UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE ECONOMIA

“GENERACION DE SCORECARDS, CONCEPTOS BASICOS Y PUNTOS DE VALIDACION”

TESINA
QUE PARA OBTENER EL TITULO DE
LICENCIADO EN ECONOMIA
P R E S E N T A :
EDUARDO ALEJANDRO LOPEZ FLORES

ASESOR DE TESINA:
MTRO. OMAR CONTRERAS CLEOFAS

MEXICO, CIUDAD DE MEXICO, MAYO DE 2016

Índice de contenido

Índice de tablas

Índice de gráficas

1.1. Introducción

El presente trabajo muestra de forma general la metodología que se tiene para la definición de los factores de evaluación de un nuevo crédito para las instituciones financieras, en particular, los bancos.

Contiene un desarrollo sobre la base teórica que se debe tomar en cuenta al momento de realizar los ejercicios correspondientes para las scorecards, con el fin de que el desarrollo que se realice cumpla con su objetivo, el cual es poder discriminar los casos buenos de los malos.

Las scorecards son comunes en la industria financiera, las cuales son ocupadas para gestionar el riesgo crediticio que puedan tener las nuevas originaciones de crédito. Existen diferentes tipos de modelos que se pueden ocupar para generar dichas herramientas, los cuales pueden ser redes neuronales, árboles de decisión y modelo Logit. En este trabajo se ocupara el modelo Logit.

Una vez finalizada la teoría de la construcción de las scorecards se pasará a analizar los puntos que se consideran importantes en el desarrollo de la herramienta, dando una breve explicación sobre el por qué se debe poner atención en dichos puntos.

1.2. Justificación

Las scorecards son modelos que se utilizan en la industria bancaria que ayudan en la gestión del riesgo, los cuales tienen como objetivo el diferenciar a los clientes buenos de los malos al momento de otorgar un crédito.

Para llevar a cabo estas técnicas se pueden requerir metodologías desde lo más complejas (como son la generación de clusters, optimización de la transformación de viables basándose en medidas de discriminación, la estimación de una red neuronal), partiendo de lo más básico a la exploración de las variables con base en la variable objetivo, así como el uso del modelo Logit.

Tomando en cuenta lo anterior, el primer paso para la generación de una scorecard se requiere de conocimientos teóricos de los elementos utilizados en el desarrollo.

1.3. Objetivo

El objetivo general es dar una introducción sobre lo qué son las scorecards, las cuales pueden ser ocupadas como un elemento base para la decisión de la otorgación de un crédito.

El objetivo particular es mostrar las técnicas básicas que se ocupan para la generación de una scorecard, las cuáles consisten en la generación de pruebas estadísticas sobre los datos y sobre el resultado final.

1.4. Historia de los credit scoring

Como se menciona en los primeros capítulos del libro <<*Credit Scoring and its applications*>> de Lyn C. Thomas y compañía, la historia de los credit scorign es de aproximadamente 60 años. En si un credit scoring es un apoyo que sirve para identificar diferentes grupos en una población cuando no se puede ver una característica definida de cada grupo, es decir, se ocupa el conjunto de variables para realizar dicha diferenciación.

A lo largo de los años se han tenido diferentes modificaciones sobre la metodología utilizada para su estimación. La primera persona que utilizó este concepto de realizar la separación de grupos de una población con un enfoque estadístico fue Ronald Fisher en 1936.

La principal aplicación práctica de los credit scoring se presentó en al año 1930, cuando algunas compañías las utilizaron para comparar los resultados que se tenían en la otorgación de créditos por parte de los analistas financieros. Fue en 1960 cuando algunos bancos y otros otorgadores de créditos se dieron cuenta de la utilidad de los credit scoring. Sin embargo, se tenía el problema de su aplicación, el incremento de solicitudes generaba problemas para la capacidad de análisis que se necesitaba para la toma de decisiones, lo cual se solucionó con el desarrollo de la informática y la computación, es decir las nuevas tecnologías ayudaron a expandir la capacidad de análisis y por tanto de decisión.

Con estos avances se tuvieron mejoras en las tasas de moratoria de las instituciones crediticias, por lo cual fueron usadas cada vez por más compañías. En 1980 se introdujeron técnicas como la regresión logística y la programación lineal, hoy en día lo más actual es el uso de técnicas de inteligencia artificial como lo son las denominadas redes neuronales.

2. Scorecards

Una scorecard es una herramienta que ocupan instituciones financieras para evaluar a los posibles nuevos clientes o clientes actuales y tomar la decisión de otorgar el crédito solicitado o no, para lo cual se basan en la probabilidad de incumplimiento que se estima.

Las herramientas de calificación se pueden aplicar para créditos minoristas y mayoristas. Este trabajo se enfocará en lo que se denominan créditos minoristas, los cuales son créditos que se le otorgan a algún particular, autónomo o pequeño negocio. En general, los créditos que se consideran como minoristas son los conocidos como crédito al consumo, tarjeta de crédito, crédito nómina, hipotecas y créditos Pymes.

2.1. Tipos de scorecards

Cuando una persona piensa en solicitar un crédito con un banco se inicia lo que se conoce como fase de ciclo de riesgo, el cual es el tiempo de vida que tiene el crédito dentro de la institución financiera. Las fases que puede tener el crédito son las siguientes:

- Admisión – Reactivo
- Seguimiento – Comportamentales y proactivas
- Impagados – Cobranza
- Mora – Recuperación

Dependiendo de en qué fase se encuentre el crédito, la institución financiera puede ocupar diferentes tipos de scorecards o herramientas. Durante cada fase del ciclo de riesgo se toman decisiones, por lo cual, en cada fase se ocupan diferentes herramientas.

La construcción de una scorecard es de una estructura homogénea, sin importar si la herramienta es usada para créditos minoristas o mayoristas, lo mismo sucede para las herramientas que se ocupan en diferentes fases del ciclo crediticio, la principal diferencia que se tiene entre estos tipos de herramientas es el tipo de variables que se ocupan, así como la definición de la variable objetivo.

2.2. Fases de construcción de Score Card

Existen distintas bibliografías donde se mencionan las diferentes fases que se deben de llevar a cabo para la generación de una scorecard, a lo largo de este trabajo se ocuparan los libros <<*Credit Risk Scorecards, Developing and Implementing Intelligent Credit Scoring*>> de Naeem Siddiqi, y el libro <<*Credit Scoring and its application*>> de Lyn C. Thomas y compañía, en donde se desarrollan las siguientes fases:

1) Definición del Universo
2) Segmentación
3) Tratamiento de los datos
4) Estimación del Modelo
5) Validación

A continuación se definirán cada uno de estos puntos.

2.2.1. Definición del Universo

Para definir el universo con el cual se va a trabajar se debe considerar lo siguiente:

- ***Definir la población objetivo***

Cuando se genera un análisis como el del scorecard se debe asegurar que el modelo esté dirigido a evaluar a la población en cuestión, donde se tiene que identificar si existen subpoblaciones a las cuales se debe diferenciar y no englobar en una sola herramienta. En el caso de que existan estos grupos diferenciados se tiene que generar una scorecard por cada subpoblación.

- ***Definición de la estructura de la base de datos***

Aquí se define el tipo de modelo que se va a generar, se tiene que elegir si va a ser a nivel cliente o a nivel crédito, además se tiene que tomar en cuenta si existen efectos de estacionalidad o de volatilidad de las posibles variables a utilizar en el desarrollo.

En el desarrollo de las scorecards se utiliza el supuesto de que el comportamiento futuro está reflejado por el comportamiento del pasado, es decir, el análisis que se realiza es determinar qué tan probable es que una evaluación sea buena o mala, dependiendo de la definición del objetivo a evaluar. Para alcanzar el objetivo antes planteado se debe tomar una ventana de tiempo y monitorear el comportamiento de estas operaciones en otro tiempo para definir si son buenos o malos, para lo cual se tienen que agrupar dentro de una tabla de datos las variables características que se espera sean útiles en el análisis de la variable objetivo (la cual refleja el evento que se pretende analizar).

Considerando lo anterior, se pueden definir tres puntos importantes a considerar:

- el tiempo de muestra,
- el periodo de comportamiento y
- el tiempo de evaluación.

El primero punto hace referencia a la acumulación de todas las variables de las operaciones que se han tenido como históricos, teniendo en consideración que el tiempo de la información no se debe prolongar, con el objetivo de que las variables tengan algún efecto de estacionalidad o distorsión de los datos. El periodo de comportamiento es el tiempo en el cual se le da un seguimiento, mientras que en el tiempo de evaluación se define si la operación es considerada como buena o como mala.

- ***Definir las fuentes de información***

Es indispensable contar con algún repositorio de la información donde se pueda obtener la mayor cantidad de datos de los clientes u operaciones, ya que dependiendo de la fuente los datos se pueden ver afectados los análisis, así como de las conclusiones del proyecto.

- ***Definir la variable a predecir***

La variable a predecir es indispensable en el modelo, para lo cual se deben tomar en cuenta que la definición de la variable dependiente puede variar en razón del análisis del modelo y de las fuentes de información que se tengan disponibles.

Por ejemplo, para efectos de este trabajo se toma en cuenta que se pretende predecir el incumplimiento, es decir, que un cliente u operación no cumpla con los pagos que se acordaron al momento de la otorgación del crédito, dado a que la definición de un incumplido puede ser muy amplia, lo que se puede ocupar como respaldo es algún tipo de estudio.

En el caso de las instituciones financieras el apoyo que ocupa para la definición del incumplimiento es el Acuerdo de Basilea II, donde se menciona que se considera como vencida una operación o cliente en el momento en que esta tenga 90 días de impago.

Se puede partir del supuesto de que existe un estudio en el cual se considera que se es más restrictivo en el análisis del incumplimiento de las operaciones si en lugar de considerar los 90 días se toman solo 70 días. El analista deberá decidir sobre cuál fuente de información debe definir su variable objetivo.

