pequeña tendrá un ICV grande.

Notar que tanto la media como la varianza muestral son variables aleatorias, por lo que es necesario conocer su distribución, para entender como se comporta ICV. Sean x_1, x_2,…, x_n valores muestrales independientes e idénticamente distribuidos (i.i.d.) de una variable aleatoria x. Se define la media muestral como:

$$\bar{x}_n = \sum_{i=1}^{n} \frac{x_i}{n}$$

Si la distribución de población tiene como esperanza μ y varianza σ^2 entonces, la esperanza y la varianza de la media muestral corresponden a:

$$E(\bar{x}_n) = E\left(\sum_{i=1}^{n} \frac{x_i}{n}\right) = \mu$$

$$Var(\bar{x}_n) = Var\left(\sum_{i=1}^{n} \frac{x_i}{n}\right) = \frac{\sigma^2}{n}$$

Por otro lado, el teorema central del límite, indica que en condiciones muy generales (y cuando la desviación estándar es conocida), la suma de n variables aleatorias independientes se aproxima bien a una distribución normal cuando n es grande, y por lo tanto es posible asegurar que la distribución de la media muestral (dado que se trabaja con una gran cantidad de datos) es normal, de parámetros:

$$\bar{x}_n \sim N(\mu, \frac{\sigma^2}{n})$$

La varianza muestra se define de la siguiente manera:

$$S_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Cuya esperanza y varianza se muestran a continuación:

$$E(S_n^2) = \sigma$$

$$Var(S_n^2) = \frac{\mu_4 - \sigma^4}{n} \to 0$$

Donde μ_4 es el momento teórico de orden 4 de la variable aleatoria. Notar que la varianza de la varianza muestral tenderá a 0 cuando n sea lo suficientemente grande.

Por lo tanto, la distribución de la varianza muestral no es única y depende de la distribución de la variable de origen, por ejemplo, si las variables de origen son normales entonces la distribución de la varianza muestral es una chi cuadrado, pero no se puede asegurar nada cuando se habla de una distribución de las variables cualquiera.

Si se definen nuevas variables aleatorias iguales a $(x_i - \bar{x})^2$ entonces se podría asumir que la suma de dichas variables sigue aproximadamente una distribución normal, cuyos parámetros serían la esperanza y la varianza de la varianza muestral. Este supuesto es un tanto débil ya que para que se cumpla dicha propiedad es necesario que las variables sean i.i.d. lo cual difícilmente se cumple en este caso, pero sirve de aproximación.

Por lo tanto, la distribución de ICV (con todas las restricciones anteriores) corresponde al cuociente entre dos distribuciones normales, lo que se conoce como distribución de Cauchy[1], cuya función de densidad de probabilidad es la siguiente:

[1] Wikipedia, The Free Enciclopedia. Cauchy Distribution.

$$p_Z(z) = \frac{b(z)c(z)}{a^3(z)} \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y}\left[2\Phi(\frac{b(z)}{a(z)} - 1\right] + \frac{1}{a^2(z)\pi\sigma_x\sigma_y} e^{-\frac{1}{2}(\frac{\mu_x^2}{\sigma_x^2}+\frac{\mu_y^2}{\sigma_y^2})}$$

Donde

$$a(z) = \sqrt{\frac{1}{\sigma_x^2}z^2 + \frac{1}{\sigma_y^2}}$$

$$b(z) = \frac{\mu_x}{\sigma_x^2}z + \frac{\mu_y}{\sigma_y^2}$$

$$c(z) = e^{\frac{1b^2(z)}{2a^2(z)} - \frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2}+\frac{\mu_y^2}{\sigma_y^2}\right)}$$

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

3.4 CARACTERÍSTICAS DE LOS MODELOS PROPUESTOS.

3.4.1 Test ICV-1

Una primera forma de medir el cambio en la distribución de dos muestras es mediante la resta de sus ICV, luego de determinar esta diferencia, se verifica contra el intervalo de cambio máximo determinado por los coeficientes beta de la regresión y se concluye en base a ello. Entonces, el test corresponde a:

$$ICV_1 - ICV_2 = \left|\frac{\bar{x}_1}{S_1} - \frac{\bar{x}_2}{S_2}\right| \in \left[\frac{\beta_i - 2\sigma_{\beta_i}}{\beta_i}, \frac{\beta_i + 2\sigma_{\beta_i}}{\beta_i}\right]$$

Notar que cuando las muestras son exactamente las mismas, es decir, tienen la misma media y varianza muestral, entonces el valor del estadístico es igual a 0, es por ello que mientras más cercana a 0 sea la diferencia entre los ICV de las muestras, más similares serán sus distribuciones.

Dado que ICV (para el caso de variables normales) tiene una distribución de Cauchy entonces por propiedades de dicha distribución, la resta también será Cauchy. El problema de determinar la distribución de este estadístico no será abordado en este trabajo ya que la complejidad del cálculo matemático excede la capacidad de esta

memoria y por tanto quedan propuestas como trabajo futuro. Sin embargo, cabe destacar que existe una publicación titulada “Asymptotic sampling distribution of inverse coefficient-of-variation and its applications” de K. Sharma, que podría servir para profundizar en este tema, la cual no se encuentra disponible (de manera gratuita o mediante el vínculo institucional de la Universidad de Chile) por lo que no ha sido consultada.

3.4.2 Test ICV-2

Una segunda forma de medir el cambio en la distribución de dos muestras es mediante la razón de sus ICV, luego de determinar este valor, se verifica contra el intervalo de cambio máximo determinado por los coeficientes beta de la regresión y se concluye en base a ello. Entonces, el test corresponde a:

$$\frac{ICV_1}{ICV_2} = \left. \frac{\bar{x}_1}{S_1} \middle/ \frac{\bar{x}_2}{S_2} \right. \in \left[\frac{\beta_i - 2\sigma_{\beta_i}}{\beta_i}, \frac{\beta_i + 2\sigma_{\beta_i}}{\beta_i} \right]$$

Notar que cuando las muestras son exactamente las mismas, es decir, tienen la misma media y varianza muestral, entonces el valor del estadístico es igual a 1, es por ello que mientras más cercana a 1 sea la razón entre los ICV de las muestras, más similares serán sus distribuciones.

Encontrar la distribución de este estadístico también es un problema en si mismo, pero se sabe que si dos variables independientes X e Y siguen una distribución de Cauchy de media 0 y factor de forma igual a a y b respectivamente, entonces la distribución del cuociente de dichas variables es[2]:

$$p_z(z|a,b) = \frac{ab}{\pi^2(b^2z^2 - a^2)} \ln\left(\frac{b^2z^2}{a^2}\right)$$

[2] Wikipedia: The Free Enciclopedia. Ratio Distribution

CAPÍTULO 4

MODELO DE REGRESIÓN LOGÍSTICA

4.1 DESCRIPCIÓN GENERAL DE LA BASE DE DATOS

La base de datos utilizada para el desarrollo del modelo de regresión logística fue extraída del sitio web "*Kaggle*"[3], una plataforma de competencia de modelos predictivos, donde las empresas ponen a disposición bases de datos y estadísticos de todo el mundo compiten para crear los mejores modelos.

Los datos escogidos corresponden a los de una institución financiera, que cuenta con 150000 registros. La variable objetivo es de carácter binario, donde el valor 1 corresponde a aquellos clientes que presentarán 90 o más días de mora y el valor 0 corresponde a aquellos clientes que no lo harán. Hay un 93,3% de clientes buenos y un 6,6% de clientes malos.

Se cuenta con 10 variables explicativas continuas, donde la principal información que se entrega es a cerca del historial crediticio con la institución financiera, pero también hay variables demográficas y de características de los clientes (como la edad y el número de dependientes). En la tabla a continuación se describe cada una de las variables:

Tabla 1: Descripción de las Variables del Modelo

Variable	Descripción	Tipo
Target	Persona con 90 o más días de mora	Dicotómica
Uso_línea	Saldo total de las tarjetas y líneas de crédito personales, excepto bienes raíces y la deuda a plazo; dividido por la suma de los límites de crédito.	Porcentaje
Edad	Edad en años	Entero
Mora3060	Número de veces que el prestatario ha tenido entre 30-59 días de retraso (pero no peor que eso) en los últimos dos años	Entero
DebtRatio	Los pagos mensuales de la deuda, pensiones alimenticias, los costos de vida, dividido por los ingresos brutos mensuales	Porcentaje
Ingreso	Ingreso mensual	Real
OtrosProd	Número de créditos abiertos y líneas de crédito	Entero
Mora90	Número de veces que el prestatario ha tenido 90 días o más de deuda	Entero
NumHip	Número de hipotecas y préstamos de bienes raíces, incluidas las líneas de crédito hipotecario	Entero
Mora6090	Número de veces que el prestatario ha tenido entre 60-89 días de retraso (pero no peor que eso) en los últimos dos años.	Entero
Dependientes	Número de dependientes en la familia (cónyuge, hijos, etc.) excluyéndose a si mismo	Entero

[3] www.kaggle.com

4.2 TRATAMIENTO DE LOS DATOS

Existen dos tipos de errores en los datos que pueden causar problemas al momento de generar un modelo: los valores faltantes y los valores fuera de rango. Se entienden por valores faltantes a la ausencia del valor de cierta variable que describe al cliente y por valores fuera de rango a valores que escapan de los rangos permitidos para una variable en particular. Lidiar con este tipo de datos es el objetivo de la etapa de pre-procesamiento y existen diversas técnicas para ello, como por ejemplo: eliminación de registros, llenado con promedios y modas, utilización de modelos predictivos, entre otras.

4.2.1 Valores fuera de Rango

La mayoría de las variables presenta valores fuera de rango, algunas en una proporción muy pequeña de registros y otras en una mayor cantidad. El tratamiento que se les dio fue el siguiente:

- Se eliminaron los registros que tuviesen valores fuera de rango en las variables: Mora3060, Mora6090, Mora90 y OtrosProd. Estas variables tenían una proporción muy pequeña de valores fuera de rango, por lo que eliminar los registros no significaba una modificación muy sustancial de la base de datos.

- La variable Edad presentaba 6,75% de valores fuera de rango, donde el rango considerado válido fue entre 21 y 75 años. Los valores que estaban fuera de dicho rango podrían haber sido remplazados con el promedio, pero esto distorsionaba la distribución de la variable por lo que se consideró que era mejor eliminar aquellos registros.

- Se eliminaron los valores fuera de rango de las variables UsoLínea y DebtRatio (expresadas en porcentaje), exceptuando aquellos que tenían un valor faltante en la variable Ingreso, esto porque se consideró ajustarlos posteriormente.

4.2.2 Valores Faltantes

En la base de datos hay dos variables que presentan valores faltantes: Ingreso (19,8%) y Dependientes (2,61%). Además si el registro tiene valor faltante en la variable Dependientes también lo tiene en la variable Ingreso (lo contrario no es siempre cierto).

La variable ingreso presenta registros con valor 0 o muy pequeños que generan inconsistencias en la variable DebtRatio, que claramente está relacionada. Todos los ingresos menores o iguales a 500, que corresponden a un 1,9%, serán tratados como valores faltantes, y por lo tanto el porcentaje a tratar en esta variable asciende a 21,7%.

Al analizar el comportamiento de los clientes que presentan valores faltantes en la variable ingreso y los que no, se puede construir la siguiente tabla:

Tabla 2: Frecuencia de la variable target en muestras con y sin valores faltantes

		Target		Total
Ingreso		1	0	
Si missing	1	1782	30777	32559
No missing	0	8244	109197	117441
Total		10026	139974	150000

Tabla 3: Porcentaje de la variable target en muestra con y sin valores faltantes

		Target	
Ingreso		1	0
Si missing	1	5,47%	94,52%
No missing	0	7,02%	92,98%

En la segunda tabla se puede apreciar los porcentajes de clientes buenos y malos en cada muestra de la base original. Se concluye que cuando no hay valores faltantes los clientes son levemente más malos que cuando se incorporan los registros con valores faltantes (7,02% versus 6,6% de clientes malos respectivamente).

Si se supone que la distribución “teórica” de la muestra, en términos de la proporción de clientes buenos y malos, es la de la base completa y se realiza un test chi-cuadrado de ajuste, considerando como distribución observada aquella presente en la base sin valores faltantes, el p-valor arrojado por el test es prácticamente cero y por lo tanto se podría concluir que el eliminar los registros con valores faltantes en ingreso no afecta de manera sustancial el modelo final. Dado lo anterior y considerando que la imputación de datos distorsiona la distribución de las variables se optó por eliminar también estos registros.

El resultado final de esta etapa fue una base de datos limpia y sin inconsistencias, la cantidad de registros disminuyó a 103367 registros.

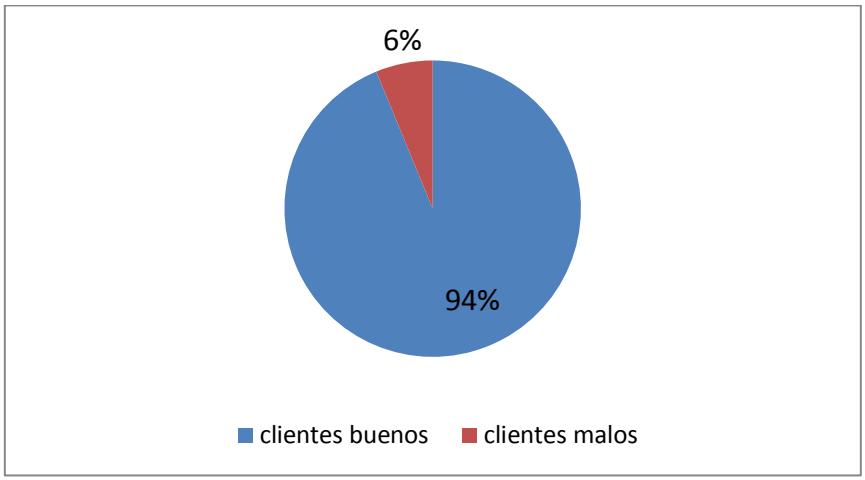
En la etapa de transformación de variables se categorizaron las variables Mora3060, Mora6090, Mora90, OtrosProd, NumHip y Dependientes. Finalmente, en la etapa de selección de variables se calcularon las correlaciones de las variables independientes con la variable objetivo, encontrando que todas son significativas para el análisis, razón por la cual no se eliminó ninguna de ella.

4.3 DESCRIPCIÓN DE LAS VARIABLES

Esta sección está dedicada a analizar en detalle cada una de las variables presentes en la base de datos. Conocer la distribución de cada una de ellas es importante al aplicar la metodología propuesta y por lo tanto es fundamental realizar este análisis.

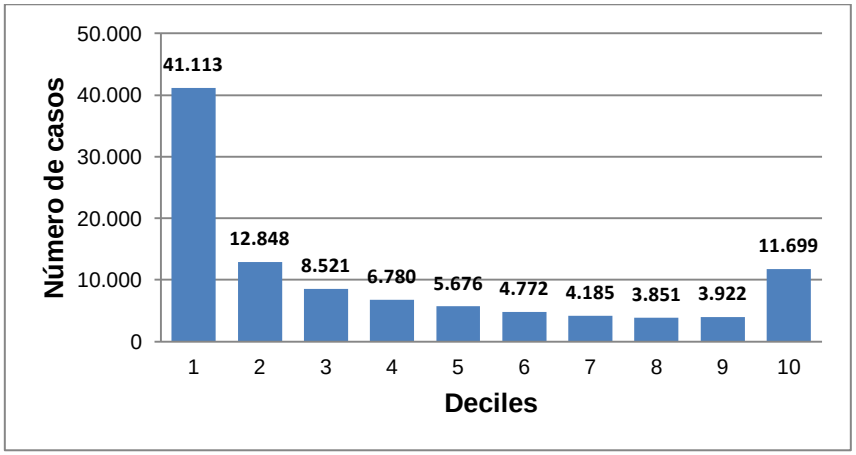
La primera variable en estudio es la variable dependiente del modelo, llamada Target. Su naturaleza es dicotómica y en la siguiente figura se muestra la proporción de clientes que caen en default y los clientes que no.

Ilustración 3: Distribución de la Variable Target



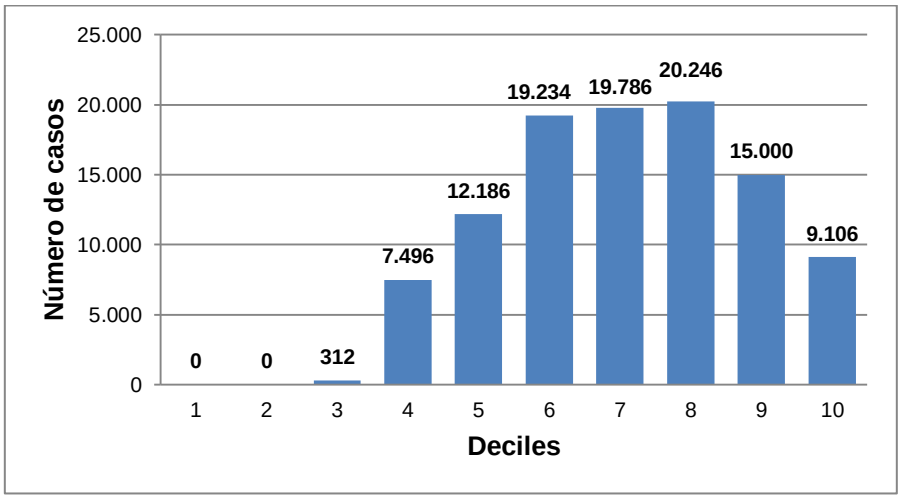
La variable UsoLínea, que indica el saldo total utilizado de las tarjetas y líneas de crédito personales, excepto bienes raíces y la deuda a plazo; dividido por la suma de los límites de crédito, por lo tanto es una variable en porcentaje. El histograma frecuencias es el siguiente.

Ilustración 4: Histograma de frecuencias de la variable UsoLínea



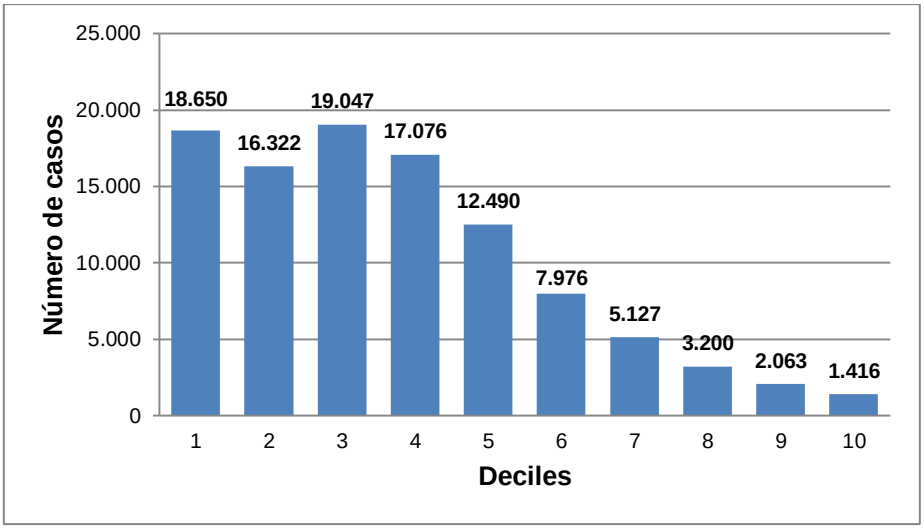
La Variable Edad, que indica la edad en años de cada uno de los clientes es una variable continua en el rango 21-75 años, tiene el siguiente histograma de frecuencias:

Ilustración 5: Histograma de frecuencias de la Variable Edad



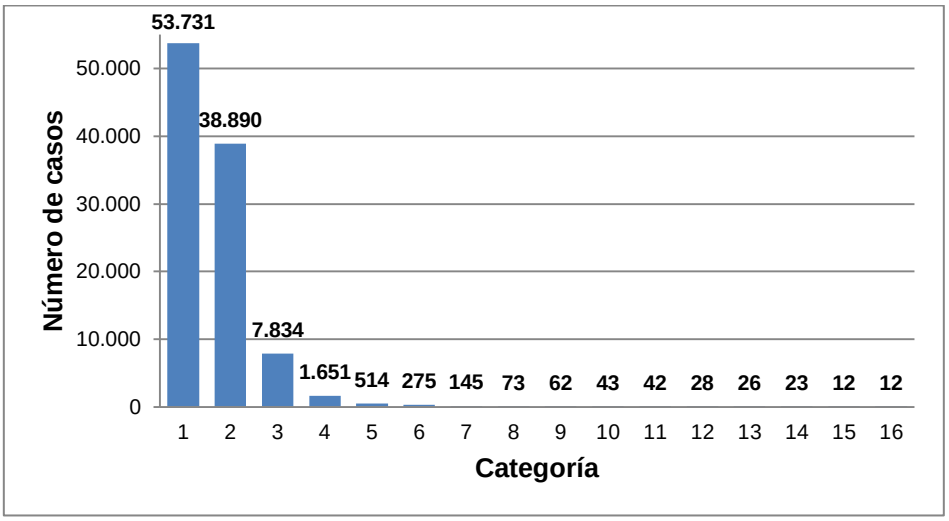
La variable DebtRatio representa los pagos mensuales de la deuda, pensiones alimenticias, los costos de vida, dividido por los ingresos brutos mensuales, por lo tanto es una variable en porcentaje altamente correlacionada con el ingreso. El histograma de frecuencias es el siguiente:

Ilustración 6: Histograma de frecuencias de la variable DebtRatio



La variable Ingreso representa el ingreso mensual de cada cliente, es una variable continua en el rango 500-100000, cuya media es 6670 aproximadamente. El histograma de frecuencias es el siguiente:

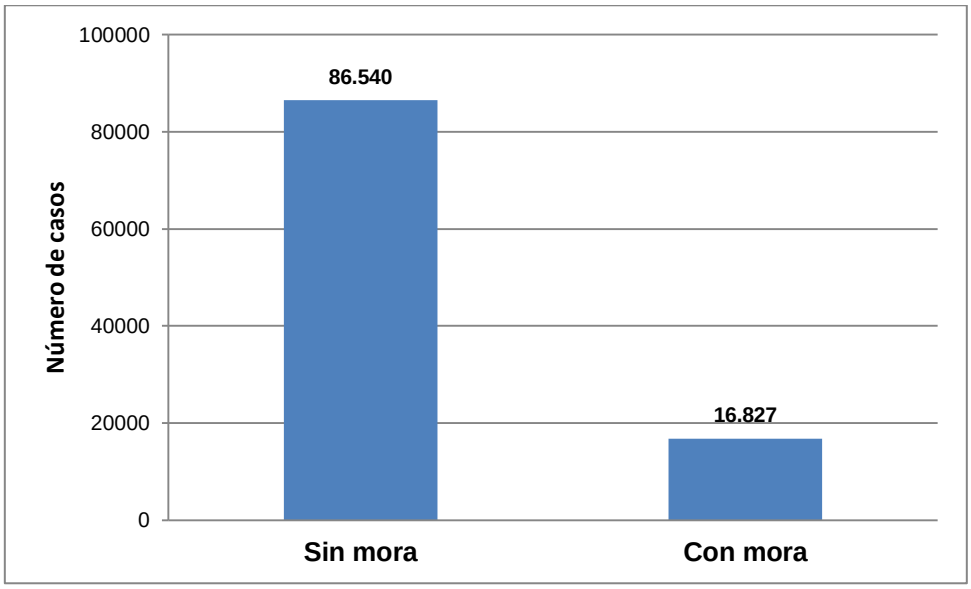
Ilustración 7: Histograma de frecuencias de la variable Ingreso



Se puede observar que la mayoría de los clientes está en los primeros segmentos, donde el primero de ellos, que representa ingresos menores a 6000, agrupa aproximadamente al 52% de la muestra, mientras que el segundo, correspondiente a ingresos entre 6000 y 12000, agrupa a cerca del 38% de la muestra.

La variable Mora 3060 fue categorizada, para ello se crearon 2 clases: clientes sin mora y clientes con mora. La frecuencia de cada categoría se presenta en la figura n° 7, donde claramente se puede apreciar que los clientes morosos son muy pocos en comparación con lo que no presentan mora (un 16,3% versus un 83,7%).

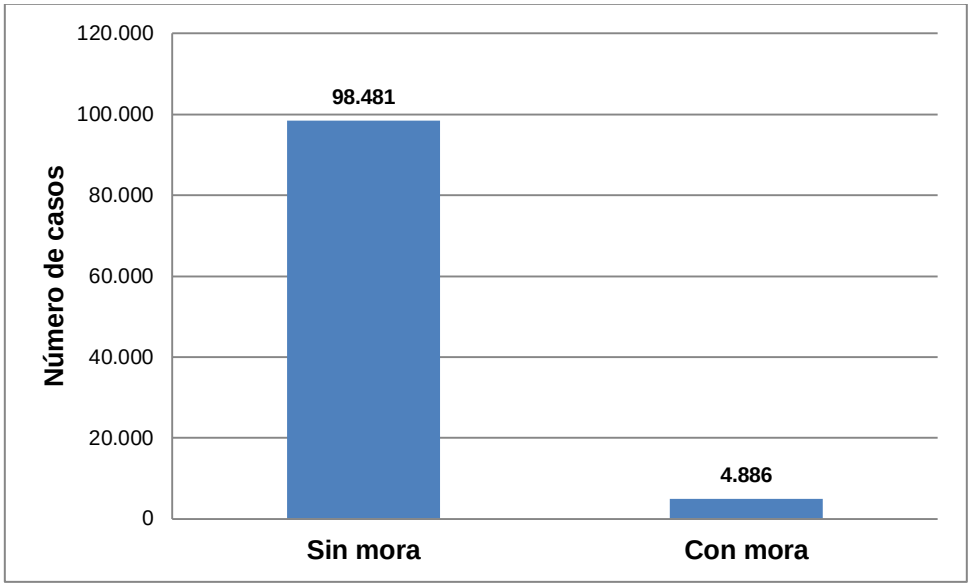
Ilustración 8: Histograma de frecuencias de la variable Mora3060



La variable Mora 6090 fue categorizada de la misma manera que la variable anterior. Nuevamente la proporción de clientes morosos es mucho menor a la de clientes sin

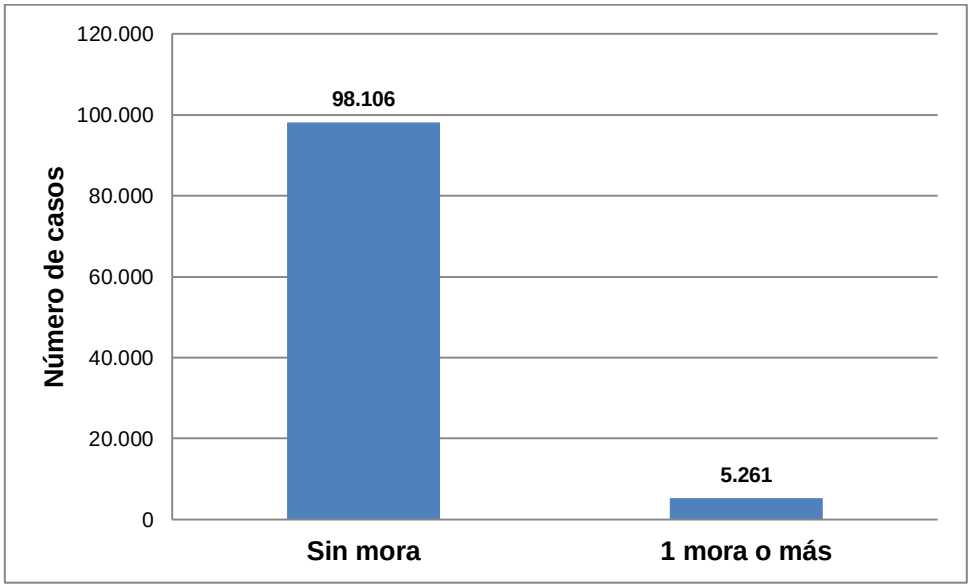
mora e incluso se mejora respecto a la distribución anterior, como se puede observar en la siguiente figura:

Ilustración 9: Histograma de frecuencias de la variable Mora6090



La variable Mora90 fue categorizada de la misma manera que las dos variables anteriores. Nuevamente la proporción de clientes morosos es mucho menor a la de clientes sin mora. En la siguiente figura se detalla la cantidad de clientes con mora y sin mora:

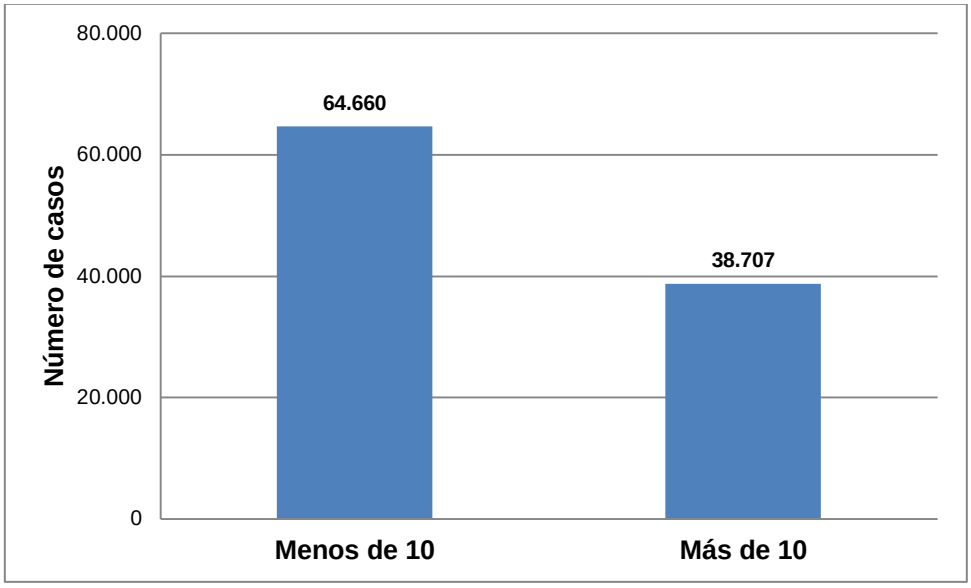
Ilustración 10: Histograma de Frecuencias para la variable Mora90



La variable OtrosProd indica el número de créditos abiertos y líneas de crédito que tiene un cliente y también fue categorizada en las siguientes clases: Menos de 10 productos,

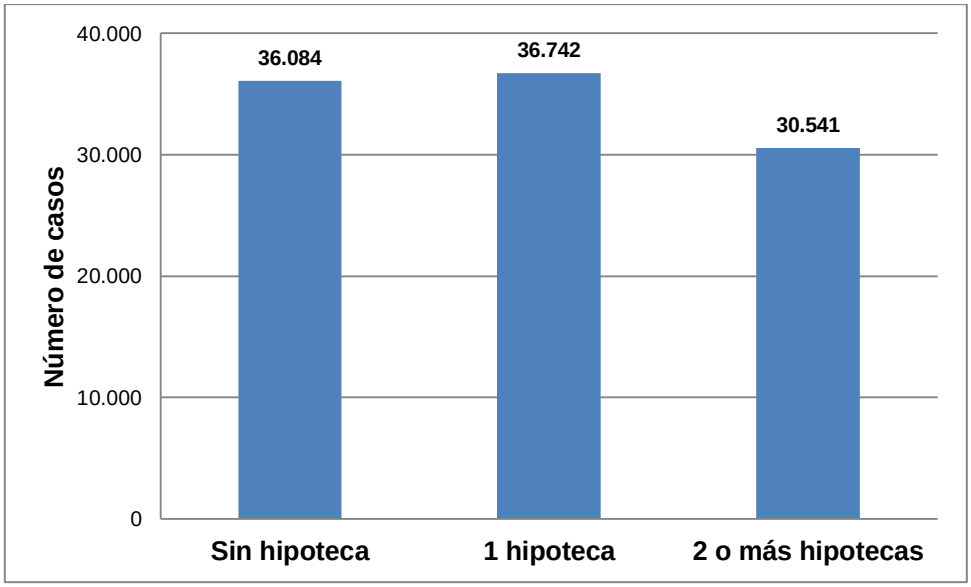
entre 10 y más de 10 productos. La frecuencia de cada categoría se presenta a continuación.