- ***Análisis de distribución inicial y análisis de madurez***

Se debe contemplar la distribución que tienen las variables a utilizar, así como la madurez de la información con la que se cuenta. También se tiene que considerar que entre mayor información se tenga para la generación del modelo se puede tener mayor confianza en las pruebas estadísticas de estabilidad de los datos.

Cuando se habla de madurez de la información se hace referencia a que se debe tener un periodo tal que sea suficiente para generar la definición de la variable objetivo.

- ***Seleccionar una muestra de modelización***

Al momento de generar un modelo se toma como una práctica común la generación de dos muestras, una de desarrollo y otra de validación. Por lo general, la muestra de desarrollo está compuesta por el 80% de la información disponible, esto en el caso de tener una buena cantidad de información, en la práctica se pueden considerar mil casos como una cantidad aceptable; mientras que la muestra de validación es el 20% restante, la cual va a servir para ratificar los resultados que se obtengan a partir de la primera.

En ciertos ejercicios se puede tener el escenario en el cual no se cuenta con suficiente cantidad de información, por lo cual no es recomendable dividir la información en el 80% de desarrollo y el 20% de validación, es decir, se debe ocupar toda la información en la parte de desarrollo, así como en la parte de validación.

2.2.2. Segmentación del Universo

Una vez definido el universo se deben realizar análisis para determinar si existen grupos dentro de este universo que valgan la pena separarlos ya que se comportan de forma diferente entre ellos, sin embargo, esto se debe realizar para aquellos casos en donde el tamaño del universo lo permita. Es decir, lo que se busca es detectar los atributos o características que optimicen la diferenciación de cada uno de los grupos.

Existen tres tipos de criterios que se pueden ocupar para generar estas agrupaciones.

- ***Tipología de información disponible***

Es conveniente diferenciar a las agrupaciones según a la información ya que esta puede ser diferente entre ellos, siempre y cuando la diferenciación sea relevante y pueda ayudar a determinar el perfil de riesgo de la operación.

- ***Cambios estructurales***

Es el grupo de operaciones que presenten un comportamiento diferenciado ante el evento que se está analizando.

- ***Criterios de negocio***

En ocasiones el negocio es quien determina la forma en que se deben llevar a cabo las segmentaciones, a este tipo de acciones que se hacen sin tomar en cuenta un análisis en la distribución de las variables se denomina criterio experto. En estos casos se requiere realizar un análisis sobre las segmentaciones finales con el fin de corroborar que el efecto de esto sea el buscado por el negocio.

Para llevar a cabo la segmentación del universo se debe tomar en cuenta la cantidad de observaciones que se tienen en el universo, en el caso de tener suficiente información, la segmentación no debe tener complicación. Sin embargo, se debe tomar en cuenta que aunque el universo tiene los casos suficientes, los grupos que se están segmentando no deben tener poca población ya que esto puede generar que los resultados encontrados para estos últimos no sean confiables al momento de ocuparlos en la práctica.

2.2.3. Tratamiento de los datos

En los puntos anteriores se mencionó la importancia de la cantidad de información que se debe tener disponible para el desarrollo de la herramienta, sin embargo, existe otro punto a considerar, la cual es la calidad de la información.

Este apartado es igual de importante que el de la generación del universo, ya que en este se podrá identificar el tipo de información con la que se estará trabajando en pasos posteriores, así mismo se podrán detectar los posibles errores y falta de información.

Esta revisión se hace a nivel variable, con el fin de determinar qué tipo de acción se realizará en el caso de encontrar algún tipo de error en la información. Dicho análisis contiene las siguientes fases:

- ***Análisis descriptivo***

Se buscará la existencia de valores vacíos y determinar el porcentaje de estos con respecto al total, identificar si existe una elevada concentración en ciertos valores de la variable, así como de errores de información.

- ***Análisis temporal***

Se realiza un análisis sobre las variables con respecto al tiempo, para identificar si existen cambios en la distribución de la variable.

- ***Coherencia en la información***

Al momento de realizar la distribución de las variables se tiene que identificar si la distribución que se tiene es coherente con lo que se espera de la información. En el caso de no encontrar esta coherencia con la variable objetivo, se debe descartar del desarrollo ya que puede generar resultados no acordes a lo buscado.

- ***Depuración de datos / Creación de nuevas variables***

De los análisis antes mencionados se debe tomar una decisión sobre qué hacer con los casos vacíos, es decir, para aquellos datos que no tengan una distribución estable y que tampoco tienen una coherencia con la variable objetivo.

Una de las técnicas que se puede ocupar es la eliminación de los casos que presenten estas características, en otros, como es el caso de los casos vacíos, se les puede imputar valores muy extremos para ser incorporados en la distribución. Esta decisión la tiene que tomar el analista al momento de la selección de las variables durante el desarrollo de la scorecard.

Una vez que se determinó que hacer en el proceso de revisión de información, se puede analizar la posibilidad de generar nuevas variables a partir de las que ya se cuenta, esto con el fin de ampliar la cantidad de estas, las cuales puedan ser usadas para definir la variable objetivo.

2.2.4. Estimación de los modelos

Una vez definido el universo, las posibles variables a ocupar y la variable objetivo, el último paso consiste en seleccionar las variables que van a ser usadas en el desarrollo final del modelo.

2.2.4.1. Análisis bivariante

Este se puede generar para todas las variables que se pretenden usar dentro del modelo, el análisis se debe realizar usando diferentes métricas para determinar si son usadas o no para el desarrollo.

- ***Análisis gráfico***

En el apartado de tratamiento de datos se mencionó la coherencia de la información, en el cual solo se analizó la distribución que puede llegar a tener cada variable, en esta ocasión lo que se analizará es si la **Tasa de Mora (TM)** de los datos de cada una de las variables es consistente con lo que se busca explicar.

Para generar este análisis gráfico se debe generar la TM, su fórmula es la siguiente:

$$TM = \frac{Casos\ malos_i}{Total\ de\ casos_i} \quad (1)$$

Dónde:

Casos malos = Cantidad de casos malos (en base a la variable objetivo) en el grupo i de la variable

Total de casos = Casos totales de la variable del grupo i

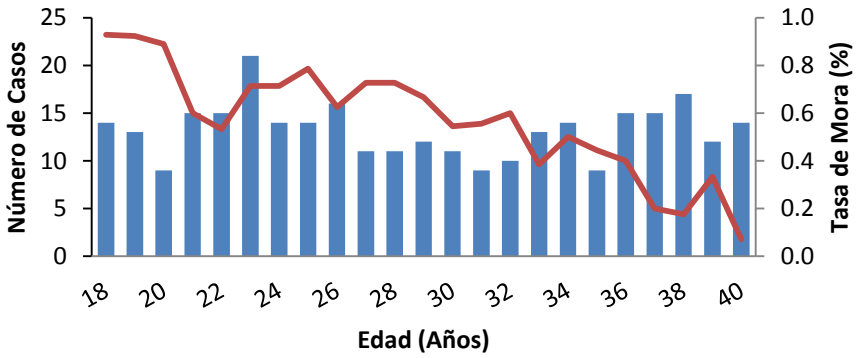
De los valores obtenidos en la TM se puede concluir lo siguiente:

- Entre más alto es el valor de TM más casos malos se encuentran en este grupo, por lo cual ese grupo tiene una mayor probabilidad de ser malo.
- Entre más bajo es el valor de la TM menos casos malos se tienen, por lo cual ese grupo tiene una menor probabilidad de ser malo.

Para ejemplificar lo anterior, en una hoja de Excel se generó una distribución aleatoria de números que tienen valores de entre 18 y 50 años, simulando ser la edad de los clientes de un banco, donde al igual se le define de forma aleatoria la variable de incumplimiento. Por otra parte, supongamos que se tiene un estudio que determina que las personas con menor edad son más susceptibles a incumplir los pagos de un crédito, por lo cual entre menor es la edad mayor se puede esperar que sean malos pagadores.

La siguiente gráfica representa esta situación, donde se muestra el análisis gráfico de lo que se espera encontrar con la TM de la variable.

Gráfica 1. Distribución TM de la variable edad simulada



Fuente: Elaboración propia

La distribución de la variable no tiene problemas ya que se cuenta con una buena cantidad de casos en cada uno de sus valores, por otra parte la tasa de mora tiene una distribución que se esperaba. En el caso de que la TM tuviera una tendencia inversa se consideraría que la variable no tiene la coherencia con respecto a la variable objetivo, por lo cual se debe descartar del modelo final.

Otro análisis gráfico es el de la generación del Peso de la Evidencia (en inglés Weight of Evidence [WOE]), la forma de estimarlo es mediante la siguiente fórmula:

$$WOE = \ln\left(\frac{\%buenos_i}{\%malos_i}\right) \quad (2)$$

Dónde:

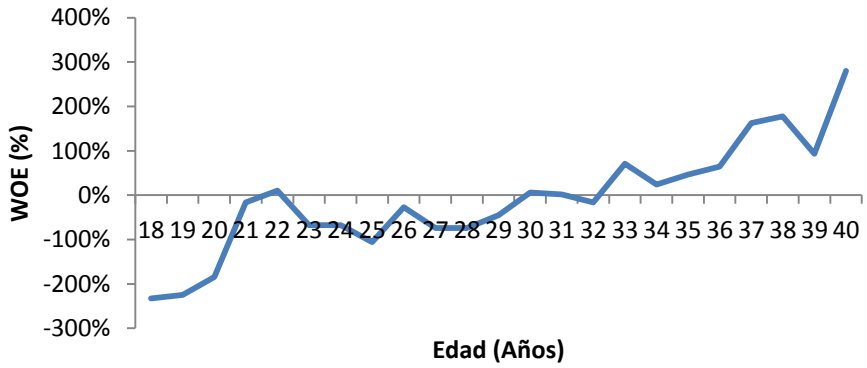
% buenos = cantidad de casos buenos del grupo i de la variable entre la cantidad de casos buenos de la variable.

% malos = cantidad de casos malos del grupo i de la variable entre la cantidad de casos malos de la variable.