Ilustración 11: Histograma de frecuencias de la variable OtrosProd



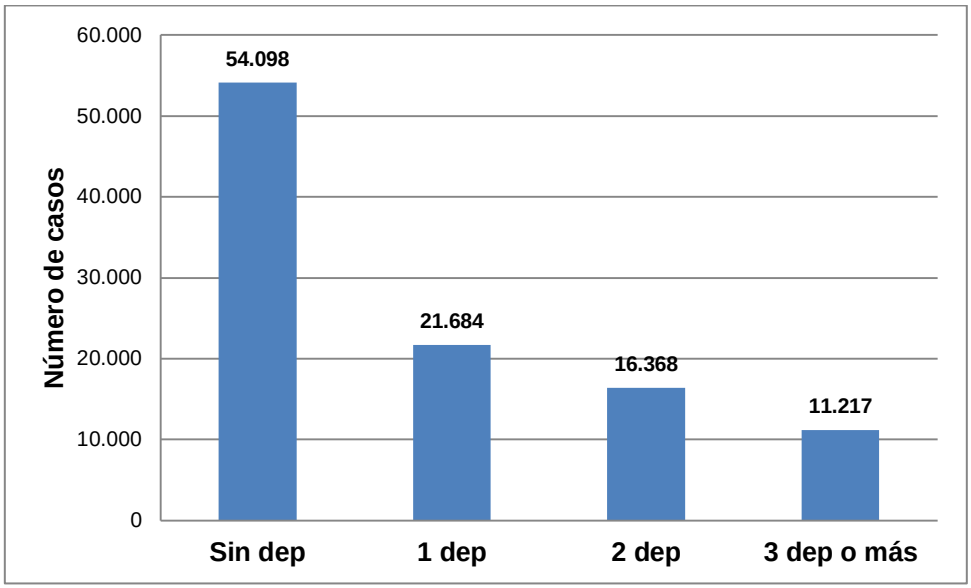
La variable NumHip representa el número de hipotecas y préstamos de bienes raíces, incluidas las líneas de crédito hipotecario y fue categorizada en las siguientes clases: sin hipoteca, 1 hipoteca, 2 o más hipotecas. La frecuencia de cada categoría se presenta en el siguiente gráfico.

Ilustración 12: Histograma de frecuencias para la variable NumHip



Por último, la variable Dependientes, que indica el número de personas dependientes en la familia (cónyuge, hijos, etc.) excluyendo al mismo cliente fue categorizada en las siguientes clases: sin dependientes, 1 dependiente, 2 dependientes y 3 o más dependientes. La frecuencia de cada categoría se presenta en el siguiente gráfico.

Ilustración 13; Histograma de Frecuencias para la variable Dependientes



Finalmente, se utilizó el software Easy Fit 5.5 para ajustar distribuciones de probabilidad conocidas a cada una de las variables continuas. En la siguiente tabla se resume la distribución de cada una de las variables del modelo:

Tabla 4: Distribución de probabilidad de las variables del modelo

Variable	Tipo	Distribución
UsoLínea	Continua	Beta(0,30032 ; 0,63821)
Edad	Continua	Johnson SB(-0,12247 ; 1,0443 ; 62,672 ; 16,985)
DebtRatio	Continua	Johnson SB (1,2534 ; 1,3222 ; 1,5001 ; -0,13309)
Ingreso	Continua	Burr (1,4005 ; 2,4919 ; 6927,4 ; 0)
Mora3060	Dicotómica	Bernoulli (0,162)
Mora6090	Dicotómica	Bernoulli (0,047)
Mora90	Dicotómica	Bernoulli (0,05)
NumHip	Categórica	Multinomial (3; 0,35; 0,35; 0,3)
OtrosProd	Dicotómica	Bernoulli(0,37)
Dependientes	Categórica	Multinomial(4; 0,52; 0,21; 0,16; 0,11)

4.4 CONSTRUCCIÓN DEL MODELO DE REGRESIÓN

Posterior al procesamiento de la base de datos y antes de calibrar el modelo de regresión, es necesario balancear la base de datos. Se puede constatar que hay una desproporción entre clientes buenos (aproximadamente 94%) y clientes malos (aproximadamente 6%), por lo que el desempeño del modelo sería sesgado, reconociendo de mejor manera a la clase dominante en desmedro de la clase menos numerosa. De hecho si se realiza una regresión logística de prueba para ver cómo se comporta el modelo, el error de predicción de la clase menos numerosa (los clientes malos) es de 83,3% lo cual es muy elevado.

Existen varias técnicas que permiten manejar este problema, dentro de las cuales se pueden distinguir dos enfoques: en el primero se opta por la asignación de un costo diferencial a las instancias de entrenamiento según las frecuencias de las clases, mientras que en el segundo se remuestrea el conjunto de datos originales, ya sea agregando casos repetidos de la clase minoritaria o submuestreando las clases mayoritarias (Drozdowicz *et al*., 2007).

Se adoptó el segundo criterio para balancear la base de datos, es decir, realizar un remuestreo de la base y obtener de esta forma un conjunto más equitativo para que el modelo no sea sesgado. Se hicieron pruebas aumentando la proporción de los casos menos numerosos hasta alcanzar un porcentaje de la clase más numerosa y también tomando submuestras de la clase más numerosa para adaptarlas al tamaño de la clase menos numerosa, también manteniendo una proporción definida. El detalle de estas pruebas puede ser consultado en el anexo B, acá se presenta el modelo que se consideró tenía mejor rendimiento (aquel con *undersampling* en una proporción 40% malos y 60% buenos).

4.4.1 Parámetros del Modelo

El modelo fue calibrado en el software SPSS Statistics 19.0, utilizando los métodos *backward, forward y wrapper,* para determinar las variables más significativas. Todos los métodos convergieron al mismo modelo, cuyos parámetros calibrados fueron los siguientes:

Tabla 5: Variables en la ecuación

Variable	Tipo	Beta	E.T.	p-valor de Wald
UsoLínea	Continua	1.820	0,059	0
Edad	Continua	-0,17	0,002	0
Mora3060	Dicotómica	1.054	0,045	0
DebtRatio	Continua	1.068	0,116	0
Ingreso	Continua		0	0,443
OtrosProd	Dicotómica	-0,5075	0,115	0
Mora90	Dicotómica	1,840	0,072	0
NumHip	Categórica			0
Categoría 1		0,381	0,069	0
Categoría 2		-0,075	0,053	0,157
Mora6090	Dicotómica	1,355	0,075	0
Dependientes	Categórica			0,140
Categoría 1		-0,143	0,064	0,025
Categoría 2		-0,096	0,07	0,173
Categoría3		-0.074	0,073	0,314
Constante		-0,908	0,177	0

La Tabla n°5 muestra para cada variable su tipo, el valor del parámetro calibrado, su error estándar y su significancia estadística en base al test de Wald, que contrasta la hipótesis de que el parámetro asociado a la variable tiene valor cero y el valor del parámetro calibrado. Se observa que ciertas variables tienen un p-valor que no permite rechazar la hipótesis nula del test de Wald y que por lo tanto no son significativas en el modelo, estas variable son: Ingreso, la categoría 2 de NumHip y Dependientes (donde solo sería significativa la categoría 1, que corresponde a ningún dependiente).

4.4.2 Ajuste del Modelo

El software SPSS calcula una serie de indicadores del ajuste y desempeño del modelo, para ello divide la totalidad de los datos en un subconjunto de entrenamiento (70% de la base) y un subconjunto de validación (30% de la base). El procedimiento es entonces calibrar un modelo en base al conjunto de entrenamiento y luego predecir los valores de la variable objetivo del conjunto de validación.

Una vez obtenidas las predicciones, se calculan los indicadores de bondad de ajuste, que se describen a continuación.

Bondad de Ajuste de Hosmer-Lemeshow

Este indicador permite evaluar el ajuste global del modelo, es decir, que tanto coincide lo predicho con la salida real observada. Agrupa los casos en deciles de riesgo y compara las frecuencias observadas con las esperadas dentro de cada subconjunto utilizando el test χ^2.

Tabla 6: Prueba de Hosmer-Lemeshow

Paso	Chi-cuadrado	g.l.	Sig.
1	94,186	8	0

Tabla 7: Tabla de contingencias para la prueba de Hosmer y Lemeshow

	Target = 0		Target = 1		Total
	Observado	Esperado	Observado	Esperado	
1	1516	1475,126	89	129,874	1605
2	1458	1421,360	147	183,640	1605
3	1388	1365,676	217	239,324	1605
4	1275	1287,670	330	317,330	1605
5	1164	1170,758	441	434,242	1605
6	967	1021,444	638	583,556	1605
7	828	853,012	777	751,988	1605
8	583	624,889	1022	980,111	1605
9	299	323,727	1306	1281,273	1605
10	153	87,338	1454	1519,662	1605

En la tabla se observa que las desviaciones son bastante pequeñas dentro de los grupos. Esto queda ratificado por el resultado del test de Hosmer-Lemeshow (Tabla n°6) el cual arrojó un p-valor de 0, lo cual valida el ajuste del modelo.

R^2 Cox-Snell y R^2 de Nagelkerke

Este indicador determina la proporción de varianza de la variable dependiente explicada por las variables predictoras (independientes). Se basa en la comparación del logaritmo de la función de verosimilitud para el modelo con respecto al logaritmo de la verosimilitud para un modelo de línea base. Los valores de este indicador varían entre 0 y 1, donde 0 significa que las variables independientes no explican de manera alguna la varianza de la variable dependiente y 1 significa que las variables independientes explican la totalidad de la varianza de la variable dependiente.

Al igual que el indicador anterior, la R^2 de Nagelkerke se utiliza para estimar la proporción de varianza de la variable objetivo que es explicada por las variables independientes. Es una versión corregida de la R^2 de Cox-Snell, pues ésta tiene un valor máximo inferior a 1, incluso para un modelo perfecto. La R^2 de Nagelkerke corrige la escala del estadístico para cubrir el rango completo de 0 a 1.

Los resultados obtenidos son:

Tabla 8: Resumen del modelo

Paso	-2log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	15520,726	0,316	0,427

Matriz de Confusión

Es una herramienta que permite evaluar el desempeño de un modelo de clasificación comparando las predicciones de éste con los valores reales observados en la muestra de datos.

Cada columna de la matriz de confusión representa el número de predicciones de cada clase, mientras que cada fila representa el número de instancias en la clase real. Una matriz de confusión es de la forma:

Tabla 9: Matriz de confusión genérica

Observado	Pronosticado	
	Clase 1	Clase2
Clase 1	A	B
Clase 2	C	D

- A es el número de observaciones de la clase 1 predichas correctamente
- B es el número de observaciones de la clase 1 predichas incorrectamente
- C es el número de observaciones de la clase 2 predichas incorrectamente
- D es el número de observaciones de la clase 2 predichas correctamente.

Con esta información es posible calcular indicadores de error en la predicción, como los descritos a continuación.

Error total: Mide el desempeño global del modelo, asumiendo que el costo de clasificar de manera errónea las observaciones de ambas clases tiene el mismo costo. Se calcula como:

$$Error\ Total = \frac{B + C}{n°\ total\ observaciones}$$

Error de tipo I: Es la proporción de observaciones clasificadas erróneamente de la clase 1. En este caso corresponde a clasificar mal a los clientes buenos, es decir, se les rechazaría el crédito- Se calcula como:

$$Error\ tipo\ I = \frac{B}{A+B}$$

Error de tipo II: Es la proporción de observaciones clasificadas erróneamente de la clase 2. En este caso corresponde a clasificar mal a los malos clientes, es decir, se le asignaría un crédito a un cliente *default,* lo que implica pérdidas monetarias directas por el incumplimiento del pago de la deuda. Se calcula como:

$$Error\ tipo\ II = \frac{C}{C+D}$$

Para este caso, la matriz de confusión es la siguiente:

Tabla 10: Matriz de confusión del Modelo

Observado	Pronosticado		
	Target		Porcentaje correcto
	0	1	
Target 0	7649	1982	79,4
Target 1	1757	4664	72,6
Porcentaje global			76,7

Con ello es posible calcular los errores del modelo de regresión:

- Error Global: 23,3%
- Error de tipo I: 20,6%
- Error de tipo 2: 27,4%

Área bajo la Curva ROC

Como medida de la capacidad de discriminación del modelo también se usa el área bajo la curva ROC, conocida como AUC. La curva ROC (Receiver Operating Characteristic) grafica la proporción of salidas clasificadas incorrectamente y la proporción de salidas correctamente clasificadas, dado cierto punto de corte (Stevenson, 2008).

El área bajo la curva ROC (AUC), cuyo valor fluctúa entre 0 y 1, provee una medida de la habilidad del modelo para discriminar entre aquellos registros que experimentan el resultado de interés (en este caso la mora) versus aquellos que no (Hosmer & Lemeshow, 1999). En particular:

- $AUC \in [0{,}5\ ;\ 6)$ indica que el modelo es malo
- $AUC \in [0{,}6\ ;\ 0{,}75)$ indica que el modelo es regular

- $AUC \in [0{,}75\,;\,0{,}9)$ indica que el modelo es bueno
- $AUC \in [0{,}9\,;\,0{,}97)$ indica que el test es muy bueno
- $AUC \in [0{,}97\,;\,0{,}1)$ indica que el modelo es excelente

Para este caso, al calcular AUC, se obtuvo el siguiente resultado:

Tabla 11: AUC

Área	Error Típico	Sig. Asintótica	Intervalo de confianza asintótico al 95%	
			Límite inferior	Límite superior
0,840	0,003	0	0,834	0,846

De lo anterior se puede concluir que el modelo es bueno, dado que su AUC es de 0,840.

CAPÍTULO 5

MODELOS DE SEGUIMIENTO

5.1 APLICACIÓN DE LOS MODELOS DE SEGUIMIENTO

En este capítulo se realizará la aplicación de las metodologías de seguimiento para determinar cambios en la distribución de las variables escogidas y descritas en el capítulo 2 junto con la aplicación de los dos modelos propuestos.

Como no se contaba con la evolución de las variables en el tiempo, la forma de abordar este problema fue realizar perturbaciones (variaciones) a los parámetros de la distribución de cada una de las variables del modelo, obteniendo con ello nuevas muestras de datos que se utilizaron como input en cada una de las metodologías de seguimiento. Se utilizó el programa EasyFit 5.5 que permite generar números aleatorios de acuerdo a la distribución que el usuario seleccione. Este programa es capaz de generar 5000 números, por lo que esa será la cantidad de datos que habrá en cada muestra, tanto la original como la generada mediante las perturbaciones.

Uno de los objetivos de este trabajo era crear una metodología de seguimiento que reconociese cambios significativos en la distribución de las variables, esto es, que no fuese extremadamente sensible reconociendo cambios ínfimos en la distribución que finalmente no tienen un impacto significativo en el modelo. Es por ello que se realizaron varios tipos de perturbaciones, con el fin de testear la sensibilidad de cada una de las metodologías. Se entiende por perturbación a una variación de los parámetros que definen la distribución de probabilidad de una variable, sumándole o restándole ruido al valor original de dicho parámetro.

Para cada variable del modelo ser realizaron 3 perturbaciones:

- Perturbación pequeña
- Perturbación mediana
- Perturbación grande

Para los test Beta 1, ICV-1 e ICV-2 es necesario calcular el intervalo de cambio máximo permitido para cada variable en función de los parámetros encontrados al realizar la regresión logística y la desviación estándar de dicha estimación. Dichos intervalos se presentan en la siguiente tabla:

Tabla 12: Intervalo de cambio máximo permitido por variable

Variable	Tipo	Coeficiente	Desviación	Límite Inferior	Límite superior
UsoLínea	Continua	1.820	0,059	0,935	1,064
Edad	Continua	-0,17	0,002	0,976	1,023
DebtRatio	Continua	1.068	0,116	0,782	1,217
Ingreso	**Continua**		**0**		
Mora3060	Dicotómica	1.054	0,045	0,914	1,085
Mora6090	Dicotómica	1,355	0,075	0,889	1,110
Mora90	Dicotómica	1,840	0,072	0,921	1,078
OtrosProd	Dicotómica	-0,5075	0,115	0,546	1,453
NumHip	**Categórica**				
Categoría 1		**0,381**	**0,069**	**0,637**	**1,362**
Categoría 2		**-0,075**	**0,053**	**-0,413**	**2,413**
Dependientes	**Categórica**				
Categoría 1		**-0,143**	**5,050**	**-69,629**	**71,629**
Categoría 2		**-0,096**	**1,854**	**-37,625**	**39,625**
Categoría3		**-0,074**	**1,014**	**-26,405**	**28,405**

Las variables destacadas son aquellas no significativas de acuerdo al test de Wald tras el resultado de la regresión, por lo que no serán sometidas a perturbaciones.

5.1.1 Test ICV-1

A cada variable se le realizaron 3 perturbaciones, una pequeña, una mediana y una grande, en el Anexo C se pueden observar los gráficos de las perturbaciones por variable, para tener una noción más clara de los cambios en la distribución. Los resultados obtenidos al aplicar este test son los siguientes:

Tabla 13: Resultados del test ICV-1

Variable	ICV-1 perturbación 1	ICV-1 perturbación 2	ICV-1 perturbación 3	Límite inferior	Límite superior
UsoLínea	**0,45**	**0,22**	**0,10**	0,935	1,064
Edad	**0,01**	**0,09**	**2,47**	0,976	1,023
DebtRatio	**4,92**	**1,17**	**26,20**	0,782	1,217
Mora3060	**0,03**	**0,23**	**0,84**	0,914	1,085
Mora6090	**0,06**	**0,28**	0,93	0,889	1,110
Mora90	**0,08**	**0,36**	0,93	0,921	1,078
OtrosProd	**0,07**	0,9	**3,4**	0,546	1,453

Los números presentados en la tabla corresponden al valor del estadístico para cada caso. En rojo se indican los valores del test que caen fuera del intervalo de cambio máximo permitido, lo que significa que se considera que el cambio en las distribuciones es significativo.

5.1.2 Test ICV-2

Al igual que el caso anterior, a cada variable se le realizaron 3 perturbaciones, una pequeña, una mediana y una grande. Los resultados obtenidos al aplicar este test son los siguientes:

Tabla 14: Resultados del test ICV-2

Variable	ICV-2 perturbación 1	ICV-2 perturbación 2	ICV-2 perturbación 3	Límite inferior	Límite superior
UsoLínea	0,86	**0,93**	**0,97**	0,935	1,064
Edad	0,98	**0,79**	**0,11**	0,976	1,023
DebtRatio	1,17	**4,92**	**26,2**	0,782	1,217
Mora3060	0,97	**0,84**	**0,59**	0,914	1,085
Mora6090	0,95	**0,79**	**0,53**	0,889	1,110
Mora90	0,93	**0,74**	**0,53**	0,921	1,078
OtrosProd	0,957	0,638	**0,319**	0,546	1,453

Los números presentados en la tabla corresponden al valor del estadístico para cada caso. En rojo se indican los valores del test que caen fuera del intervalo de cambio máximo permitido, lo que significa que se considera que el cambio en las distribuciones es significativo.

5.1.3 Test Beta 1

A cada variable se le realizaron 3 perturbaciones, una pequeña, una mediana y una grande. Los resultados obtenidos al aplicar este test son los siguientes:

Tabla 15: Resultados del test Beta 1

Variable	ICV-2 perturbación 1	ICV-2 perturbación 2	ICV-2 perturbación 3	Límite inferior	Límite superior
UsoLínea	**0,89**	**0,90**	0,95	0,935	1,064
Edad	1,01	**0,93**	**0,94**	0,976	1,023
DebtRatio	**0,62**	0,89	**0,43**	0,782	1,217
Mora3060	**0,87**	**0,54**	**0,32**	0,914	1,085
Mora6090	**0,47**	**0,18**	**0,09**	0,889	1,110
Mora90	**0,40**	**0,16**	**0,09**	0,921	1,078
OtrosProd	0,933	0,622	**0,466**	0,546	1,453

Nuevamente, los números presentados en la tabla corresponden al valor del estadístico para cada caso y en rojo se indican los valores del test que caen fuera del intervalo de cambio máximo permitido, lo que significa que se considera que el cambio en las distribuciones es significativo.

5.1.4 Stability Index

Stability Index es un indicador de la similitud entre la muestra observada en el periodo t, y la muestra utilizada para construir el modelo (Baesens et al., 2010). Se calcula como:

$$SI = \sum_{i=1}^{m} (E_i - P_i) ln \frac{E_i}{P_i}$$

Donde m es el número de clases consideradas, E_i representa el porcentaje de observaciones de la clase m en el conjunto de entrenamiento del modelo (muestra original) y P_i corresponde al porcentaje de observaciones de la clase m del conjunto de prueba del modelo (nueva muestra).

Los criterios para determinar si existe un cambio significativo en la distribución de las variables son:

- $SI < 0{,}1$ Ha ocurrido un cambio no significativo
- $0{,}1 < SI < 0{,}25$ Ha ocurrido un cambio menor
- $SI > 0{,}25$ Ha ocurrido un cambio significativo

Frente a las perturbaciones realizadas los resultados obtenidos fueron los siguientes:

Tabla 16: Resultados del test Stability Index

Variable	SI perturbación 1	SI perturbación 2	SI perturbación 3
UsoLínea	0,03	0,02	0,01
Edad	0,004	**5,681**	**2,09**
DebtRatio	**0,19**	**0,74**	**4,13**
Mora3060	0,0006	0,07	**0,49**
Mora6090	0,002	**0,19**	**0,96**
Mora90	0,005	**0,23**	**0,85**
OtrosProd	0,001	**0,183**	**0,319**

En rojo se indican los valores del Stability Index que determinan que el cambio sufrido en la distribución de las variables debido a la perturbación es significativo, mientras que en azul se indican aquellas perturbaciones que producen un cambio menor, de acuerdo a la definición de este índice.

5.1.5 Test Kolmogorov-Smirnov

El test KS para variables continuas, se utiliza para determinar si hay divergencia entre las distribuciones de probabilidad de dos muestras independientes y para contrastar la distribución empírica de una muestra contra una distribución teórica (Cieslak et al., 2007).

Se aplicó el test a las variables continuas del modelo, es decir, a Usolínea, Edad y DebtRatio, frente a las tres perturbaciones realizadas, calculando el valor del siguiente estadístico:

$$Z_{KS} = max_i\{|D_i|\}$$

Donde $D_i = F_1(X_i) - F_2(X_i)$ y $F_1(X_i)$ es la frecuencia relativa acumulada de la variable original y $F_2(X_i)$ es la frecuencia relativa acumulada de la variable perturbada. En la siguiente tabla se muestran los valores del estadístico, obtenidos para cada perturbación de cada variable:

Tabla 17: Estadísticos de prueba del Test K-S

Variable	Perturbación 1	Perturbación 2	Perturbación 3
UsoLínea	**0,08**	**0,07**	**0,05**
Edad	**0,02**	**0,09**	**0,22**
DebtRatio	**0,06**	**0,316**	**0,77**

El valor teórico de $D_i = F_1(X_i) - F_2(X_i)$ para estas muestras que contienen 5000 registros es de 0,019, por lo tanto, la regla de decisión es la siguiente:

- Si Valor del estadístico < 0,019 se acepta la hipótesis nula, es decir, no hay divergencia entre las dos muestras
- Si Valor del estadístico > 0,019 se rechaza la hipótesis nula, es decir, hay divergencia entre las muestras.

En este caso todos valores del estadístico calculados para cada una de las perturbaciones de las variables son mayores a 0,019, por lo tanto se concluye que todas las perturbaciones son significativas pues generan divergencia entre las muestras.

5.1.6 Distancia de Hellinger

La distancia de Hellinger es un indicador de la divergencia entre la distribución de dos variables categóricas. Sean X e Y dos poblaciones con p categorías, la distancia de Hellinger entre ellas se calcula como:

$$Hellinger\,(X,Y) = \sqrt{\sum_{j=1}^{p}\left(\sqrt{\frac{X_j}{n_X}} - \sqrt{\frac{Y_j}{n_Y}}\right)^2}$$

Donde X_j corresponde al número de observaciones de la clase j en la variable X, Y_j corresponde al número de observaciones de la clase j en la variable Y, n_X y n_Y son el número total de observaciones de la variable X e Y, respectivamente. Este indicador alcanza un valor mínimo 0, lo que significa que las distribuciones son idénticas y un valor máximo de $\sqrt{2}$, en cuyo caso las distribuciones son totalmente divergentes.

Se aplicó el test a las variables categóricas del modelo, es decir, a Mora3060, Mora6090 y Mora 90, frente a las tres perturbaciones realizadas, obteniendo los siguientes resultados:

Tabla 18: Resultado del Test de Distancia de Hellinger

Variable	Perturbación 1	Perturbación 2	Perturbación 3
Mora3060	0,03	**0,164**	**0,377**
Mora6090	0,099	**0,303**	**0,559**
Mora90	0,12	**0,34**	**0,55**
OtrosProd	0,027	**0,22**	**0,44**

Considerando los valores límite de la distancia de Hellinger, se considera que no hay cambios significativos en la distribución de las variables tras la perturbación 1, hay cambios menores en las distribuciones tras la perturbación 2 y hay cambios significativos tras la perturbación 3.

5.1.7 Test Chi-cuadrado

El test χ^2 puede ser utilizado para determinar si una distribución empírica de una variable coincide con una distribución de probabilidad conocida para esa variable (Zeira et al., 2005)**.** Si el conjunto de entrenamiento es lo suficientemente grande, entonces es correcto asumir que la verdadera distribución de la variable será bien estimada por éste.

Por lo tanto, este test permitirá comparar la distribución de la variable que tienen las muestras perturbadas (distribución empírica) con la del conjunto de entrenamiento (verdadera distribución de la variable).

El estadístico de prueba se calcula como:

$$\chi_p^2 = n_K \sum_{i=1}^{j} \frac{\left(\frac{x_{i,K}}{n_K} - \frac{x_{i,K-1}}{n_{K-1}}\right)^2}{\frac{x_{i,K-1}}{n_{K-1}}}$$

Este estadístico sigue una distribución χ^2 de $j-1$ grados de libertad. Por lo tanto, si el valor del estadístico χ^2_{j-1} es mayor que el valor crítico de la distribución χ^2, entonces se rechazará la hipótesis nula, es decir, la variable no es estacionaria entre períodos y si presenta cambios en su distribución. Los resultados obtenidos al aplicar este test se presentan en la siguiente tabla:

Tabla 19: Resultados del test chi-cuadrado

Variable	Valor del Estadístico			Valor Crítico
	Perturbación1	Perturbación2	Perturbación3	
UsoLínea	**185,7**	**112,7**	**67,7**	7,81
Edad	**22,3**	**608,5**	**3506,5**	7,81
DebtRatio	**146,4**	**3900,8**	**30457**	7,81
Mora3060	**19,84**	**3824,9**	**4429,1**	3,84
Mora6090	**292,2**	**22973,2**	**23821,4**	3,84
Mora90	**528,6**	**21879,8**	**22619,4**	3,84
OtrosProd	**15,35**	**2848,6**	**3890,3**	3,84

Notar que para todas las perturbaciones realizadas el valor del estadístico de prueba supera al valor crítico de la distribución chi-cuadrado, por lo tanto se rechaza la hipótesis nula en todos los casos y se concluye que la distribución de las variables ha cambiado al someterlas a las perturbaciones.

CAPÍTULO 6

ANÁLISIS DE RESULTADOS

6.1 MODELO DE REGRESIÓN LOGÍSTICA

El modelo de regresión logística resulta se bastante satisfactorio en términos del ajuste a los datos, por lo cual se puede concluir que representa bastante bien el fenómeno modelado. En la siguiente tabla se resumen los indicadores de ajuste del modelo calibrado:

Tabla 20: Indicadores de Ajuste del Modelo de Regresión

R^2 de Nagelkerke	42,7%
Error Global	23,3%
Error Tipo I	20,6%
Error tipo II	27,4%

El test de Hosmer-Lemeshow, que corresponde a un test de bondad de ajuste, evalúa qué tanto corresponde la probabilidad que arroja el modelo para cada registro con la información real que se tiene de ellos. Ambas distribuciones, la esperada y la observada son contrastadas mediante un test chi-cuadrado. El resultado en este caso fue un p-valor igual a cero, lo cual indica que el ajuste global del modelo es significativo.

Por otra parte, el R^2 de Nagelkerke, que indica la cantidad de varianza de la variable target que es explicada por las variables independientes toma un valor de 42,7%, lo que se considera satisfactorio más aun teniendo en cuenta que se trabaja con una gran cantidad de datos (sobre 100.000).