El resultado que se tiene con el WOE es inverso en comparación con la TM, es decir, entre menor sea el número obtenido mayor incumplimiento encontraremos en esos valores de la variable, y viceversa, cuando el número obtenido es más grande se presentaran menores casos de incumplimiento.

Para efectos de la generación del modelo final, la tendencia que se obtenga del WOE debe ser continua, ya sea con pendiente positiva o negativa, sin embargo, la gráfica debe tener coherencia con lo esperado de dicha variable tomando como eje la variable objetivo.

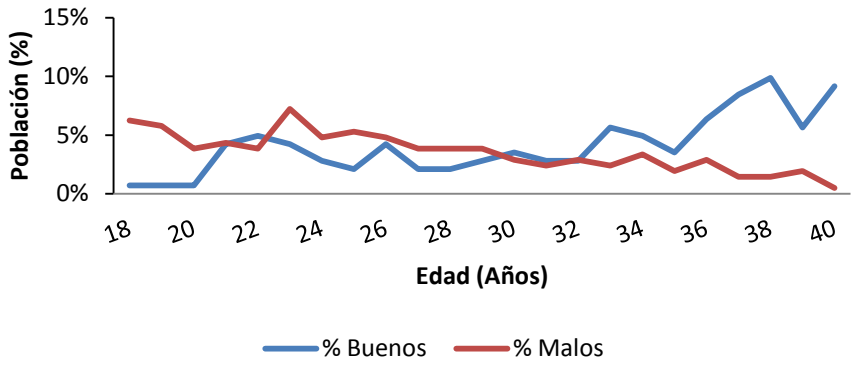
Gráfica 2. Distribución WOE de la variable edad simulada



Fuente: Elaboración propia

Otra prueba que se puede utilizar para determinar qué variable elegir es generando un análisis gráfico donde se comparan las distribuciones de los porcentajes de casos buenos y malos, con esto se busca encontrar distribuciones que se separen entre sí. Usando los mismos datos del ejemplo anterior se obtiene la siguiente distribución:

Gráfica 3. Distribución casos buenos y malos de la variable edad simulada

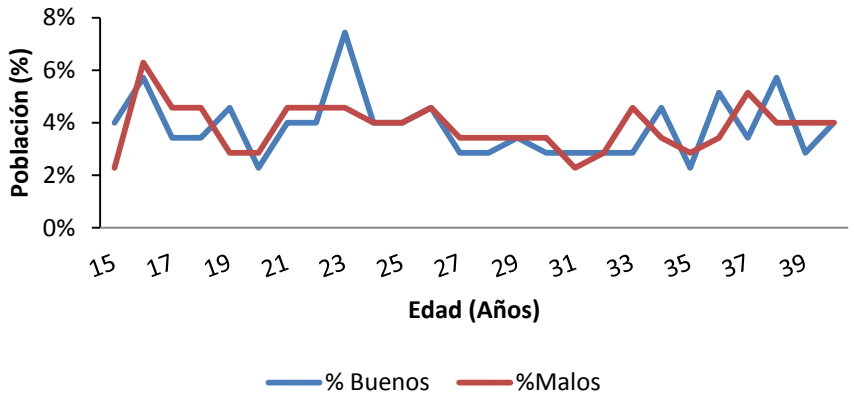


Fuente: Elaboración propia

La distribución que se tiene de casos buenos y casos malos es distinta, es decir, se puede sospechar que el comportamiento de los casos malos esté relacionado con la edad que pueda tener el sujeto a evaluación.

La siguiente gráfica muestra el caso en el cual la distribución de los casos buenos y malos son semejantes. Si el comportamiento de las distribuciones no muestra diferencias, se puede tomar como sospecha que la variable analizada puede ser no significativa en el desarrollo del modelo.

Gráfica 4. Distribución casos buenos y malos de la variable edad simulada



Fuente: Elaboración propia

- ***Indicadores de capacidad discriminante***

Se pueden ocupar distintas métricas para llevar a cabo un análisis de capacidad discriminante, en este trabajo se mencionarán tres:

- ***Índice de Gini***

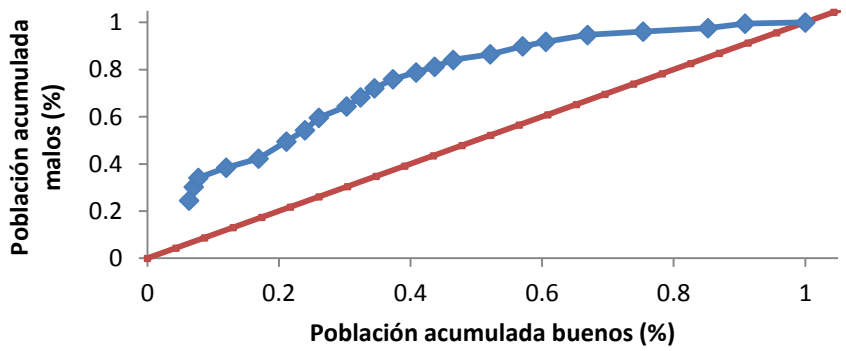
El Índice de Gini es un indicador de capacidad discriminante, el cual hace uso de la curva de Lorenz, esta curva se genera usando los porcentajes acumulados de casos buenos y malos, y se muestran en un gráfico de dispersión, esta curva se compara con una recta de 45°, y en donde, si la curva es muy semejante a la recta se considera que la variable no cuenta con poder discriminatorio.

Como esto se puede dejar a diferentes percepciones, se genera el índice de Gini, el cual es el área bajo la curva de Lorenz. Existen diferentes formas de calcular dicho índice, para este trabajo se usará la siguiente metodología:

Se estima el área de los rectángulos que se tiene por el porcentaje acumulado de cada uno de los valores de la variable, se suman las áreas obtenidas, a las cuales se les resta 0.5 y posteriormente se multiplica por 2, el valor que se tenga es el Índice de Gini de la variable, el cual va a permitir determinar la fuerza de discriminación de esta.

Continuando con el ejemplo de la Edad se tiene lo siguiente:

Gráfica 5. Curva de Lorenz



Fuente: Elaboración propia

Se considera que existe discriminación en la variable ya que la Curva de Lorenz tiene una buena separación con la recta de 45°, para tener un valor cuantitativo se estima el Índice de Gini, y los resultados se muestran en la siguiente tabla:

Tabla 1. Distribución de los casos de la variable edad simulada

Edad	Casos	Buenos	Malos	% Buenos	% Malos	% Acum Buenos	% Acum Malos	Base	Altura	Rectángulo
18	14	1	13	1%	8%	1%	8%	0.01	0.08	0.00
19	13	1	12	1%	7%	1%	15%	0.01	0.15	0.00
20	9	1	8	1%	5%	2%	19%	0.01	0.19	0.00
21	15	6	9	4%	5%	7%	25%	0.04	0.25	0.01
22	15	7	8	5%	5%	12%	29%	0.05	0.29	0.02
23	21	6	15	4%	9%	16%	38%	0.04	0.38	0.02
24	14	4	10	3%	6%	19%	44%	0.03	0.44	0.01
25	14	3	11	2%	6%	22%	51%	0.02	0.51	0.01
26	16	6	10	4%	6%	26%	56%	0.04	0.56	0.03
27	11	3	8	2%	5%	28%	61%	0.02	0.61	0.01
28	11	3	8	2%	5%	31%	66%	0.02	0.66	0.01
29	12	4	8	3%	5%	34%	71%	0.03	0.71	0.02
30	11	5	6	4%	4%	37%	74%	0.04	0.74	0.03
31	9	4	5	3%	3%	40%	77%	0.03	0.77	0.02
32	10	4	6	3%	4%	43%	81%	0.03	0.81	0.02
33	13	8	5	6%	3%	49%	84%	0.06	0.84	0.05
34	14	7	7	5%	4%	54%	88%	0.05	0.88	0.05
35	9	5	4	4%	2%	58%	90%	0.04	0.90	0.03
36	15	9	6	7%	4%	65%	94%	0.07	0.94	0.06
37	15	12	3	9%	2%	74%	95%	0.09	0.95	0.09

Edad	Casos	Buenos	Malos	% Buenos	% Malos	% Acum Buenos	% Acum Malos	Base	Altura	Rectángulo
38	17	14	3	10%	2%	84%	97%	0.10	0.97	0.10
39	12	8	4	6%	2%	90%	99%	0.06	0.99	0.06
40	14	13	1	10%	1%	100%	100%	0.10	1.00	0.10
Total general	304	134	170						**Suma del área de los rectángulos**	0.76
									Índice de Gini	0.51

Fuente: Elaboración propia

Dónde:

Base = % Acumulado de buenos en el valor g, menos el % acumulado de buenos en el valor g-1

Altura = % Acumulado de malos

Rectángulo = Base por altura

El Índice de Gini de la variable es del 51%, siendo este un valor aceptable para tomar en cuenta en el modelo. Empíricamente si una variable tiene un Índice de Gini por arriba del 20% se puede considerar para ser usada en la estimación del modelo.

Es de resaltar que el índice de Gini es usado como la base de selección de variables, sin embargo se pueden ocupar otras pruebas como a continuación se presentan.

- ***Prueba de Hipótesis***

Para realizar la prueba de Hipótesis se toma como supuesto que se están evaluando dos muestras distintas, donde se trata determinar si la diferencia entre ellas es significativa o es probable que se deba a la causalidad. Se parte del supuesto de que los dos grupos a evaluar son independientes entre sí, además se debe comprobar que las distribuciones son normales[1], en el caso de que lo anterior no sea así, la prueba de hipótesis no se puede llevar a cabo. En el caso contrario, en el cual se puede afirmar lo anterior (haciendo las pruebas correspondientes), se genera lo siguiente:

La prueba de hipótesis es la siguiente:

- $H_0 = \mu_{(\bar{x}_1 - \bar{x}_2)=0}$ (Las muestras son homogéneas)
- $H_1 = \mu_{(\bar{x}_1 - \bar{x}_2)\neq 0}$ (Las muestras no son homogéneas)

La fórmula que se ocupa para generar la prueba de hipótesis es:

[1] En el Anexo 2 se tiene detalle de una prueba de bondad de ajuste a la distribución normal

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}} \quad (3)$$

Dónde:

$\bar{x}$=es el valor medio del grupo 1 y 2

S = es la desviación estándar del grupo 1 y 2

n = es la cantidad de casos del grupo 1 y 2

En el evento en que la aplicación de la fórmula (3) se cuente con 30 o más casos el valor calculado se va a contrastar contra valores obtenidos con base a un nivel de significancia (siendo en lo general un valor de 0.05), con el cual haciendo uso de una tabla de distribución acumulada normal estándar se obtienen los niveles críticos que se usan para determinar si se rechaza o se acepta la hipótesis nula.