El error global del modelo es de un 23,3%, lo que significa que aproximadamente 2 de cada 10 clientes son mal clasificados. El error de tipo I, que corresponde a clasificar como mal cliente (aquel que presenta una mora de más de 90 días) a uno que es bueno es de 20,6%, lo que significa que 8 de cada 10 clientes de la institución financiera son clasificados correctamente. Por otro lado, el error de tipo II, que corresponde a clasificar como buen cliente a uno que es malo, asciende a un 27,4%, lo que significa que alrededor de 3 de cada 10 clientes malos serán clasificados como buenos y por ende no se tomarán medidas preventivas en cuanto a ellos.

Lo anterior podría ser preocupante, pero el minimizar el error de tipo II en un modelo de regresión logística implicaría penalizar la clasificación de clientes buenos, provocando incluso un sobreajuste en los datos (hacia los clientes malos). Más aun, si se considera que la cantidad de clientes malos en general es muy baja respecto de la cantidad de clientes buenos, generar modelos que predigan con exactitud el a los clientes morosos podría significar la pérdida de oportunidades con los clientes no morosos. Sin duda es

un problema que ha de ser evaluado con información a cerca de los costos que conlleva clasificar mal a uno u otro cliente, en base a ello se puede mejorar el modelo de clasificación.

Con este modelo se comprueba que la regresión logística es una buena herramienta par modelar fenómenos del ámbito financiero, en particular la probabilidad de mora de los clientes. Sin embargo existen técnicas más recientes de minería de datos que combinadas con este tipo de modelos podrían hacer del proceso de clasificación algo mucho más eficiente de lo que ya es. No es cuestionable la utilidad de los modelos de regresión logística para modelar fenómenos dicotómicos, ya que por un lado la implementación es muy sencilla y por otro la lectura de los resultados es de muy fácil interpretación (Hosmer & Lemeshow, 2000).

6.2 METODOLOGÍAS DE SEGUIMIENTO

Se aplicaron 5 modelos de seguimiento, además de los propuestos, para detectar cambios significativos en la distribución de las variables del modelo. Debido a que no se contaba con una base de datos temporal en la que se pudiese observar la evolución de cada variable a través del tiempo, se optó por encontrar la distribución actual de las variables y realizar perturbaciones, mediante la variación de los parámetros de dichas distribuciones. El objetivo primordial era estudiar como se comportaban las metodologías propuestas frente a cambios de diversa magnitud en la distribución de las variables y además compararlas con los modelos que ya existen. Los resultados obtenidos fueron los siguientes.

Tabla 21: Comportamiento de los modelos dada una perturbación pequeña

	Test ICV-1	Test ICV-2	Test Beta 1	Stability Index	Test KS	Distancia de Hellinger	Test chi-cuadrado
UsoLínea	X	O	X	O	X		X
Edad	X	O	O	O	X		X
DebtRatio	X	O	X	X	X		X
Mora3060	X	O	X	O		O	X
Mora6090	X	O	X	O		O	X
Mora90	X	O	X	O		O	X
OtrosProd	X	O	O	O		O	X

Tabla 22: Comportamiento de los modelos dada una perturbación mediana

	Test ICV-1	Test ICV-2	Test Beta 1	Stability Index	Test KS	Distancia de Hellinger	Test chi-cuadrado
UsoLínea	X	X	O	O	X		X
Edad	X	X	X	X	X		X
DebtRatio	X	X	X	X	X		X
Mora3060	X	X	X	O		X	X
Mora6090	X	X	X	O		X	X
Mora90	X	X	X	O		X	X
OtrosProd	O	O	O	O		X	X

Tabla 23: Comportamiento de los modelos dada una perturbación grande

	Test ICV-1	Test ICV-2	Test Beta 1	Stability Index	Test KS	Distancia de Hellinger	Test chi-cuadrado
UsoLínea	X	X	X	O	X		X
Edad	X	X	O	X	X		X
DebtRatio	X	X	X	X	X		X
Mora3060	X	X	X	X		X	X
Mora6090	O	X	X	X		X	X
Mora90	O	X	X	X		X	X
OtrosProd	X	X	X	X		X	X

Las tablas representan el comportamiento de cada test frente a una perturbación pequeña, mediana y grande, donde las cruces indican que el test respectivo determinó que había un cambio significativo en la distribución de la variable, dada la perturbación correspondiente y los círculos indican que el cambio en la distribución de las variables no fue significativo, y que por ende no afectaría el rendimiento del modelo original al utilizarlo en la nueva muestra.

Se observa que el primero de los test propuestos en este trabajo, el test ICV-1, determina (en su mayoría) que los cambios en la distribución de las variables son significativos para cualquiera de las perturbaciones que aquí se realizaron. Esto podría indicar dos cosas, o es extremadamente sensible o no funciona correctamente. Si el problema fuese la sensibilidad, entonces el test debería comportarse bien si se compara la distribución de dos muestras idénticas, en ese caso, el valor del estadístico debiese estar en el intervalo de cambio máximo definido en función de los parámetros de la regresión. Si esto no sucediese, la conclusión obvia sería que el test no funciona.

Para comprobar lo anterior se revisó el caso de una de las variables analizadas en este trabajo. Sea la variable Mora3060, cuya distribución de probabilidad es una Bernoulli de parámetro p = 0,162. Considérense dos muestras exactamente iguales con las siguientes características:

Tabla 24: Test ICV-1 para dos muestras idénticas

	Media	Varianza	ICV	Valor del test
Muestra 1	0,162	0,13	1,19	
Muestra 2	0,162	0,13	1,19	0

El valor del test corresponde a la diferencia de ICV entre cada una de las muestras. Por otra parte, el intervalo de confianza para la variable Mora3060 es [0,914 ; 1,085], por lo tanto el valor del estadístico cae fuera del intervalo, lo que significa que el cambio en la distribución de la variable entre las dos muestras es significativo, lo cual es una contradicción ya que las muestras son exactamente iguales.

Para entender por qué sucede esto es necesario revisar la forma en que se definió el intervalo de cambio máximo permitido:

$$C_i \in \left[\frac{\beta_i - 2\sigma_{\beta_i}}{\beta_i}, \frac{\beta_i + 2\sigma_{\beta_i}}{\beta_i}\right]$$

Si el modelo de regresión logística es bueno, la desviación estándar asociada al parámetro será pequeña y por lo tanto el intervalo de cambio máximo estará centrado en 1. Esto causa problemas con el test ICV-1, dado que al contrastar dos muestras mediante la resta de sus ICV, cuando éstas sean muy pequeñas el valor del estadístico será cercano a cero, lo que obviamente caería fuera del intervalo de cambio máximo.

Esto es posible de constatar al revisar las tablas anteriores donde el test reconoció que para una perturbación pequeña de todas las variables estos cambios eran significativos, también reconoció que para perturbaciones medianas solo la distribución de la variable OtrosProd no cambiaba de forma significativa, lo cual entra en contradicción con el resultado de la perturbación anterior. Y finalmente determinó que para perturbaciones grandes las variables Mora6090 y Mora90 no veían afectada su distribución de manera importante, lo que nuevamente entra en contradicción con los resultados de las perturbaciones anteriores.

Si se analiza los resultados obtenidos por el segundo de los test propuestos, ICV-2, se puede notar, en primer lugar, que para perturbaciones pequeñas (tabla n°21), se comporta de manera consistente con el Stability Index y con la distancia de Hellinger, lo cual habla de la robustez del método propuesto.

Si se le compara con el test Beta 1, que es muy similar en su formulación (de hecho se considera que ICV-2 es una corrección de este test), se observa que, para la misma perturbación pequeña, ICV-2 es menos sensible al reconocer cambios significativos que Beta 1. De hecho, ICV-2 no reconoce ningún cambio significativo en la distribución de las variables, mientras que Beta 1 reconoce que 5 de las 7 variables si han sufrido cambios importantes.

Ahora, si se analiza la tabla de perturbaciones medianas (tabla n°22) se observa que ICV-2 reconoce cambios significativos en 6 de las 7 variables, mientras que Beta 1 lo hace en 5 de las 7 variables y el Stability Index en solamente 2. Se podría pensar (si solo se mirara esta tabla) que ICV-2 es más sensible que Beta 1, pero se considera que los resultados no son concluyentes. El test Beta 1 presenta un problema en su formulación y es que al calcular su estadístico solo en función de las medias muestrales, se puede dar el caso que las distribuciones sean muy distintas en forma, pero que su media sea muy similar (como se ilustró en el Capítulo 3) y que por lo tanto el test arroje que el cambio en la distribución no es significativo, cuando claramente lo sería. Este tipo de distorsiones podrían generar conclusiones erradas respecto al funcionamiento de los test.

Al comparar con el Stability Index, se observa que el test ICV-2 es más sensible, ya que reconoce cambios significativos en todas las variables, mientras que Stability Index reconoce solo dos. Al comparar con la distancia de Hellinger, que en el caso anterior se comportaba de manera similar, se constata que tienen el mismo comportamiento para esta perturbación, lo cual es consistente con en análisis anterior.

Por último, al observar los resultados de la tabla de perturbaciones grandes, se puede observar que solo el Test Beta 1 y Stability Index reconocen cambios no significativos en la distribución de las variables, mientras que los demás test si lo hacen. El caso del Stability Index es consistente con los resultados de las tablas anteriores, mientras que el caso del test Beta 1 no lo es, dado que para una perturbación mediana a la variable edad éste había determinado que era significativa y ahora para una perturbación mayor determinó que no lo era. Esta inconsistencia puede deberse, como se explicó anteriormente, a la forma del estadístico del test Beta 1.

Notar que tanto el test KS como el Test chi-cuadrado detectan cambios en todas las variables del modelo. La explicación proviene de la definición de los estadísticos de contraste de estos test. Se observa que a medida que aumenta el tamaño de la muestra, las pruebas se hacen más sensibles (lo que se conoce también como la potencia de la prueba estadística), pues su varianza es una función inversa del número de observaciones; a mayor número, menos variabilidad y más sensibilidad. El efecto es que dichos test determinarán un cambio incluso si éste es insignificante para el modelo solo por el hecho de contar con un número elevado de observaciones (Tolvett, 2010).

Además, se realizó una segunda prueba con una base de datos considerablemente más pequeña para verificar el funcionamiento de las metodologías propuestas en condiciones diversas (ver Anexo D). Dado que se demostró anteriormente que el test ICV-1 está mal formulado, no ha sido probado con esta nueva base de datos. Los resultados obtenidos se resumen en las siguientes tablas:

Tabla 25: Comportamiento de los modelos dada una perturbación pequeña

	Test ICV-2	Test Beta 1	Stability Index	Test KS	Distancia de Hellinger	Test chi-cuadrado
Age	O	O	O	O		X
Shape	O	O	O		O	X

Tabla 26: Comportamiento de los modelos dada una perturbación mediana

	Test ICV-2	Test Beta 1	Stability Index	Test KS	Distancia de Hellinger	Test chi-cuadrado
Age	X	O	X	X		X
Shape	O	O	X		X	X

Tabla 27: Comportamiento de los modelos dada una perturbación grande

	Test ICV-2	Test Beta 1	Stability Index	Test KS	Distancia de Hellinger	Test chi-cuadrado
Age	X	X	X	X		X
Shape	O	O	X		X	X

Se observa que para perturbaciones pequeñas (tabla 25) el test ICV-2 se comporta de manera consistente con las metodologías ya conocidas, reconociendo que el cambio en la distribución, para ese nivel de perturbación, no es significativo. El único test que reconoce cambios significativos es chi-cuadrado, que es extremadamente sensible, como se explico anteriormente.

Para perturbaciones medianas y grandes (tablas 26 y 27 respectivamente) se observa que el test ICV-2 pierde consistencia, si se le compara con los resultados de los demás test, entregando conclusiones confusas. El modelo de regresión logística que dio origen a los intervalos de cambio máximo permitido es un buen modelo, lo que queda ratificado con los diversos indicadores que sirven para evaluar su desempeño, por lo tanto, se descarta que la definición de los intervalos sea la causa de las inconsistencias de ICV-2.

Ahora bien, para calcular el estadístico de ICV-2, se utiliza la media muestral y la varianza muestral, que dependen exclusivamente del tamaño de la muestra. Por lo tanto, mientras mayor sea el número de observaciones, menos distorsión habrá en los valores de dichos indicadores. Al ser este el caso de una muestra relativamente pequeña (en comparación con las usualmente utilizadas en data mining) es natural que el estadístico esté un poco distorsionado y por lo tanto un análisis en base a él puede ser peligroso ya que arrojaría resultados confusos o no del todo ciertos.

CAPÍTULO 7

CONCLUSIONES

7.1 SOBRE LOS RESULTADOS DEL TRABAJO REALIZADO

El principal objetivo de este trabajo de título era plantear una metodología para modelos de regresión logística capaz de detectar cambios significativos en la distribución de las variables, entendiendo que un cambio significativo es aquel que modifica la capacidad discriminante del modelo y que por tanto es peligroso y no se le debe ignorar. Cuando se identifican este tipo de cambios es necesario recalibrar el modelo para incorporar el efecto de las nuevas distribuciones de las variables y así garantizar un funcionamiento óptimo del modelo.

Existen variados modelos que cumplen este cometido, de los cuales se seleccionaron 5 que presentaban características similares a la metodología propuesta. De estos modelos, y posterior al análisis de perturbaciones se puede concluir lo siguiente.

Una de las metodologías propuesta en este trabajo, ICV-2 presentó resultados similares a los del Stability Index y la Distancia de Hellinger para perturbaciones pequeñas. Para perturbaciones medianas o grandes el test ICV-2 se asemeja en resultados a la Distancia de Hellinger, siendo entonces más sensible que el Stability Index, que aguanta hasta la perturbación 2 sin declarar que los cambios son significativos. Al observar la tabla de resultados correspondiente al Stability Index se puede ver que para la perturbación 2, el test identifica cambios menores en la distribución, por lo que la diferencia de sensibilidad entre este test e ICV-2 tampoco es tan abismante. De hecho, para esta perturbación el valor del estadístico de ICV-2 cae fuera del intervalo de confianza máximo por muy poco, lo que podría indicar que el test no es tan sensible como parece a simple vista, además que los resultados obtenidos en este análisis se restringen al modelo de regresión logística calibrado, si éste es deficiente, entonces el test ICV-2 no funcionará bien.

Es lógico que los resultados del Stability Index y de la distancia de Hellinger sean similares, ya que ambos entregan una medida escalar de la divergencia de la distribución de dos muestras. Ambos test son alternativas confiables para realizar seguimiento.

El test ICV-2, es una prueba de contraste de hipótesis y se diferencia de los indicadores anteriores porque entrega un resultado dicotómico, es decir, afirma o desmiente la hipótesis de que no hay cambios significativos entre dos muestras de datos. En ese sentido se asemeja más a los test kolmogorov-Smirnov y Chi-cuadrado, que también son pruebas de hipótesis.

Otra particularidad de este test es que determina un cambio por sobre un umbral, que queda determinado por el intervalo de confianza asociado al parámetro calibrado de la variable, lo que significa que el test captura una diferencia significativa solo si esta

impacta al modelo de tal manera que el parámetro calibrado deja de ser válido en las nuevas observaciones.

Por lo mismo, una de las desventajas de este test es que al depender de los parámetros de un modelo de regresión, requiere para su funcionamiento óptimo que éste sea muy bien calibrado, de lo contrario no funcionará bien. Además solo se puede aplicar para hacer seguimiento a modelos de regresión logística, lo cual también es una restricción. Esto no sucede con los modelos más genéricos, como Stability Index, Distancia de Hellinger, etc. que al ser construidos en base a proporciones de las observaciones de las muestras no se restringen a un caso particular de modelos.