En el caso contrario en el que se tengan menos de 30 casos a evaluar el valor obtenido a través de la fórmula (3) se hace uso de un nivel de significancia y ciertos grados de libertad, con los cuales se obtienen los valores críticos para rechazar o no rechazar la hipótesis nula. Los grados de libertad se estiman con la siguiente fórmula:

$$gl = \frac{\left[\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)\right]^2}{\frac{\left(\frac{s_1^2}{n1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n2}\right)^2}{n_2 - 1}} \quad (4)$$

Para el ejemplo de la edad que se ha estado utilizando se consideran los siguientes datos:

- Se evaluará si la distribución de casos buenos y malos son iguales, por lo tanto se tiene la siguiente prueba de hipótesis:
  - $H_0 = \mu_{(\bar{x}_1 - \bar{x}_2)=0}$
  - $H_1 = \mu_{(\bar{x}_1 - \bar{x}_2)\neq 0}$

Para tal evaluación, se consideran los siguientes estadísticos

Tabla 2. Estadísticos de la variable edad simulada

Tipo	Media	Desviación	Casos
Bueno	32.04	6.42	134
Malo	26.39	5.99	170

Fuente: Elaboración propia

Utilizando la fórmula antes mencionada se obtiene el siguiente resultado:

$$z = \frac{(32.04 - 26.39)}{\sqrt{\left(\frac{6.42^2}{134} + \frac{5.99^2}{170}\right)}} = 7.859497 \quad (5)$$

Dado que se tienen más de 30 casos se utiliza una tabla normal acumulativa. Si se evalúa el valor a 0.05 de significancia, los valores críticos usados para la regla de decisión para rechazar o no la hipótesis son -1.96 y 1.96. Como el valor estimado queda por afuera de estos dos intervalos, se considera que se rechaza la hipótesis nula antes planteada, es decir, se considera que las muestras no son homogéneas.

- ***Valor de la información (Information Value [IV])***

Es una métrica de ordenación que se usa para evaluar la predicción de las variables. A continuación se muestra la fórmula:

$$IV = \sum_i (malos_i - buenos_i) * WOE_i = \sum_i (malos_i - buenos_i) * log\left(\frac{malos_i}{buenos_i}\right) \quad (6)$$

Dónde:

- $Buenos_i$ *son el porcentaje de observaciones buenas sobre total de buenos en el tramo* i
- $Malos_i$ *son el porcentaje de observaciones malas sobre total de malos en el tramo* i

Para determinar si el valor obtenido es bueno o malo se usa la siguiente escala:

- No predictiva = **IV** < 0.02
- Predicción débil = 0.02 <= **IV** < 0.10
- Predicción Media = 0.10 <= **IV** < 0.30
- Predicción fuerte = 0.30 <= **IV** < 0.50
- Sobre predicción = 0.5 <= **IV**

A continuación se muestra el **IV** de la variable Edad que se ha estado trabajando anteriormente.

Tabla 3. Estadísticos de la variable edad simulada

Edad	Casos	Buenos	Malos	% Buenos	% Malos	WOE	**IV**
18	14	1	13	0.75%	7.65%	-232.70%	0.1606
19	13	1	12	0.75%	7.06%	-224.69%	0.1418
20	9	1	8	0.75%	4.71%	-184.15%	0.0729
21	15	6	9	4.48%	5.29%	-16.75%	0.0014
22	15	7	8	5.22%	4.71%	10.44%	0.0005

Edad	Casos	Buenos	Malos	% Buenos	% Malos	WOE	**IV**
23	21	6	15	4.48%	8.82%	-67.83%	0.0295
24	14	4	10	2.99%	5.88%	-67.83%	0.0197
25	14	3	11	2.24%	6.47%	-106.13%	0.0449
26	16	6	10	4.48%	5.88%	-27.29%	0.0038
27	11	3	8	2.24%	4.71%	-74.29%	0.0183
28	11	3	8	2.24%	4.71%	-74.29%	0.0183
29	12	4	8	2.99%	4.71%	-45.52%	0.0078
30	11	5	6	3.73%	3.53%	5.56%	0.0001
31	9	4	5	2.99%	2.94%	1.48%	0.0000
32	10	4	6	2.99%	3.53%	-16.75%	0.0009
33	13	8	5	5.97%	2.94%	70.80%	0.0214
34	14	7	7	5.22%	4.12%	23.80%	0.0026
35	9	5	4	3.73%	2.35%	46.11%	0.0064
36	15	9	6	6.72%	3.53%	64.34%	0.0205
37	15	12	3	8.96%	1.76%	162.43%	0.1168
38	17	14	3	10.45%	1.76%	177.84%	0.1544
39	12	8	4	5.97%	2.35%	93.11%	0.0337
40	14	13	1	9.70%	0.59%	280.29%	0.2554
Total general	304	134	170			**IV variable**	1.1319

Fuente: Elaboración propia

El valor que se obtiene es mayor a 0.5 por lo cual se considera como una variable que puede funcionar dentro del modelo.

2.2.4.2. Partición de la muestra

Para generar este punto depende del tamaño de muestra con que se cuente. Lo que se busca es evitar un sobreajuste del modelo, por lo cual se toma una muestra representativa con el fin de generar el modelo, la cual se denomina muestra de entrenamiento, y se hace uso del resto de la información para verificar que los resultados obtenidos con la primera muestra sean estables, la muestra de validación.

En el caso de que se esté trabajando con poca información es recomendable hacer uso del 100% de la población, ya que en este sentido la poca información puede provocar que los resultados generados no sean estables.

2.2.4.3. Agrupación de las variables

Este es un punto importante ya que una mala generación de tramos de las variables puede generar resultados dudosos. En este punto se generan nuevas variables a partir de las ya existentes con el objetivo de que las generadas tengan un alto poder discriminante.

Una de las metodologías que se ocupa en la industria financiera es basar la generación de los grupos a partir del cálculo del WOE. Como es de recordar, en el apartado de análisis bivariante se mencionó dicho cálculo, al momento de estimar el WOE para las variables sin transformación por lo general no se va a tener una tendencia continua, es decir se van a tener picos a lo largo de la distribución, como se mostró en la gráfica número 2 que se utilizó como ejemplo.

En esta ocasión lo que se busca es que la tendencia sea continua, es decir, que no exista algún tipo de distorsión a lo largo de la distribución de la nueva variable. Lo que se pretende es tener grupos de población cuyo comportamiento sea homogéneo ante el incumplimiento.

2.2.4.4. Regresión Logística

La herramienta que se ocupa para generar el modelo de probabilidad es un modelo Logit ya que dicha función ayuda en la estimación de valores que entren dentro de las posibilidades de una probabilidad, 0 y 1.

Logit es un modelo de respuesta binaria, es decir, la variable que se trata de explicar tiene dos valores, cero o uno, para lo cual se ocupan variables de interés:

$$P(y = 1|x) = P(y = 1|x_1, x_2, \dots, x_k) \tag{7}$$

Donde “x” son las variables a usar.

Aquí es donde entra la relevancia de todo el análisis de las variables independientes y de la definición de la variable objetivo.

La definición de la variable objetivo va a determinar qué tanta relevancia pueden tener las variables independientes, lo cual afecta al análisis de selección de variables, así como el agrupado que se genera para cada variable.

Dado que lo que se busca es que el valor obtenido se encuentre dentro del intervalo de 0 y 1, se ocupa la siguiente función.

$$P(y = 1|x) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = G(\beta_0 + x\beta) \tag{8}$$

Donde G es una función que asume valores estrictamente entre cero y uno: $0 < G(z) < 1$, para todos los números reales, asegurando que las probabilidades de respuesta estimadas están estrictamente entre cero y uno.

Una de las funciones no lineales que se usan como función G(z) es el utilizado en el modelo Logit, y cuya fórmula es la siguiente:

$$G(z) = \frac{e^z}{1 + e^z} = \Lambda(z) \tag{9}$$

2.2.5. Discriminación del modelo

En este punto se generan las pruebas para determinar que el modelo que se estima a partir de la metodología que se menciona en el apartado anterior permite discriminar la población haciendo uso de la variable objetivo que se definió en un inicio.

La prueba más común que se puede obtener para este punto es la del Índice de Gini, para lo cual se hace uso de la probabilidad que se estima usando el modelo Logit y la variable objetivo.

2.2.6. Tarjetas de puntuación

Dependiendo de la forma en que se quiera dar a conocer el Score, se pueden usar diferentes formatos, usar valores con puntos decimales o con números decrecientes. La elección que se haga no va a afectar el resultado, ya que el escalado de los valores ayuda a mostrar los resultados.

Esta variación de la forma en que se muestra la scorecard se puede deber a las siguientes causas:

- Dependiendo del tipo de Software que se ocupe en la empresa.
- Facilidad para que el personal lo entienda y lo utilice.
- Llevar una continuidad con lo que ya se ha estado trabajando en la empresa.

En general la relación entre probabilidades y el score se puede representar como una transformación lineal:

$$\text{Puntuación} = \text{Offset} + \text{Factor}\ln(\text{probabilidad}) \tag{10}$$

Cuando una scorecard es desarrollada usando probabilidades específicas hacia un score y especificando “puntos para doblar las probabilidades” (pdo en inglés), el factor y el offset se calculan usando las siguientes fórmulas:

$$\text{Factor} = \text{pdo}/\ln(2) \tag{11}$$
$$\text{Offset} = \text{puntuación} - (\text{Factor} * \ln(\text{probabilidad base})) \tag{12}$$

Por ejemplo, si la scorecard es escalada donde el usuario busca probabilidades base de 50:1 en 600 puntos y busca las probabilidades al doble cada 20 puntos (pdo=20), el factor y el offset van a ser:

$$\text{Factor} = \frac{20}{\ln(2)} = 28.8539 \tag{13}$$
$$\text{Offset} = 600 - (28.8539 * \ln(50)) = 487.123 \tag{14}$$

Y cada score correspondiente a cada conjunto de probabilidades (para cada atributo) se calcula de la siguiente forma:

$$\text{Puntuación} = 487.123 + 28.8539 * \ln(probabilidad) \tag{15}$$

Esta fórmula se puede usar para generar cualquier tipo de score, usando cualquier modelo que pueda estimar una probabilidad de malo.

Cuando se trabaja con el WOE como insumo para generar la probabilidad, la fórmula anterior se modifica en lo siguiente:

$$\text{Puntuación} = \ln(\text{odds}) * \text{factor} + \text{offset} \tag{16}$$

$$= -\left(\sum_{j,i=1}^{k,n}\left(\text{woe}_j * \beta_i\right) + a\right) * \text{Factor} + \text{offset}$$

$$= -\left(\sum_{j,i=1}^{k,n}\left(\text{woe}_j * \beta_i\right) + \frac{a}{n}\right) * \text{Factor} + \text{offset}$$

$$= \sum_{j,i=1}^{k,n}\left(-\left(\text{woe}_j * \beta_i\right) + \frac{a}{n}\right) * \text{Factor} + \frac{\text{offset}}{n}) \tag{17}$$

Dónde:

WOE = peso de evidencia para cada atributo agrupado
β = el coeficiente de la regresión para cada variable
a = el intercepto de la regresión logística
n = número de variables
k = número de grupos en cada variable

La fórmula calcula el score que va a ser asignado a cada grupo de atributos para cada variable que se usa en el desarrollo del scorecard, por lo cual sumando cada una de estas puntuaciones se obtiene el score final.

Por ejemplo, supongamos que tenemos los siguientes datos para la variable edad que se ha puesto en ejemplo en los apartados anteriores:

De la scorecard que se está construyendo se buscan probabilidades base de 40:1 en 600 puntos y se busca un pdo de 30, por lo cual al aplicar las fórmulas anteriores tenemos como valores del factor y del offset lo siguiente:

$$\text{Factor} = 30/\ln(2)$$
$$= 43.28 \tag{18}$$

$$\text{Offset} = 600 - (43.28 * \ln(40)) = 440.34 \tag{19}$$

Suponemos que el modelo logit que se tiene al final del desarrollo contiene 8 variables que son significativas, el intercepto del modelo es 0.354 y la beta asociada a la variable edad es -0.864.

Con estos datos se puede construir la tarjeta de puntuación para la variable edad, los valores de los WOEs asociados para cada agrupado de la variable y la puntuación final se presenta en la siguiente tabla

Tabla 4. Puntuaciones para la variable edad simulada

Grupo	WOE	WOE*Beta(edad)	Puntos
1	-0.965	0.8338	17

Grupo	WOE	WOE*Beta(edad)	Puntos
2	-0.427	0.3689	37
3	-0.125	0.1080	48
4	0.234	-0.2022	62
5	0.767	-0.6627	82

Fuente: Elaboración propia

Para dar una mayor claridad a esto a continuación se pondrá la fórmula y sustitución de valores con lo cual se genera la puntuación del grupo 1:

$$\text{Grupo1} = (-(\text{woe}_1 * \beta_{edad}) + \frac{\text{a}}{\text{n}}) * \text{Factor} + \frac{\text{offset}}{\text{n}}) \quad (20)$$

$$\text{Grupo1} = (-(-0.965 * -0.864) + \frac{0.354}{8}) * 43.28 + \frac{440.34}{8}) \quad (21)$$

$$\text{Grupo1} = 17$$

3. Validación del desarrollo del Scorecard

En el capítulo anterior se describió de forma general cómo se construye una scorecard, sin embargo, si se desea profundizar en este tema existen dos bibliografías que tratan de forma detallada cada uno de estos temas. El libro <<*Credit Risk Scorecards, Developing and Implementing Intelligent Credit Scoring*>> de Naeem Siddiqi proporciona temas teóricos sobre la construcción de una scorecard, también ofrece diferentes opciones sobre el cómo armar estas herramientas desde el inicio hasta el cómo presentar los resultados. Por otra parte, el libro <<*Credit Scoring and its applications*>> de Lyn C. Thomas y compañía, muestra un desarrollo de las herramientas de forma técnica, donde incluye diferentes propuestas para llevar a cabo los ejercicios que se deben de realizar en el desarrollo.

Cuando una persona toma como base a estos autores y empieza a ver el desarrollo de lo que es una herramienta se puede percatar de que existen puntos importantes sobre los cuales se debe tener control en la definición y desarrollo de ejercicios, debido a que de lo contrario pueden obtener resultados diferentes a lo esperado.

Considerando lo anterior, en este apartado se va a unir la parte teórica con la parte práctica con el fin de identificar los puntos importantes a revisar en un desarrollo, ya sea para una revisión de la herramienta por un área de control interno o auditoría, o para la validación del resultado por parte del área desarrolladora.

3.1. Puntos a revisar

Tomando en cuenta lo que ya se ha revisado en la parte teórica se podrán identificar los siguientes puntos como relevantes:

- Definición de la variable objetivo
- Tendencia lógica de las variables en base a la variable objetivo
- Agrupación de las variables
- Muestra de desarrollo del modelo Logit
- Discriminación del modelo

En la experiencia que se ha tenido en la revisión de los modelos, así como en el desarrollo de modelos internos se considera que estos puntos son pasos intermedios a considerar ya que pueden representar puntos de cambio los cuales pueden alterar el resultado final.

3.1.1. Definición de la variable objetivo

Como se mencionó en la parte teórica, la definición de esta variable es el punto central del trabajo. Tomando como ejemplo a las instituciones financieras, que es en donde se utilizan con mayor frecuencia las scorecards, el determinar la variable es indispensable y depende de qué tan estrictos se quiere ser.

Para un banco una scorecard se puede ocupar para originar créditos, los cuales pueden ser para personas de los cuales el banco cuenta con su información tales como la sociodemográfica, información de pagos de las personas o incluso el buró de crédito, o puede ser para nuevos clientes, de los cuales obtiene información la puede obtener a partir de la solicitud del crédito. Las scorecards ayudan a definir a quien darle un crédito, ya sea cliente o no. En este sentido, la variable objetivo va a estar definida en razón sobre lo que es malo para el banco, con lo cual inicia la discusión sobre qué es bueno y qué es malo.

Para llevar a cabo la selección de lo que es bueno o malo se puede utilizar el criterio del desarrollador, el cual debe tomar en cuenta cuál es el objetivo de la herramienta, o también se puede ocupar la definición de lo que se usa en el mercado, e incluso el tema regulatorio puede entra en esta definición.

Solo para ampliar un poco el tema, en los acuerdos de Basilea II se menciona una definición de lo que se considera como default o incumplimiento:

“El deudor se encuentra en situación de mora durante más de 90 días con respecto a cualquier obligación crediticia significativa frente al grupo bancario“

Lo establecido en los Acuerdos de Basilea II es considerado por los bancos centrales de cada país como una señal de buenas prácticas, las autoridades monetarias y financieras dentro de sus atribuciones de regulación pueden obligar a los bancos a adoptar esta definición al momento de generar sus herramientas.

3.1.2. Tendencia de las variables

Es importante que las variables que se pretenden ocupar dentro del modelo tengan una tendencia que sea coherente con la variable objetivo, para lo cual se hace uso de la información histórica que se tenga para definir si el comportamiento de la variable que hoy en día se comporta como lo ha venido haciendo a lo largo del tiempo.

La coherencia de la variable, en el desarrollo de estas herramientas, se refiere a que, poniendo el ejemplo de la capacidad de pago, se espera que entre mayor capacidad de pago el cliente puede hacer frente a sus obligaciones bancarias, y por el contrario entre menor capacidad se espera que se incumpla en los pagos.

En el apartado teórico de la scorecard se ocupó el ejemplo de la variable edad, la cual se considera que esta tiene un comportamiento determinado ante el incumplimiento, entre menor edad tenga la persona es más propenso a ser incumplido, en comparación a una persona mayor. Si se considera que esta variable se tiene que comportar de esta forma sin importar el tipo de segmentación que se esté realizando dentro del modelo, se puede definir que una persona con menor edad tiene mayor probabilidad de ser incumplida en comparación con una de mayor edad.

Considerando como ejemplo que la institución este desarrollando una nueva herramienta, la misma que definió que entre menor edad se tiene mayor probabilidad de incumplimiento, pero se encuentra que la edad de los clientes tiene un comportamiento diferente (mayor edad significa que se tiene una mayor probabilidad de incumplimiento) se debe evaluar si se deja de lado esta variable, ya que no figura dentro de la lógica de lo que se tiene históricamente. Este ejercicio se tiene que realizar en todas las variables que se están evaluando para ser usadas dentro del modelo.

El problema de las variables que no tienen una relación coherente con la variable objetivo por lo general se debe a que se cuenta con una muestra de desarrollo pequeña, en el caso de tener poca información se debe realizar la revisión de las variables, ya que existe la posibilidad de que se tenga una variable que funciona perfecto para el modelo pero que no tiene una relación lógica con la variable objetivo.