Otra consideración que hay que tener es que para construir este test se utilizaron supuestos de normalidad en muestras grandes, lo que significa que su aplicación será efectiva para modelos con gran cantidad de registros, como se demostró en este trabajo. Al hacer pruebas con una base de datos más pequeña los resultados no fueron coherentes, por lo tanto se podría declarar que una de las restricciones de las metodologías propuestas es que requiere una gran cantidad de datos para su correcto funcionamiento.

Otra ventaja que tiene este test es que es fácil de implementar y es menos sensible que otras metodologías que existen, como el test Kolmogorov-Smirnov y el test Chi-cuadrado, que demostraron ser extremadamente sensibles a cualquier tipo de perturbación. Esto se debe a que la capacidad de detectar un cambio en la distribución depende del tamaño de la muestra. Si el número de observaciones supera las 10.000, entonces los test se vuelven demasiado sensibles, lo cual los hace poco confiables. Si se considera que la mayoría de las aplicaciones de modelos de regresión logística utilizan grandes bases de datos para su calibración, el uso de estos test no sería recomendable.

Con respecto al test Beta 1, el test ICV-2 parece ser menos sensible, lo cual es una mejora si se considera que este último es una corrección de Beta 1, que trata de incorporar el concepto de forma de la distribución al momento de hablar de cambios significativos. Se observaron algunas inconsistencias en los resultados del test Beta 1, pero esto se puede deber a que como utiliza de input los parámetros del modelo de regresión logística, depende mucho de que éste sea bueno.

El otro test que se planteó en esta memoria, ICV-1 no funcionó como se esperaba, la razón es que al definirlo como una diferencia en vez de una división y al ser el intervalo de cambio máximo una medida porcentual era imposible que los valores del test estuvieran dentro de los límites permitidos. Seguramente una redefinición del intervalo de cambio máximo permitido podría mejorar este problema.

En general se puede hablar de un buen cumplimiento de los objetivos de este trabajo de título, se plantearon dos metodologías de seguimiento para modelos de regresión logística, que posteriormente fueron aplicadas a un modelo de riesgo de crédito, que por lo demás presentó un buen rendimiento, en términos de su ajuste y de su capacidad de predicción, lo cual agrega confiabilidad a los resultados obtenidos por los test.

Se realizó en benchmark quedando conforme con los resultados, ya que uno de los modelos se comportó de manera similar a dos metodologías de seguimiento muy robustas: Stability Index y Distancia de Hellinger, lo cual es una muy buena señal, que indica que el modelo planteado es efectivo, al menos para un caso como este, donde se trabaja con muestras grandes.

7.2 CONSIDERACIONES PARA TRABAJOS FUTUROS

Una primera consideración es la formalización matemática de la distribución de los test planteados. Existen pocos trabajos que estudien la distribución del inverso del índice de variación, de hecho se encontró solo una publicación[4] y no estaba disponible para ser consultada. El realizar dicho análisis escapaba a los objetivos de esta memoria y es por ello que se decidió focalizar los esfuerzos en tareas más concretas, como la construcción de un modelo de regresión logística robusto. Sin duda formalizar el test sería un tremendo aporte en el ámbito de la estadística.

Respecto al test ICV-1 podría estudiarse una manera de redefinir el intervalo de cambio máximo permitido de manera que fuese comparable con el estadístico de esta metodología. Ya se comprobó que dada la definición que se le dio al intervalo de cambio máximo este test no funcionaba, pero eso no significa que la idea tras de é deba ser descartada.

Finalmente, una última línea de investigación futura sería la integración de un parámetro que de cuenta de los costos y beneficios de recalibrar el modelo. Con ello se podrá determinar el momento preciso en que un modelo que presenta cambios en la distribución de sus variables deberá ser recalibrado, correspondiendo al instante en que las pérdidas asociadas a los errores de clasificación sean mayores al costo asociado a la recalibración. Este instante no necesariamente coincidirá con lo que las metodologías de seguimiento indiquen, debido a que los criterios de evaluación son distintos, uno es estadístico y otro monetario. Estudiando la interacción entre estos criterios posible encontrar un valor óptimo que responda tanto a los requerimientos estadísticos como a los intereses económicos de la entidad en cuestión.

[4] Asymptotic sampling distribution of inverse coefficient-of-variation and its applications" de K. Sharma

BIBLIOGRAFÍA

[1] ALTMAN, E. AVERY, R. EISENBEIS, A. SINKEY, J. 1981. Application of Classification Techniques in Business, Banquing and Finances. Greenwich. Jai Press. pp. 2-5.

[2] BAESENS, B. CASTERMANS, G. MARTENS, D. HAMERS, B. VAN GESTEL, T. 2010. An Overview and Framework for PD Backtesting and Benchmarking. Journal of the Operational Research Society. pp. 5-10.

[3] BRAVO, C. MALDONADO, S. WEBER, R. 2009. Model Follow-Up in Logistic Regression Models. Chile, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas. 5p.

[4] BRAVO, C. MALDONADO, S. WEBER, R. 2010. Experiencias Prácticas en la Medición de Riesgo Crediticio de Microempresarios utilizando Modelos de Credit Scoring. Chile, Universidad de Chile, Facultad de ciencias Físicas y Matemáticas. pp 20.

[5] CIESLAK, D. CHAWLA, N. 2007. Detecting Fractures in Classifier Performance. En: Seventh IEEE International Conference on Data Mining. University of Notredame, Department of Computer Science and Engineering. pp. 123-132.

[6]CHITARRONI, H. 2002. La regresión logística. EN: Área empleo y Población, Instituto de Investigación en Ciencias Sociales. pp 12.

[7] DEPARTMENT OF TRADE AND INDUSTRY (DTI) 2004. Fair, Clear and Competitive: The Consumer Credit Market in the 21st Century. Inglaterra. Pp. 1-15.

[8] DROZDOWICZ, B. EVIN, D. HADAD, A. 2007. Modelo para el Tratamiento de Datos Desbalanceados basado en Redes Neuronales Autoorganizadas. Universidad Nacional de Entre Ríos, Facultad de Ingeniería. pp 1-3.

[9] FAYYAD, U. PIATETSKY-SHAPIRO, G. SMYTH, P. 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of The ACM. 39(11): 27-34.

[10] GNEDENKO, B. V. KOLMOGOROV, A. N. 1968. Limit Distributions for Sums of Independent Random Variables. Rusia, Addison-Wesly Publishing Company. pp298.

[11]GREENE, W. 2004. Basic Econometrics. 4°Edición. The McGraw-Hill Companies. Pp 1003.

[12] HOSMER, D. LEMESHOW, S. 2000. Applied Logistic Regression. 2ª ed. United States, Wiley. pp. 1-70

[13] LACOURLY, N. 2004. Estadística. Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas. Pp 1-53.

[14] LITTLE, R. RUBIN, D. 1987. Statistical Analysis with Missing Data. New York. John Wiley & Sons. pp 278

[15] NORVING, P. RUSSELL, S. 2010. Artificial Intelligence: A Modern Approach. 3ª ed. New Jersey, Prentice Hall. 22p.

[16] SIDDIQUI, N. 2005. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring, United States, Wiley. Pp. 1-30

[17] STEVENSON, M. 2008. An Introduction to Logistic Regression. En: EpiCentre, IVABS, Massey University. Pp. 21.

[18] THOMAS, L. EDELMAN, D. CROOK, J. 2002. Credit Scoring and It´s Application. United States, SIAM. Pp. 1-25

[19] THOMAS, L. 2009. Consumer Credit Models: Pricing, profit, and Portfolios. United States, Oxford University Press. Pp. 1-9.

[20] TOLVETT, CARLOS. 2010. Estudio de metodologías para el seguimiento de modelos de credit scoring utilizando regresión logística. Memoria para optar al título de Ingeniero Civil Industrial. Santiago. Universidad de Chile. Pp.109.

[21] UNIVERSIDAD DE BARCELONA, Pruebas para dos muestras independientes. [en línea] http://www.ub.edu/aplica_infor/spss/cap6-2.htm. [Febrero 2012]

[22] WESTERN HEMISPHERE CREDIT 6 LOAN REPORTING INITIATIVE 2007. Credit and Loan Reporting Systems in Chile. Pp. 1-10.

[23] WIKIPEDIA: THE FREE ENCICLOPEDIA. Cauchy Distribution. [en línea] http://en.wikipedia.org/wiki/Cauchy_distribution [Octubre 2011]

[24] WIKIPEDIA: THE FREE ENCICLOPEDIA. Coefficient of Variation. [en línea] http://en.wikipedia.org/wiki/Coefficient_of_variation [ENERO 2012]

[25] WIKIPEDIA: THE FREE ENCICLOPEDIA. Ratio Distribution. [en línea] http://en.wikipedia.org/wiki/Ratio_distribution [Octubre 2011]

[26] ZEIRA, G. LAST, M. MAIMON, O. 2005. Segmentation of Continous Data Streams Based on a Change Detection Methodology. EN: Advanced Techniques in Knowledge Discovery and Data Mining Springer. pp. 103-126.

ANEXO A

MATRIZ DE CORRELACIONES

Tabla A 1: Matriz de Correlaciones de las Variables del Modelo

	Target	UsoLínea	Edad	DebtRatio	Ingreso	Mora3060	Mora90	Dependientes	Mora6090	NumHip	OtrosProd
Target	1	0,238	-0,089	0,053	-0,044	0,211	0,307	0,039	0,235	-0,043	-0,12
UsoLínea	0,238	1	-0,229	0,137	-0,084	0,211	0,247	0,069	0,175	-0,118	-0,106
Edad	-0,089	-0,229	1	0,002	0,141	-0,055	-0,077	-0,166	-0,059	0,163	0,187
DebtRatio	0,053	0,137	0,002	1	-0,107	0,086	-0,028	0,085	0,025	0,547	0,278
Ingreso	-0,044	-0,084	0,141	-0,107	1	-0,006	-0,061	0,156	-0,036	0,312	0,200
Mora3060	0,211	0,211	-0,055	0,086	-0,006	1	0,206	0,053	0,228	0,008	0,054
Mora90	0,307	0,247	-0,077	-0,028	-0,061	0,206	1	0,027	0,260	-0,100	-0,074
Dependientes	0,039	0,069	-0,166	0,085	0,156	0,053	0,027	1	0,030	0,143	0,027
Mora6090	0,235	0,175	-0,059	0,025	-0,036	0,228	0,260	0,030	1	-0,043	-0,023
NumHip	-0,043	-0,118	0,163	0,547	0,312	0,008	-0,100	0,143	-0,043	1	0,336
OtrosProd	-0,012	-0,106	0,187	0,278	0,200	0,054	-0,074	0,027	-0,023	0,336	1

ANEXO B

RESULTADOS DE LA REGRESIÓN LOGÍSTICA FRENTE A DIVERSOS BALANCEOS

Se presentan a continuación los resultados obtenidos al calibrar modelos de regresión logística variando el modelo de balanceo de la base de datos.

B.1 BASE ORIGINAL

Se realizó una regresión con la base desbalanceada original para observar el comportamiento del modelo. Dicha proporción es de 93,8% de clientes buenos y 6,2% de clientes malos. Los resultados obtenidos son los siguientes:

Tabla B 1: Resumen del Modelo Base Original

Paso	-2log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	3,7726,631	0,096	0,257

Tabla B 2: Clasificación del Modelo Base Original

Observado		Pronosticado		
		Target		Porcentaje Correcto
		0	1	
Paso 1	0	96165	781	99,2
	1	5544	877	13,7
Porcentaje global				93,9

Al observar los resultados es claro que el modelo no alcanza a capturar las características predictivas de la clase menos numerosa y por lo tanto tiene un porcentaje de predicción bajísimo, lo cual significa que es necesario balancear la base.

B.2 OVERSAMPLING 4060

Corresponde a replicar los registros de la clase menos numerosa, considerando la clase más numerosa como el 60% de la base final. Por lo tanto los registros de clientes malos fueron replicados hasta alcanzar el 40% restante. Los resultados obtenidos son los siguientes:

Tabla B 3: Resumen del modelo Base Oversampling 4060

Paso	-2log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	157422,294	0,308	0,416

Tabla B 4: Clasificación del Modelo Base Oversampling 4060

Observado		Pronosticado		
		Target		Porcentaje Correcto
		0	1	
Paso 1	0	76833	20113	79,3
	1	17790	46420	72,3
Porcentaje global				76,5

B.3 UNDERSAMPLING 4060

Corresponde a seleccionar una muestra de los registros de la variable más numerosa, considerando que la variable menos numerosa tendrá una frecuencia de 40% en la base final. Los resultados obtenidos son los siguientes:

Tabla B 5: Resumen del modelo Base Undersampling 4060

Paso	-2log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	15520,726	0,316	0,427

Tabla B 6: Clasificación del Modelo Base Undersampling 4060

Observado		Pronosticado		
		Target		Porcentaje Correcto
		0	1	
Paso 1	0	7649	1982	79,4
	1	1757	4664	72,6
Porcentaje global				76,7

B.4 OVERSAMPLING 3070

Corresponde a replicar los registros de la clase menos numerosa, considerando la clase más numerosa como el 70% de la base final. Por lo tanto los registros de clientes malos fueron replicados hasta alcanzar el 30% restante. Los resultados obtenidos son los siguientes:

Tabla B 7: Resumen del modelo Base Oversampling 3070

Paso	-2log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	123797,670	0,280	0,397

Tabla B 8: Clasificación del Modelo Base Oversampling 3070

Observado		Pronosticado		
		Target		Porcentaje Correcto
		0	1	
Paso 1	0	76649	20297	79,1
	1	11430	30149	72,5
Porcentaje global				77,1

B.5 UNDERSAMPLING 3070

Corresponde a seleccionar una muestra de los registros de la variable más numerosa, considerando que la variable menos numerosa tendrá una frecuencia de 30% en la base final. Los resultados obtenidos son los siguientes:

Tabla B 9: Resumen del modelo BaseUndersampling 3070

Paso	-2log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	18931,918	0,286	0,406

Tabla B 10: Clasificación del Modelo Base Undersampling 3070

Observado		Pronosticado		
		Target		Porcentaje Correcto
		0	1	
Paso 1	0	11886	3097	79,3
	1	1761	4460	72,6
Porcentaje global				77,3

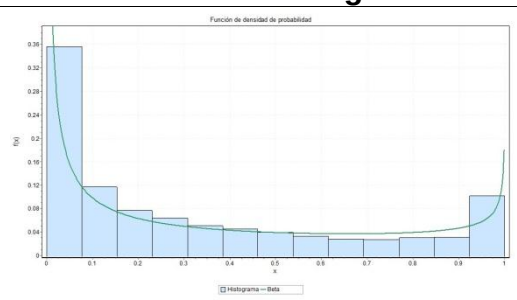
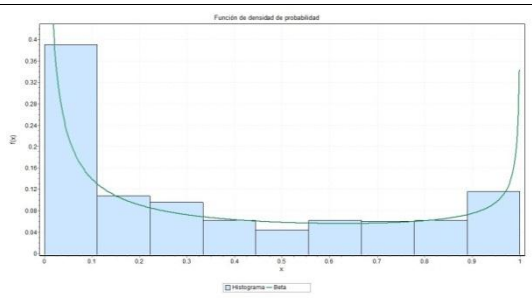
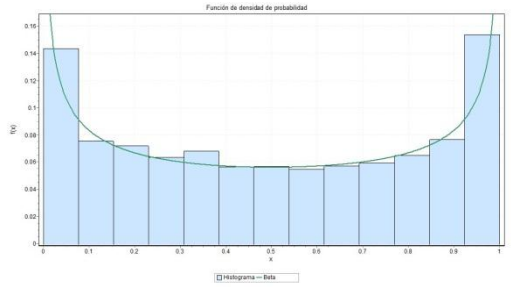
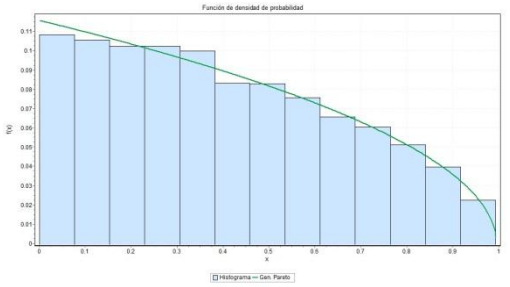
ANEXO C

GRÁFICO DE LAS PERTURBACIONES REALIZADAS A LAS VARIABLES

Se presentan a continuación las perturbaciones realizadas a las variables del modelo, de forma gráfica, para apreciar de mejor manera cómo va cambiando la distribución conforme se modifican sus parámetros.