3.1.3. Agrupado de las variables

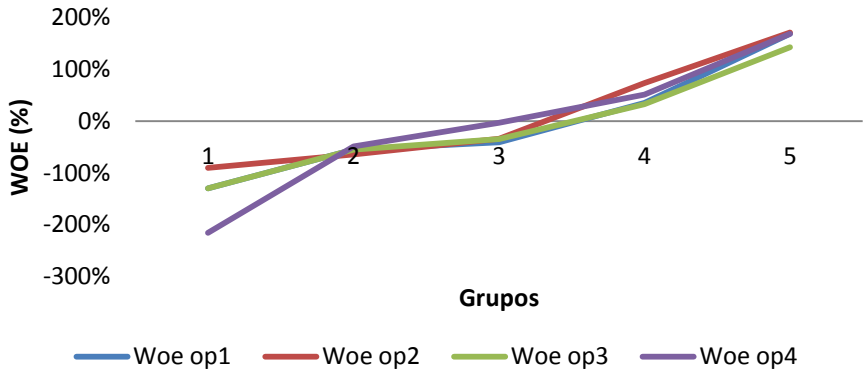
Al analizar la parte teórica se puede llegar a pensar que la selección de las variables es uno de los puntos importantes para la generación del modelo Logit, ya que puede ayudar a la discriminación del modelo, sin embargo, en el apartado de selección de variables, siendo muy estrictos, lo que ayuda es minimizar el trabajo de agrupado de las variables.

Generar el agrupado de las variables puede ser un trabajo complicado y más cuando no se cuenta con un método universal para generarlo. Considerando esto, no es lo mismo realizar este ejercicio para 10 variables, que a 20 variables e incluso mucho menos para 1,000 variables.

Por tal motivo es bueno generar la selección de variables, sin embargo en el momento en que se esté generando el modelo Logit se pueden descartar las variables no significativas, pero si se ocupan una gran cantidad de variables puede representar tiempo en el proceso para el programa que genera dicho modelo.

Utilizando la variable que ha servido de ejemplo, se estiman diferentes agrupaciones, de las cuales se debe definir cuál de ellas es la más adecuada para ser utilizada en la estimación del modelo. Para este ejemplo se generaron 4 opciones, las cuales de forma gráfica se muestran a continuación:

Gráfica 6. Ejemplos de distribución WOE para la variable edad simulada



Fuente: Elaboración propia

Como se puede ver, las cuatro opciones ahora tienen además de una tendencia continua, valores que no generan picos dentro de su distribución. Se pueden considerar como buena cualquiera de las cuatro, sin embargo hay que considerar dos factores para realizar la selección.

El primer factor radica en que dado que lo que se busca de las nuevas variables es que ayuden en la predicción de la variable objetivo, es necesario revisar el nuevo valor del **IV** que generan estos grupos, los resultados se presentan en la siguiente tabla:

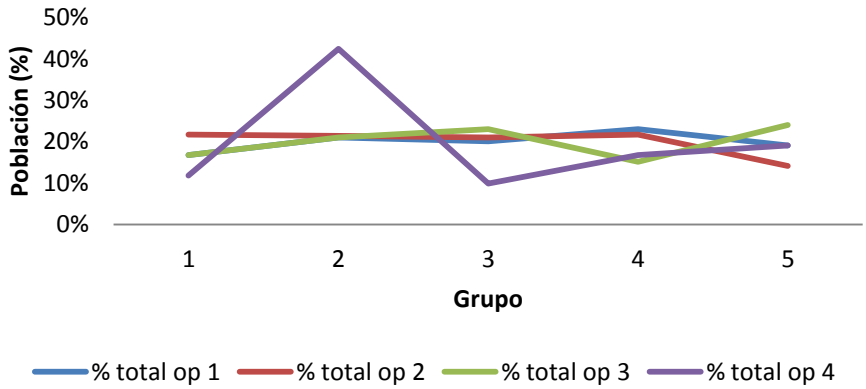
Tabla 5. IV de ejemplo de agrupación

Opción	IV	IV ABS
1	-0.84	0.84
2	-0.75	0.75
3	-0.79	0.79
4	-1.00	1.00

Fuente: Elaboración propia

De las cuatro opciones se considera como mejor la número 4 ya que es la que maximiza el **IV** de la variable. Sin embargo el siguiente factor a considerar puede ser igual de importante. Dicho factor puede ejemplificarse en la siguiente gráfica que muestra la forma en que se distribuyen el porcentaje de casos en cada uno de los grupos definidos para cada una de las opciones:

Gráfica 7. Ejemplos de distribución de casos de la variable edad simulada



Fuente: Elaboración propia

Al observar la Gráfica 7, la opción 4 se puede descartar, ya que en el segundo grupo contiene cerca del 40% de las observaciones, lo cual puede generar distorsiones en los resultados finales o sesgos en la evaluación del modelo. El caso de la opción 3 se aprecia que existen diferentes niveles de porcentaje de casos entre el grupo 4 y 5 que pueden generar dudas sobre su semejanza en los grupos, lo mismo pasa con la agrupación 2. Por último, el grupo 1, mantiene la cantidad de porcentaje de casos de entre 20% y 25% en cada uno de sus grupos.

En suma, se pueden generar diferentes agrupaciones sobre una variable, las cuales a primera vista pueden ser adecuadas para ser usadas en el modelo, sin embargo, los factores del **IV** y del porcentaje de casos en cada uno de los grupos es indispensable para determinar qué la agrupación no generen problemas al momento de generar el modelo, como lo son el sesgo de los resultados.

3.1.4. Muestra de desarrollo del modelo Logit

Este punto está muy relacionado con el siguiente. Se debe considerar este elemento ya que repercute en la discriminación del modelo.

Al momento de generar el modelo Logit se debe de determinar el tamaño de muestra que se va a ocupar. Existen dos bibliografías en particular que se toman en cuenta para este tema, una de ellas es el libro <<*Credit Scoring and Its Applications*>> de Thomas, Edelman y Crook, y por otra parte, el texto de <<*Credit Risk Scorecards Developing and Implementing Intelligent Credit Scoring*>> de Siddiqi, N., en ambos textos se menciona que es cierto que por lo general no existe una equivalencia entre la cantidad de casos de casos buenos y casos malos, considerando una buena práctica la generación del modelo Logit usando una muestra que sea lo más cercana al 50/50 o la misma proporción de casos.

Sin embargo, en la práctica se tiene con mayor frecuencia que el tamaño de las proporciones no son equivalentes, por lo cual para no perder una buena cantidad de información se ocupa toda la información disponible.

3.1.5. Discriminación del modelo

Para este punto se ocupan dos pruebas básicas para definir si existe una discriminación del modelo generado:

- Índice de Gini
- Matriz de confusión

La generación del Índice de Gini ya se había desarrollado en el apartado de la evaluación de las variables, en esta ocasión lo que se estará evaluando es la probabilidad que se estima a partir del modelo.

La matriz de confusión es un método que permite definir los casos que se acierta al momento de generar la probabilidad de la variable objetivo, y en su caso, establecer la cantidad de casos en los cuales se cayó en errores de tipo I y II.

Para generar la matriz se necesita los valores de Y que se le asignaron a cada caso, los valores de la probabilidad estimada y la definición de un punto de cohorte el cual se va a usar como frontera para determinar si la probabilidad corresponde a lo que se esperaba.

La matriz que se va a generar es de 3x3, a continuación se pone un ejemplo de esta matriz.

Tabla 6. Distribución de casos de matriz de confusión

Y REAL	Probabilidad Buenos	Probabilidad Malos	Total
0	1	2	T3
1	3	4	T4
Total	T1	T2	TT

Fuente: Elaboración propia

La información que se tiene que colocar en esta matriz se menciona a continuación:

- El punto 1 va a estar determinado por la cantidad de casos que en un inicio se habían establecido como casos buenos y como resultado de la probabilidad estimada y el punto de corte establecido se definen esos casos igualmente como buenos.
- El punto 4 es lo equivalente del punto 1, pero en lugar de casos buenos se revisan los casos malos.
- El punto 2 son los casos que la variable objetivo los marca como buenos, pero el resultado de la probabilidad estimada y haciendo uso del punto de corte definido se identifican como malos.

- El punto 3 son los casos que la variable objetivo los identifica como buenos, pero la probabilidad estimada haciendo uso del punto de corte usado anteriormente los define como malos.

Como ejemplo se utiliza la tabla **"Datos para matriz de confusión, punto de corte de 0.5"** que se presenta en el anexo 1. En un principio se cuenta con los casos buenos y malos definidos en un inicio del modelo, y adicionalmente se tienen las probabilidades estimadas. Si se agrupan los casos primero en aquellos que son buenos y malos se tiene lo siguiente:

Tabla 7. Casos buenos y malos para matriz de confusión

Y REAL	Casos
0	134
1	170

Fuente: Elaboración propia

Agregando el eje en el cual se define si los casos son buenos o malos según la probabilidad estimada se tiene lo siguiente:

Tabla 8. Distribución de casos de matriz de confusión

Y REAL	Probabilidad Buenos	Probabilidad Malos	Total
0	104	30	134
1	47	123	170
Total	151	153	304

Fuente: Elaboración propia

Para la Tabla 7 se ocupó un punto de corte de 50%, es decir que si la probabilidad que se estima es menor al 50% se considera que el caso es evaluado como bueno, de lo contrario se tomará como malo. Una vez que se generar esta variable binaria se agrega como columna en una tabla dinámica, con la cual se genera la tabla antes mostrada.

Con esta matriz se puede definir qué tan confiable puede ser el modelo generado, para lo cual, primero se debe separar los casos que se consideran como correctos e incorrectos, es decir, lo que se encuentra en las casillas antes definidas como 1 y 4, mientras que las otras dos son consideradas como incorrectas. Con esto se obtienen los porcentajes sobre el total de los casos que les corresponde y se obtiene la utilidad de la herramienta para discriminar. A continuación se muestra el cuadro resultado:

Tabla 9. Matriz de confusión

Y REAL	Prob Buenos	Prob Malos	Total
0	104	30	134
1	47	123	170
Total	151	153	304
Correcto	**104**	**123**	**227**
Incorrecto	**47**	**30**	**77**
% Correcto	78%	72%	75%
% Incorrecto	28%	22%	25%

Fuente: Elaboración propia

Como se ve en la tabla anterior, se obtiene un porcentaje de correcta estimación (en base a un punto de corte del 50%) del 75%, el cual se considera adecuado, por otra parte, se tiene un porcentaje del 25% de casos mal identificados.