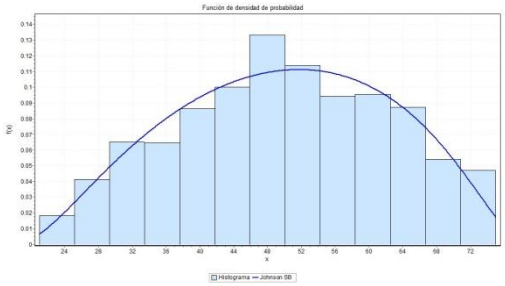
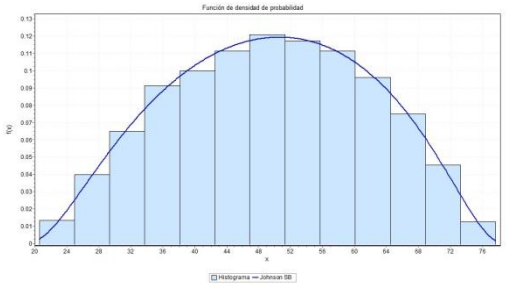
C.1 VARIABLE USOLÍNEA

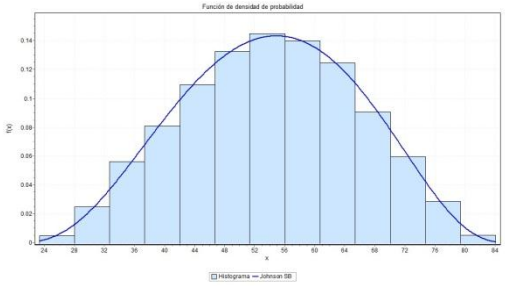
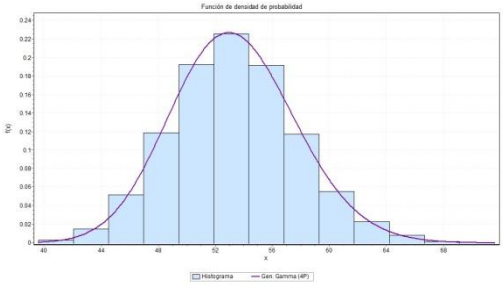
Ilustración C 1: Perturbaciones a la Variable UsoLínea

Distribución Original	Perturbación 1
	
Beta(0,30032 ; 0,63821)	Beta(0,33 ; 0,65)
Perturbación 2	**Perturbación 3**
	
Beta(0,6 ; 0,6)	Beta(1 ; 1,5)~Gen Pareto (-0,66 ; 0,66; 0)

C.2 VARIABLE EDAD

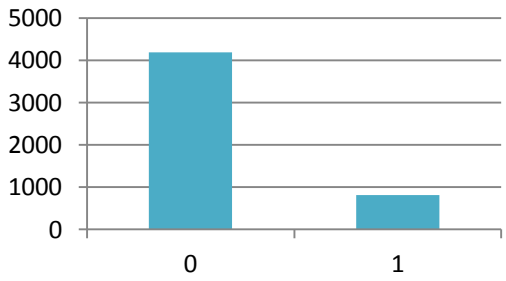
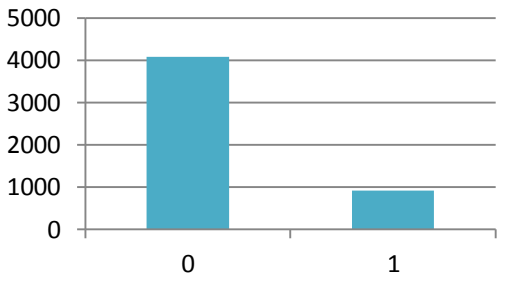
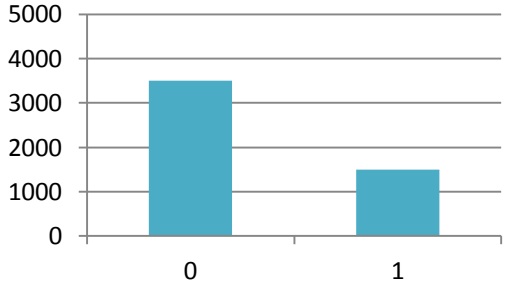
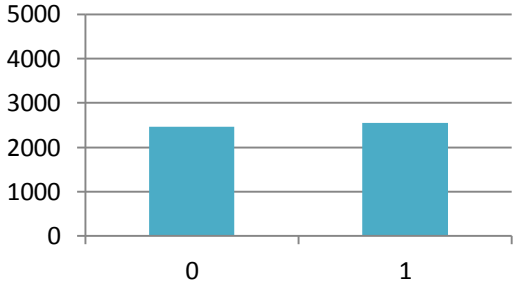
Ilustración C 2: Perturbaciones a la Variable Edad

Distribución Original	Perturbación 1
	
Johnson SB(-0,122 ; 1,044 ; 62,72 ; 16,95)	Johnson SB(-0,1 ; 1,1 ; 63 ; 17)

Perturbación 2	Perturbación 3
	
Johnson SB(0 ; 1,3 ; 68 ; 20)	Johnson SB(2;4;75;25)~Gen Gamma(1,1;79;1,03;8,4)

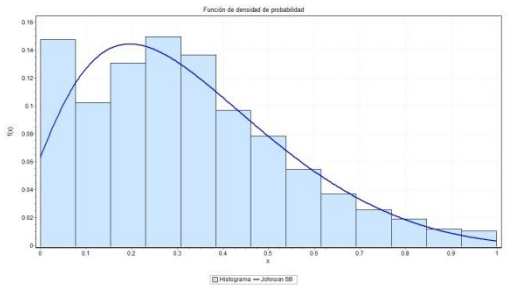
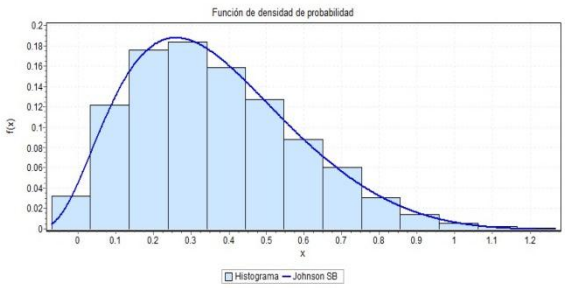
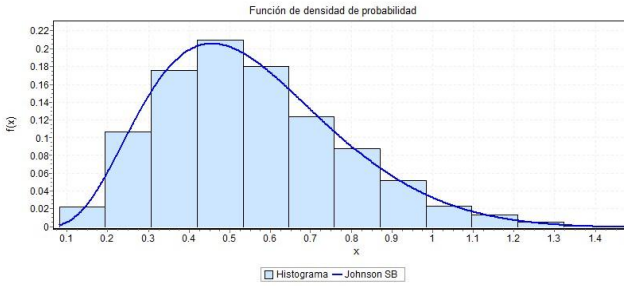
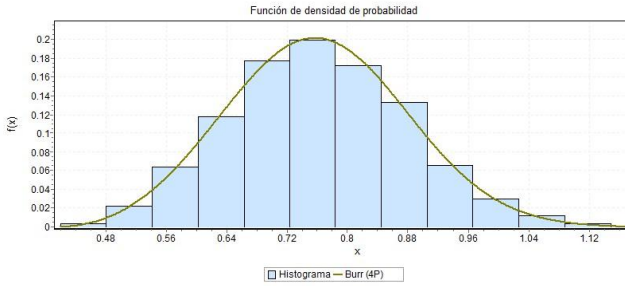
C.3 VARIABLE MORA3060

Ilustración C 3: Perturbaciones a la Variable Mora3060

Distribución Original	Perturbación 1
	
Bernoulli (0,162)	Bernoulli(0,18)
Perturbación 2	**Perturbación 3**
	
Bernoulli(0,3)	Bernoulli(0,5)

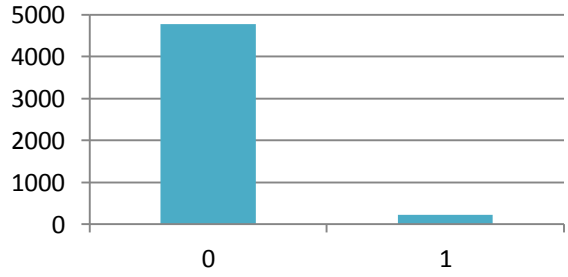
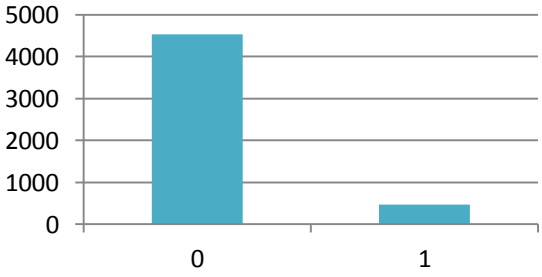
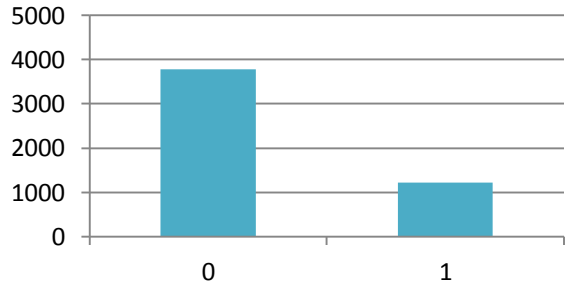
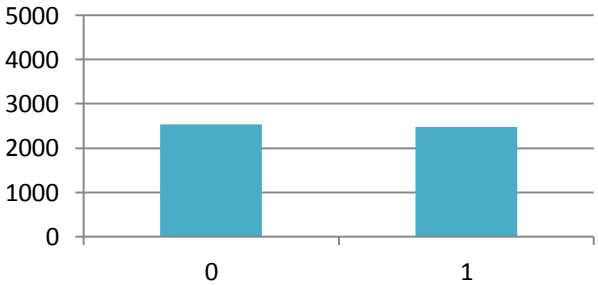
C.4 VARIABLE DEBTRATIO

Ilustración C 4: Perturbaciones a la Variable DebtRatio

Distribución Original	Perturbación 1
	
Johnson SB(1,2534; 1,322; 1,5; -0,133)	Johnson SB(1,3; 1,45; 1,6; -0,14)
Perturbación 2	**Perturbación 3**
	
Johnson SB(1,4; 1,6; 1,8; 0)	Johnson SB(2,4,2,0)~Burr 4P(11; 3,5; 0,77; 0,39)

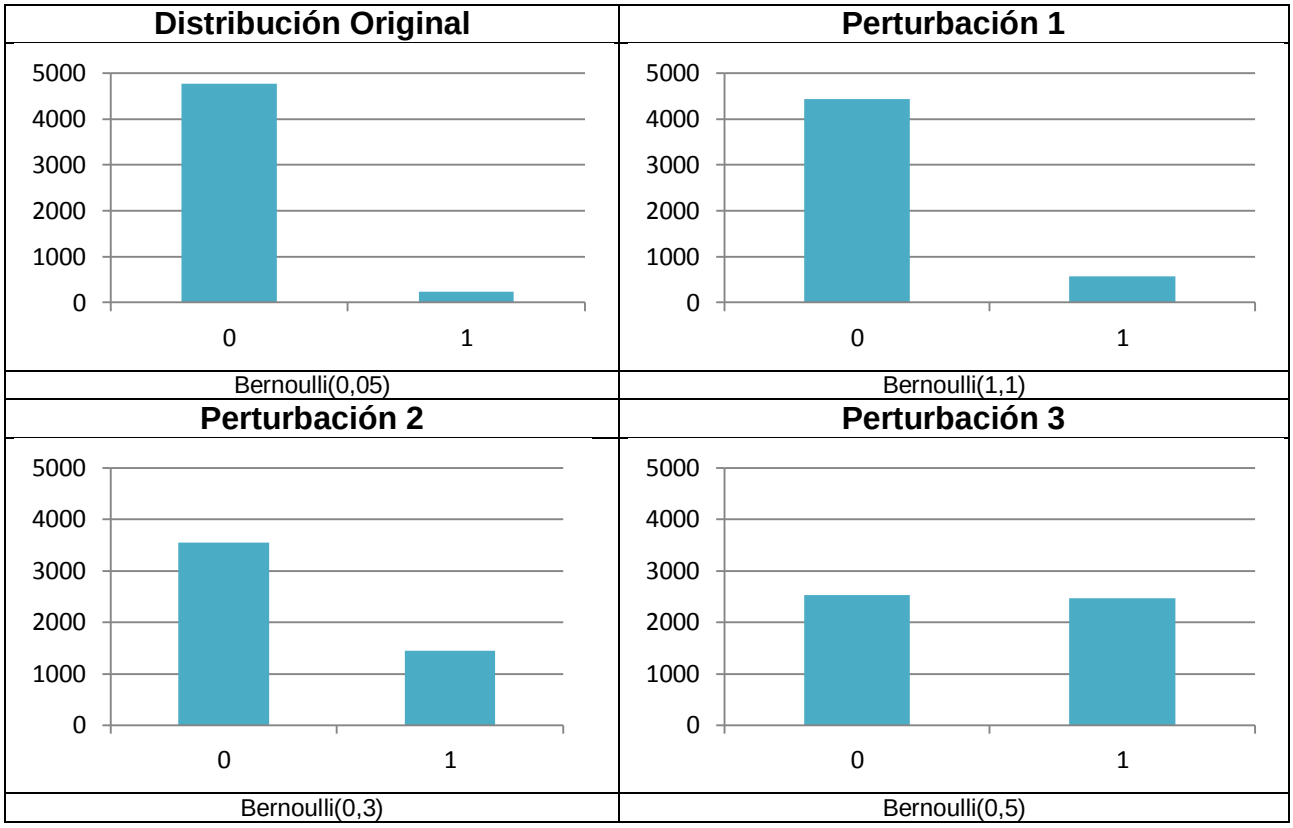
C.5 VARIABLE MORA6090

Ilustración C 5: Perturbaciones a la Variable Mora6090

Distribución Original	Perturbación 1
	
Bernoulli(0,047)	Bernoulli (0,1)
Perturbación 2	**Perturbación 3**
	
Bernoulli (0,25)	Bernoulli (0,5)

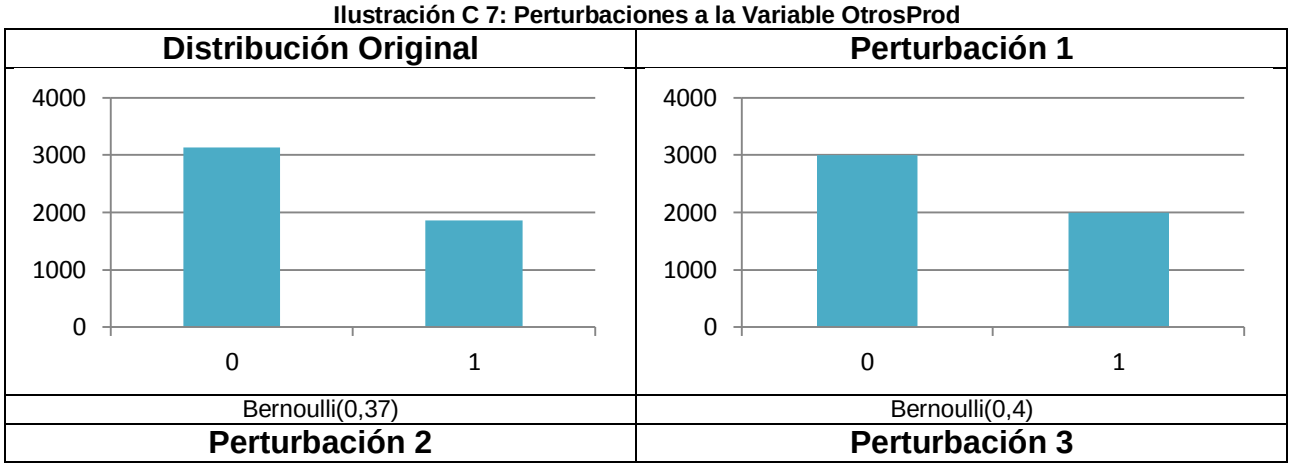
C.6 VARIABLE MORA 90

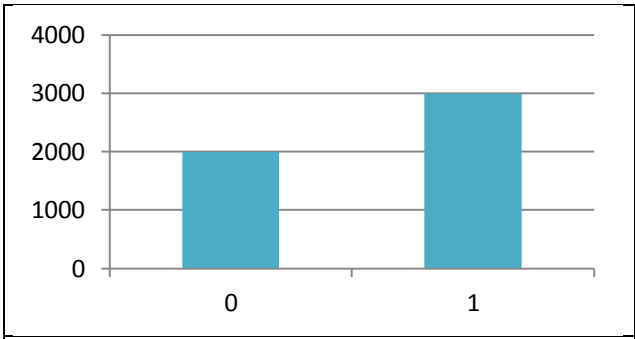
Ilustración C 6: Perturbaciones a la Variable Mora90



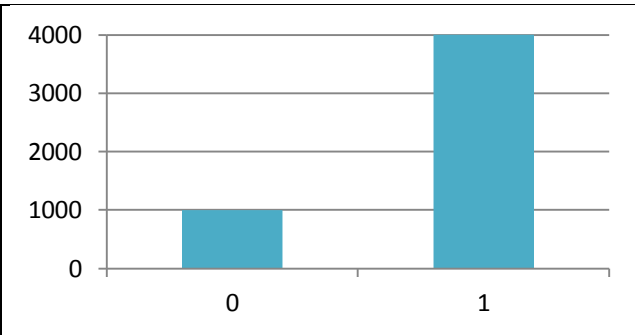
C.7 VARIABLE OTROSPROD

Ilustración C 7: Perturbaciones a la Variable OtrosProd





Bernoulli(0,6)



Bernoulli(0,8)

ANEXO D

APLICACIÓN DE LAS METODOLOGÍAS DE SEGUIMIENTO A UNA NUEVA BASE DE DATOS

Se presenta en el siguiente anexo la aplicación de las metodologías de seguimiento seleccionadas, en una nueva base de datos, con el objetivo de obtener un mejor análisis de las metodologías desarrolladas en este trabajo.

D.1 DESCRIPCIÓN GENERAL DE LA BASE DE DATOS

Los datos escogidos corresponden a información médica de una base de datos llamada "Mammographic masses*"* disponible en el sitio web *UCI Machine Learning Repository*[5], el cual es una colección de bases de datos que son usadas por la comunidad para diversos análisis.

La base de datos seleccionada cuenta con 961 registros. La variable dependiente es de carácter binario, donde el valor 1 indica que el tumor es maligno y el valor 0 indica que el tumor es benigno. Además se cuenta con 4 variables independientes, detalladas a continuación:

Tabla D 1: Descripción de las variables del modelo

Variable	Descripción	Tipo
Age	Edad de la paciente	Numérica
Shape	Forma del tumor	Categórica
Margin	Margen del tumor	Categórica
Density	Desnsidad del tumor	Categórica

D.2 CONSTRUCCIÓN DEL MODELO DE REGRESIÓN

D.2.1 Parámetros del Modelo

Luego de acondicionar la base de datos, el modelo fue calibrado en el software SPSS Statistics 19.0, utilizando los métodos *backward, forward y wrapper,* para determinar las variables más significativas. Todos los métodos convergieron al mismo modelo, cuyos parámetros calibrados fueron los siguientes:

[5] http://archive.ics.uci.edu

Tabla D 2: Variables en la ecuación

Variable	Tipo	Beta	E.T.	p-valor de Wald
Age	Numérica	0,054	0,008	0
Shape	Categórica			0
Categoría 1		-1,343	0,327	0
Categoría 2		-1,706	0,285	0
Categoría 3		-0,753	0,290	0,01
Margin	Categórica			0
Categoría 1		-1,933	0,367	0
Categoría 2		-0,268	0,573	0,639
Categoría 3		-0,816	0,327	0,013
Categoría 4		-0,435	0,284	0,126
Density	Categórica			0,4
Categoría 1		1,553	1,041	0,135
Categoría 2		0,796	0,865	0,357
Categoría 3		1,085	0,791	0,170

La Tabla anterior muestra para cada variable su tipo, el valor del parámetro calibrado, su error estándar y su significancia estadística en base al test de Wald, que contrasta la hipótesis de que el parámetro asociado a la variable tiene valor cero y el valor del parámetro calibrado.

Dado que ciertas variables tienen un p-valor que no permite rechazar la hipótesis nula del test de Wald y que por lo tanto tendrían un parámetro igual a 0, serán eliminadas del análisis de perturbaciones. Dichas variables son: Density (en todas su categorías) y Margin (categorías 2 y 4)

D.2.2 Ajuste del Modelo

Los principales indicadores del ajuste del modelo que se pueden obtener en SPSS se resumen en las siguientes tablas:

Tabla D 3: Resumen del modelo

Paso	-2log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	758,908	0,396	0,528

Tabla D 4: Matriz de confusión del Modelo

Observado	Pronosticado		
	Target		Porcentaje correcto
	0	1	
Target 0	348	98	78,0
Target 1	61	354	85,3
Porcentaje global			81,5

Con ello es posible calcular los errores del modelo de regresión:

- Error Global: 18,5%
- Error de tipo I: 22%
- Error de tipo 2: 14,7%

Tabla D 5: AUC

Área	Error Típico	Sig. Asintótica	Intervalo de confianza asintótico al 95%	
			Límite inferior	Límite superior
0,874	0,012	0	0,851	0,897

De lo anterior se puede concluir que el modelo es bueno, dado que su AUC es de 0,874.

D.3 APLICACIÓN DE LAS METODOLOGÍAS DE SEGUIMIENTO

D.3.1 Distribución de las variables

Utilizando el software Easy Fit 5.5 se ajustaron distribuciones de probabilidad conocidas a cada una de las variables del modelo. En la siguiente tabla se resumen los resultados obtenidos:

Tabla D 6: Distribución de probabilidad de las variables del modelo

Variable	Tipo	Distribución
Age	Numérica	Johnson SB(-0,85 ; 2,05 ; 130,18 ; -21,82)
Shape	Categórica	Multinomial (4 ; 0,23 ; 0,22 ; 0,1 ; 0,46)
Margin	Categórica	Multinomial (5; 0,38 ; 0,03 ; 0,13 ; 0,31 ; 0,15)
Density	Categórica	Multinomial (4; 0,01 ; 0,07 ; 0,91 ; 0,01)

D.3.2 Aplicación de las metodologías de seguimiento

En este apartado se realizará la aplicación de las metodologías de seguimiento para determinar cambios en la distribución de las variables escogidas y descritas anteriormente. Dado que se constató que el test ICV-1 presenta problemas debido a la definición del intervalo de cambio máximo permitido no será testeado.

Nuevamente la manera de probar las metodologías es mediante la perturbación de los parámetros que definen la distribución de cada variable, adicionando ruido. Para ello se han realizado tres perturbaciones.

Para los test Beta 1 e ICV-2 es necesario calcular el intervalo de cambio máximo permitido para cada variable en función de los parámetros encontrados al realizar la regresión logística y la desviación estándar de dicha estimación. Dichos intervalos se presentan en la siguiente tabla:

Tabla D 7: intervalo de cambio máximo permitido por variable

Variable	Tipo	Coeficiente	Desviación	Límite Inferior	Límite superior
Age	Numérica	0,054	0,008	0,703	1,296
Shape	Categórica				
Categoría 1		-1,343	0,327	-2,107	1,487
Categoría 2		-1,706	0,285	-3,986	1,334
Categoría 3		-0,753	0,290	-0,596	1,770

Test ICV-2

A cada variable se le realizaron 3 perturbaciones, una pequeña, una mediana y una grande. Los números presentados en la tabla corresponden al valor del estadístico para cada caso. En rojo se indican los valores del test que caen fuera del intervalo de cambio máximo permitido, lo que significa que se considera que el cambio en las distribuciones es significativo:

Tabla D 8: Resultados del test ICV-2

Variable	ICV-2 perturbación 1	ICV-2 perturbación 2	ICV-2 perturbación 3	Límite Inferior	Límite superior
Age	0,94	**0,61**	**0,34**	0,703	1,296
Shape					
Categoría 1	0,988	0,92	0,81	-2,107	1,487
Categoría 2	0,984	0,89	0,83	-3,986	1,334
Categoría 3	1,02	0,90	0,22	-0,596	1,770

Test Beta 1

A cada variable se le realizaron 3 perturbaciones, una pequeña, una mediana y una grande. Nuevamente, los números presentados en la tabla corresponden al valor del estadístico para cada caso y en rojo se indican los valores del test que caen fuera del intervalo de cambio máximo permitido, lo que significa que se considera que el cambio en las distribuciones es significativo.

Tabla D 9: Resultados del test Beta 1

Variable	ICV-2 perturbación 1	ICV-2 perturbación 2	ICV-2 perturbación 3	Límite Inferior	Límite superior
Age	0,97	0,78	**0,67**	0,703	1,296
Shape					
Categoría 1	0,92	0,76	0,6	-2,107	1,487
Categoría 2	0,95	0,73	0,62	-3,986	1,334
Categoría 3	0,83	0,05	0,045	-0,596	1,770

Stability Index

Los criterios para determinar si existe un cambio significativo en la distribución de las variables son:

- $SI < 0{,}1$ Ha ocurrido un cambio no significativo
- $0{,}1 < SI < 0{,}25$ Ha ocurrido un cambio menor
- $SI > 0{,}25$ Ha ocurrido un cambio significativo

Frente a las perturbaciones realizadas los resultados obtenidos fueron los siguientes:

Tabla D 10: Resultados del test Stability Index

Variable	SI perturbación 1	SI perturbación 2	SI perturbación 3
Age	0,007	**0,84**	**2,31**
Shape	0,01	**0,33**	**1,14**

En rojo se indican los valores del Stability Index que determinan que el cambio sufrido en la distribución de las variables debido a la perturbación es significativo, mientras que en azul se indican aquellas perturbaciones que producen un cambio menor, de acuerdo a la definición de este índice.

Test Kolmogorov-Smirnov

Se aplicó el test a la variable continua del modelo, es decir, a Age, frente a las tres perturbaciones realizadas, calculando el valor del siguiente estadístico:

Tabla D 11: Estadísticos de prueba del Test KS

Variable	Perturbación 1	Perturbación 2	Perturbación 3
Age	0,041	**0,422**	**0,56**

El valor teórico de $D_i = F_1(X_i) - F_2(X_i)$ para estas muestras que contienen 861 registros es de 0,046, por lo tanto, la regla de decisión es la siguiente:

- Si Valor del estadístico < 0,046 se acepta la hipótesis nula, es decir, no hay divergencia entre las dos muestras
- Si Valor del estadístico > 0,046 se rechaza la hipótesis nula, es decir, hay divergencia entre las muestras.

Test Chi-cuadrado

El estadístico de prueba del test chi cuadrado se calcula como:

$$\chi_p^2 = n_K \sum_{i=1}^{j} \frac{\left(\frac{x_{i,K}}{n_K} - \frac{x_{i,K-1}}{n_{K-1}}\right)^2}{\frac{x_{i,K-1}}{n_{K-1}}}$$

Este estadístico sigue una distribución χ^2 de $j-1$ grados de libertad. Por lo tanto, si el valor del estadístico χ_{j-1}^2 es mayor que el valor crítico de la distribución χ^2, entonces se rechazará la hipótesis nula, es decir, la variable no es estacionaria entre períodos y si presenta cambios en su distribución. Los resultados obtenidos al aplicar este test se presentan en la siguiente tabla:

Tabla D 12: Resultados del test chi-cuadrado

Variable	Valor del Estadístico			Valor Crítico
	Perturbación1	Perturbación2	Perturbación3	
Age	**6,19**	**633**	**1114**	3,81
Shape	**12,88**	**262**	**598**	7,81

Notar que para todas las perturbaciones realizadas el valor del estadístico de prueba supera al valor crítico de la distribución chi-cuadrado, por lo tanto se rechaza la hipótesis nula en todos los casos y se concluye que la distribución de las variables ha cambiado al someterlas a las perturbaciones.

Distancia de Hellinger

La distancia de Hellinger es un indicador de la divergencia entre la distribución de dos variables categóricas. Sean X e Y dos poblaciones con p categorías, la distancia de Hellinger entre ellas se calcula como:

$$Hellinger\,(X,Y) = \sqrt{\sum_{j=1}^{p}\left(\sqrt{\frac{X_j}{n_X}} - \sqrt{\frac{Y_j}{n_Y}}\right)^2}$$

Este indicador alcanza un valor mínimo 0, lo que significa que las distribuciones son idénticas y un valor máximo de $\sqrt{2}$, en cuyo caso las distribuciones son totalmente divergentes.

Se aplicó el test a la variable categórica del modelo, es decir a Shape, frente a las tres perturbaciones realizadas, obteniendo los siguientes resultados:

Tabla D 13: Resultado del Test de Distancia de Hellinger

Variable	Perturbación 1	Perturbación 2	Perturbación 3
Shape	0,02	**0,28**	**0,53**

Considerando los valores límite de la distancia de Hellinger, se considera que no hay cambios significativos en la distribución de las variables tras la perturbación 1 y hay cambios menores en la distribución tras la perturbación 2 y 3.