Con los mismos datos de la tabla del anexo 1 se puede generar el Índice de Gini, usando la metodología antes mencionada, donde se obtiene un valor del 86.34%.

El criterio de ocupar el punto de cohorte al 50% es por ser estándar, sin embargo, se pueden ocupar metodologías para determinar el punto de cohorte adecuado para el tipo de información que se tiene.

Para concluir, se debe considerar que la finalidad de una scorecard es discriminar los casos buenos de los casos malos, por lo cual se deben realizar los cálculos y agrupaciones necesarias para llegar a ese objetivo. Para definir si se llega o no al objetivo usualmente se utiliza el Índice de Gini, sin embargo, para dar mayor fortaleza a la conclusión que se tiene se puede utilizar la matriz de confusión, la cual nos muestra el porcentaje de casos en los cuales su probabilidad es consistente con la variable objetivo que en un inicio le fue asignada.

4. Conclusiones

Con este trabajo se puede concluir que para llevar a cabo la construcción de una scorecard no son necesarios programas especializados sobre la materia, lo que se necesita es tener claro cuál es el objetivo de la herramienta para utilizar alguna metodología que ayude a llegar al objetivo. Sin embargo esto no implica que el tener los conocimientos nos dé como resultado desarrollar una herramienta en poco tiempo, se necesita llevar a cabo un amplio análisis de la información que se dispone, así como de la exploración de las diferentes metodologías que puedan ayudar al desarrollo.

En las instituciones financieras el desarrollo de estas herramientas es de importancia ya que es el primer filtro que ocupan para dar préstamos a los clientes y nuevos clientes. Sobre el desarrollo, se espera tener un alto nivel de discriminación, ya que al llevar a cabo en la integración de los procesos se pueden disminuir las pérdidas que puedan tener las instituciones financieras al momento de otorgar un nuevo crédito. Para ser concretos, los bancos por parte de la regulación que tienen, deben tener reservas para hacer frente a las perdidas, las cuales se estiman de forma mensual, por lo cual entre menos pérdidas tengan (y haciendo uso de la metodología regulatoria o interna para calcular las pérdidas esperadas) menores serán las reservas que tengan que guardar los bancos y por lo tanto esto representa un aumento en los beneficios del banco.

Como se vio en el apartado 3 existen puntos importantes que se deben revisar para confirmar que el desarrollo de la scorecard pueda cumplir con el objetivo que se plantea, el cual es la discriminación de los casos buenos y malos. Además es importante considerar la revisión del poder discriminante de la herramienta de forma anual, ya que el cambio de las distribuciones de las variables pueden generar que las puntuaciones que se tienen al final del desarrollo ya no sirvan, lo cual representa generar una calibración a la herramienta o en el caso extremo una nueva herramienta.

Como bien ya se mencionó, el análisis de la información puede ser una tarea ardua, es por tal motivo que se pueden hacer uso de softwares especializados para el manejo de la información, en la actualidad es muy común que las instituciones financieras con alto volumen de información, así como de procesos de análisis de la información utilicen el programa Statistical Analysis System (SAS), el cual ayuda de forma significativa en los análisis debido a que estos se pueden automatizar en el caso de que se tenga un criterio establecido.

5. Anexo 1

Tabla 10. Datos para matriz de confusión, punto de corte de 0.5

Folio	y	y est	Punt CORTE
1	0	0.1	0
2	1	0.73	1
3	1	0.52	1
4	0	0.53	1
5	1	0.74	1
6	0	0.55	1
7	1	0.64	1
8	0	0.18	0
9	0	0.44	0
10	1	0.84	1
11	1	0.6	1
12	0	0.17	0
13	0	0.33	0
14	0	0.37	0
15	0	0.09	0
16	1	0.89	1
17	1	0.47	0
18	0	0.26	0
19	0	0.49	0
20	1	0.5	1
21	1	0.97	1
22	1	0.37	0
23	1	0.43	0
24	1	0.9	1
25	1	0.86	1
26	1	0.45	0
27	0	0.47	0
28	1	0.45	0
29	1	0.76	1
30	0	0.48	0
31	0	0.03	0
32	1	0.64	1
33	0	0.45	0
34	0	0.25	0
35	1	0.37	0
36	1	0.61	1
37	0	0.46	0
105	1	0.72	1
106	1	0.76	1
107	1	0.78	1
108	1	0.94	1
109	1	0.36	0
110	0	0.5	1
111	0	0.32	0
112	1	0.44	0
113	1	0.84	1
114	0	0.57	1
115	1	0.8	1
116	1	0.52	1
117	0	0.44	0
118	1	0.91	1
119	1	0.47	0
120	1	0.59	1
121	0	0.59	1
122	1	0.92	1
123	0	0.28	0
124	1	0.55	1
125	1	0.83	1
126	1	0.92	1
127	1	0.39	0
128	1	0.92	1
129	1	0.54	1
130	1	0.9	1
131	1	0.56	1
132	1	0.59	1
133	0	0.23	0
134	1	0.76	1
135	1	0.76	1
136	0	0.49	0
137	1	0.89	1
138	0	0.15	0
139	0	0.32	0
140	0	0.38	0
141	1	0.3	0
209	1	0.33	0
210	1	0.39	0
211	1	0.3	0
212	1	0.68	1
213	0	0.4	0
214	0	0.08	0
215	1	0.51	1
216	1	0.8	1
217	0	0.14	0
218	0	0.35	0
219	0	0.15	0
220	1	0.53	1
221	1	0.74	1
222	1	0.4	0
223	1	0.91	1
224	0	0.37	0
225	0	0.27	0
226	0	0.09	0
227	0	0.47	0
228	1	0.41	0
229	0	0.25	0
230	0	0.54	1
231	1	0.35	0
232	1	0.78	1
233	1	0.63	1
234	1	0.54	1
235	0	0.25	0
236	0	0.13	0
237	0	0.46	0
238	1	0.74	1
239	1	0.9	1
240	1	0.62	1
241	0	0.56	1
242	1	0.43	0
243	0	0.09	0
244	1	0.88	1
245	1	0.63	1

Folio	y	y est	Punt CORTE
38	1	0.74	1
39	0	0.47	0
40	0	0.23	0
41	1	0.95	1
42	1	0.68	1
43	0	0.33	0
44	0	0.19	0
45	0	0.03	0
46	0	0.17	0
47	1	0.33	0
48	0	0.09	0
49	1	0.77	1
50	1	0.55	1
51	1	0.91	1
52	0	0.12	0
53	0	0.39	0
54	0	0.25	0
55	0	0.53	1
56	1	0.63	1
57	0	0.18	0
58	1	0.88	1
59	0	0.27	0
60	0	0.46	0
61	1	0.96	1
62	1	0.45	0
63	1	0.57	1
64	0	0.55	1
65	1	0.59	1
66	0	0.14	0
67	1	0.77	1
68	1	0.59	1
69	0	0.06	0
70	0	0.16	0
71	1	0.55	1
72	1	0.68	1
73	1	0.47	0
74	1	0.31	0
75	0	0.11	0
76	0	0.38	0
77	0	0.45	0
142	1	0.45	0
143	0	0.34	0
144	0	0.24	0
145	0	0.37	0
146	1	0.78	1
147	1	0.32	0
148	0	0.57	1
149	1	0.83	1
150	0	0.16	0
151	1	0.59	1
152	0	0.32	0
153	1	0.32	0
154	0	0.02	0
155	1	0.62	1
156	1	0.81	1
157	0	0.13	0
158	0	0.24	0
159	1	0.48	0
160	1	0.83	1
161	0	0.27	0
162	1	0.35	0
163	1	0.45	0
164	1	0.36	0
165	1	0.67	1
166	1	0.58	1
167	0	0.42	0
168	1	0.72	1
169	1	0.59	1
170	0	0.26	0
171	0	0.05	0
172	0	0.1	0
173	1	0.58	1
174	0	0.08	0
175	0	0.53	1
176	1	0.48	0
177	0	0.55	1
178	0	0.39	0
179	0	0.28	0
180	1	0.97	1
181	0	0.18	0
246	1	0.52	1
247	0	0.09	0
248	0	0.56	1
249	1	0.35	0
250	0	0.55	1
251	1	0.96	1
252	0	0.06	0
253	1	0.36	0
254	0	0.56	1
255	1	0.78	1
256	0	0.31	0
257	1	0.47	0
258	1	0.71	1
259	0	0.12	0
260	1	0.93	1
261	1	0.97	1
262	1	0.51	1
263	0	0.44	0
264	0	0.45	0
265	0	0.53	1
266	1	0.99	1
267	1	0.77	1
268	0	0.46	0
269	1	0.9	1
270	1	0.84	1
271	0	0.18	0
272	1	0.33	0
273	0	0.52	1
274	0	0.28	0
275	1	0.54	1
276	1	0.88	1
277	0	0.5	1
278	1	0.99	1
279	1	0.34	0
280	1	0.97	1
281	0	0.24	0
282	1	0.56	1
283	0	0.31	0
284	1	0.59	1
285	1	0.92	1

Folio	y	y est	Punt CORTE
78	0	0.25	0
79	0	0.13	0
80	1	0.38	0
81	1	0.7	1
82	0	0.51	1
83	1	0.88	1
84	1	0.46	0
85	0	0.14	0
86	0	0.36	0
87	1	0.47	0
88	1	0.88	1
89	1	0.44	0
90	1	0.76	1
91	1	0.8	1
92	1	0.5	1
93	1	0.47	0
94	0	0.21	0
95	0	0.59	1
96	0	0.57	1
97	1	0.86	1
98	0	0.17	0
99	1	0.95	1
100	1	0.72	1
101	1	0.87	1
102	1	0.35	0
103	0	0.58	1
104	1	0.42	0

Folio	y	y est	Punt CORTE
182	1	0.84	1
183	1	0.44	0
184	0	0.46	0
185	0	0.35	0
186	0	0.59	1
187	1	0.88	1
188	1	0.62	1
189	1	0.54	1
190	0	0.48	0
191	0	0.26	0
192	1	0.97	1
193	0	0.55	1
194	0	0.09	0
195	1	0.51	1
196	1	0.48	0
197	0	0.36	0
198	1	0.96	1
199	0	0.6	1
200	0	0.45	0
201	0	0.49	0
202	1	0.5	1
203	1	0.79	1
204	1	0.82	1
205	1	0.76	1
206	1	0.92	1
207	0	0.29	0
208	0	0.33	0

Folio	y	y est	Punt CORTE
286	1	0.6	1
287	1	0.47	0
288	0	0.6	1
289	0	0.23	0
290	0	0.56	1
291	1	0.75	1
292	1	0.47	0
293	0	0.22	0
294	0	0.29	0
295	0	0.52	1
296	1	0.47	0
297	1	0.96	1
298	1	0.8	1
299	0	0.1	0
300	1	0.54	1
301	1	0.77	1
302	1	0.58	1
303	0	0.55	1
304	0	0.6	1

Fuente: Elaboración propia

6. Anexo 2

Prueba de bondad de ajuste de la distribución Normal

Tomando en cuenta el ejercicio de la edad que se ve a lo largo del capítulo 2 se genera una prueba en la cual se busca es probar si los datos tienen una distribución normal, para lo cual se cuenta con lo siguiente:

Tabla 11. Variable edad, prueba de bondad de ajuste a la distribución Normal

Edad	Frecuencias Observadas
18	14
19	13
20	9
21	15
22	15
23	21
24	14
25	14
26	16
27	11
28	11
29	12
30	11
31	9
32	10
33	13
34	14
35	9
36	15
37	15
38	17
39	12
40	14
Total	304

Fuente: Elaboración propia

Para este ejercicio se utilizan las siguientes hipótesis:

H_o: los datos se ajustan a una distribución normal.

H_l: los datos no se ajustan a una distribución normal.

Sobre la información de la tabla número 12 se obtiene la media y la desviación estándar usando las siguientes fórmulas:

$$Media = \frac{\sum(Marcas_i) * f_i}{N} \tag{22}$$

$$Desviación\ Estándar = \sqrt{\frac{\sum(Marcas_i)^2 * f_i}{N} - (Media)^2} \tag{23}$$

Adicional a estos cálculos se debe obtener la Probabilidad Teórica (PT) de los datos, para lo cual se debe generar un valor sobre el cual se desea la distribución haciendo uso de la media, desviación estándar y las frecuencias observadas como a continuación se muestra:

$$Z = \frac{Marca - FO}{Desviación\ Estándar} \tag{24}$$

El valor obtenido se busca dentro de una tabla de distribución acumulada Normal Estándar. Dado que la probabilidad que se obtiene es de una distribución acumulada se deben generar las diferencias entre cada una de las marcas, con excepción de la primera de ellas, ya que con esto se obtiene la probabilidad teórica entre cada una de las marcas. Los resultados de estas tablas se encuentran a continuación:

$$Media = \frac{\sum(Marcas_i) * f_i}{N} = 29 \tag{25}$$

$$Desviación\ Estándar = \sqrt{\frac{\sum(Marcas_i)^2 * f_i}{N} - (Media)^2} = 6.77 \tag{28}$$

Tabla 12. Variable edad, bondad de ajuste a la distribución Normal - Resultados

Edad	FO	Z	G(z)	PT	FT	(FO-FT)^2/FT
18	14	-1.61	0.05	0.05	16.32	0.33
19	13	-1.46	0.07	0.02	5.61	9.74
20	9	-1.31	0.10	0.02	6.98	0.59
21	15	-1.16	0.12	0.03	8.49	5.00
22	15	-1.02	0.15	0.03	9.38	3.37
23	21	-0.87	0.19	0.04	11.64	7.53
24	14	-0.72	0.24	0.04	13.26	0.04
25	14	-0.57	0.28	0.05	14.77	0.04
26	16	-0.43	0.33	0.05	14.98	0.07
27	11	-0.28	0.39	0.06	17.07	2.16
28	11	-0.13	0.45	0.06	17.80	2.60

Edad	FO	Z	G(z)	PT	FT	(FO-FT)^2/FT
29	12	0.02	0.51	0.06	18.15	2.08
30	11	0.17	0.57	0.06	18.09	2.78
31	9	0.31	0.62	0.05	16.49	3.40
32	10	0.46	0.68	0.06	16.88	2.80
33	13	0.61	0.73	0.05	15.76	0.48
34	14	0.76	0.78	0.05	14.38	0.01
35	9	0.90	0.82	0.04	12.03	0.76
36	15	1.05	0.85	0.04	11.31	1.20
37	15	1.20	0.88	0.03	9.66	2.95
38	17	1.35	0.91	0.03	8.07	9.87
39	12	1.49	0.93	0.02	6.20	5.42
40	14	1.64	0.95	0.02	5.35	13.97
Total	304			1.00	304	77.18

Fuente: Elaboración propia

Donde:

FT (Frecuencia teórica) = N*Probabilidad Teórica

El valor que se obtiene al final es la suma de la última columna de la tabla 13, es decir el valor de 77.18. Este valor se compara contra una tabla de Chi – Cuadrada, para lo cual se deben obtener los grados de libertad de la siguiente forma:

$$grados\ de\ libertad = K - R - 1 = 23 - 2 - 1 = 20 \qquad (26)$$

Donde K son el número de filas, en este ejemplo son 23, y R son las columnas, las cuales son la FO y la FT. Como se muestra en la fórmula (29) los grados de libertad son 20, y tomando en cuenta un nivel de significancia del 5%, el valor en tablas es 31.41, debido a que el valor que se tiene del ejercicio, el 77.18 que se tiene sombreado en la tabla 13, es mayor se rechaza la hipótesis nula, es decir se considera que los datos no se ajustan a una distribución normal.

7. Bibliografía

1. BBVA Informe financiero (2010). Probabilidad de incumplimiento (PD). Disponible en: http://accionistaseinversores.bbva.com/TLBB/micros/informes2010/es/Gestiondelriesgo/ProbabilidaddeincumplimientoPD.html
2. Canavos, G. (1998). Probabilidad y Estadística: Aplicaciones y métodos (1st ed.). México: Mc Graw.
3. Cheema, J., Thielbar, M. (2003). Programming II: Manipulating data with the DATA step Course Notes. Estados Unidos de América: SAS Institute Inc.
4. Cheema, J., Thielbar, M. (2003). Programming II: Manipulating data with the DATA step Course Notes. Estados Unidos de América: SAS Institute Inc.
5. Comité de supervisión Bancaria de Basilea (2006). Convergencia internacional de medidas y normas de capital. Estados Unidos de América: Banco de pagos internacionales.
6. Cosar, T., Boj, E., Fortiana, J. (2012) Bondad de ajuste y elección de puntos de corte en regresión logística basada en distancias, aplicación al problema de credit scoring, Anales del Instituto de Actuarios Españoles, 1, (18), 19-40
7. Curso de la materia de Economía Cuantitativa Estadística Aplicada a la Mercadotecnia, del semestre 2013-2, con el profesor Jacobo López Barojas.
8. Gujarati, D. (2004). Econometría (4th ed.). México: McGraw-Hill.
9. Gutiérrez, M. (2007). Modelos de credit scoring - ¿Qué, cómo, cuándo y para qué?. [en línea] Argentina: Banco Central de la República Argentina. Disponible en: http://www.bcra.gob.ar/Pdfs/Publicaciones/CreditScoring.pdf [2015, 12 de Septiembre]
10. Moral, G., Gavilá, S., Población, J. (2006). Carteras Minoristas. Sistemas de scoring: construcción y evaluación. [Diapositiva]. Madrid: Segundo seminario sobre Basilea II, 23 diapositivas.
11. Ochoa, J., Galeando, W., Agudelo, L. (2010). Construcción de un modelo de scoring para el otorgamiento de crédito en una entidad financiera. Perfil de Coyuntura Económica, 1 (16), 191-222.
12. Ramírez, M. (2014). Econometría con estimaciones para México (1st ed.). México: Facultad de Economía de la Universidad Nacional Autónoma de México.
13. Rayo, S., Lara, J., Camino, D. (2010). Un modelo de Credit Scoring para instituciones de microfinanzas en el marco de Basilea II. Journal of Economics, Finance and Administrative Science, 15 (28), 89-123.
14. SAS Institute Inc. (2012). Developing credit scorecards using credit scoring for SAS Enterprise Miner 12.1. [en línea] Estados Unidos: SAS Institute Inc. Disponible en: http://support.sas.com/documentation/cdl/en/emcsgs/66008/PDF/default/emcsgs.pdf [2015, 12 de Septiembre]
15. Siddiqi, N. (2005). Credit risk scorecards: developing and implementing intelligent credit scoring (1st ed.). Estados Unidos de América: John Wiley & Sons, Inc.
16. Simon, J. (2013). Macro Language II: Advanced Techniques. Estados Unidos de América: SAS Institute Inc

17. Simon, J., Mitterling, L. (2013). Macro Language I: Essentials Course Notes. Estados Unidos de América: SAS Institute Inc.
18. Spiegel, M., Stephens, L. (2002). Estadística (3rd ed.). México: Editorial Mc Graw Hill
19. Thomas, L., Edelman, D., Crook, J. (2002). Credit Scoring and its Applications (1st ed.). Estados Unidos de América: Society for Industrial and Applied Mathematics
20. Wooldrige, J. (2012). Introductory econometrics: a modern approach (4th ed.). Estados Unidos de América Editorial South-Western Cengage Learning.