MODELO DE APROBACIÓN DE CRÉDITO EN UNA ENTIDAD FINANCIERA.

JEFERSON MAURICIO MARIN APONTE
VIVIAN YULIETH MARULANDA AROCA

FUNDACION UNIVERSITARIA LOS LIBERTADORES
FACULTAD DE CIENCIAS BASICAS
PROGRAMA DE ESTADÌSTICA APLICADA
BOGOTA, D.C
2016

MODELO DE APROBACIÓN DE CRÉDITO EN UNA ENTIDAD FINANCIERA.

JEFERSON MAURICIO MARIN APONTE
VIVIAN YULIETH MARULANDA AROCA

Asesor
ALEX ZAMBRANO

FUNDACION UNIVERSITARIA LOS LIBERTADORES
FACULTAD DE CIENCIAS BASICAS
PROGRAMA DE ESTADÌSTICA APLICADA
BOGOTA, D.C
2016

TABLA DE CONTENIDO

Listado de Tablas e Ilustraciones

MODELO DE APROBACIÓN DE CRÉDITO EN UNA ENTIDAD FINANCIERA.

** Jeferson Mauricio Marin, Vivian Yulieth Marulanda

Resumen

En este estudio se presenta una visión general de un modelo estadístico en una base de datos simulada de una entidad financiera, utilizado como una herramienta para medir el riesgo alto o bajo al tomar la decisión de otorgar un crédito. Se basa en el análisis de dos tipos de datos referente a los clientes, datos demográficos y datos de crédito. Esta metodología da un referente para la decisión final de conceder o rechazar la solicitud de crédito.

ABSTRAC

In this study an overview of a statistical model is presented in a simulated database of a financial institution, used as a tool to measure high or low when making the decision to grant credit risk. It is based on the analysis of two types of data relating to customers, demographic data and credit data. This methodology provides a reference for the final decision to grant or reject the loan application.

Palabras clave: modelo, base simulada, metodología, modelo logit, riesgo.

1. Introducción

Una prioridad de cualquier entidad financiera es contar con criterios confiables para establecer a quién debe conceder un crédito y en qué medida hacerlo; de ahí la razón por la que es importante tener un instrumento con el cual medir el riesgo que se corre al aprobar un crédito y es el interés de estas entidades reducir lo más posible tal riesgo y de esta manera otorgar o rechazar créditos a clientes o posibles clientes.

Las decisiones financieras cotidianas persiguen la obtención de rendimiento, dentro de un ambiente de riesgo. Todas las evaluaciones en las decisiones financieras se establecen en modelos que relacionan los niveles de rendimiento con el riesgo y soportados en modelos estadísticos. (Navarro ,2003)

Este, se trata de una metodología que clasifica el nivel de riesgo en alto o bajo para aprobación o rechazo de créditos, se basa en el análisis de dos tipos de datos de los clientes y nuevos clientes: datos de crédito, como pueden ser su historial crediticio y su comportamiento en cuanto a la morosidad de pagos y datos demográficos como pueden ser edad, sexo, ingresos, situación laboral, etc. (Núñez, 2011)

El modelo aquí descrito que discrimina el riesgo en alto o bajo consiste en una fórmula. Se podrá evaluar el riesgo de otorgar un crédito si el modelo lo clasifica como alto o bajo. A partir del resultado se genera un puntaje (score) que se le asocia a las variables predictivas para indicar un nivel de riesgo. El modelo da un marco para la decisión final en el otorgamiento o rechazo de la solicitud de créditos. Esto le permite a la entidad financiera ser más objetiva al momento de decidir la aprobación o rechazo de créditos, aunque es claro que el modelo no determina con exactitud la conducta de un cliente en particular, pero si da una visión de comportamiento para clientes con características comunes.

Normalmente las entidades deciden el puntaje mínimo de riesgo para otorgar un crédito basados en su experiencia, pero en la entidad de estudio no se ha realizado un modelo de este tipo, así que la decisión la da el estadístico.

Este trabajo está estructurado en seis capítulos. En el primero se presenta el trabajo. En el segundo se hace una revisión de los conceptos generales de estadística que son utilizados en el credit scoring. En el capítulo tres se explica los pasos que llevo a lugar este estudio. El desarrollo general de la técnica credit scoring utilizada para este trabajo se encuentran en el capítulo cuatro. En el capítulo cinco se da un ejemplo donde se aplica el modelo y en el capítulo seis se concluye a partir de los resultados obtenidos.

1.1. Formulación del Problema:

¿Cómo clasificar el riesgo de crédito de los clientes de una entidad financiera con datos simulados a través de un modelo estadístico?

1.2. Justificación:

Se han realizado varias investigaciones para determinar las principales causas de mora en pagos de los créditos otorgados por diferentes bancos a nivel mundial, Núñez, G. (2011), pero en la entidad financiera en estudio no, por lo que, un trabajo de este tipo es de utilidad en esta entidad, pues, se corre el riesgo de tomar malas decisiones en cuanto a qué cliente asignar un préstamo. Así que una de las necesidades más importantes para las entidades financieras, es tener criterios confiables para determinar a qué personas deben otorgar préstamos. De esta manera, se quiere buscar un modelo estadístico que clasifique el riesgo de crédito de los clientes de una entidad financiera utilizando los datos de préstamos.

1.3. Objetivos:

Objetivo general

Clasificar el riesgo de crédito de los clientes de una entidad financiera con datos simulados a través de un modelo estadístico.

Objetivos específicos

a) Identificar cuáles son las variables de una base de datos simulada que se empleara en el diseño de un modelo estadístico que clasifique el riesgo de crédito de los clientes de una entidad financiera.
b) Generar un modelo estadístico teniendo en cuenta las variables identificadas.
c) Evaluar la clasificación de crédito a los clientes de una entidad financiera con datos simulados.

2. MARCO DE REFERENCIA

La base de datos simulada cuenta con variables demográficas e históricos de préstamos a clientes y sus comportamientos en cuanto a la morosidad de pagos.

Para clasificar los clientes según su comportamiento con los pagos es necesario identificar los tipos de clientes

En 2010, el estudio de Nieto, S. Crédito al Consumo: La Estadística aplicada a un problema de Riesgo Crediticio mostró que:

1. Se considera buen cliente a aquellos individuos que:

a. Pagan el monto de su deuda en el periodo de gracia (entre los límites de pago o de corte).
b. Cuando no cuenta con el capital para pagar la totalidad de la deuda pero pagan al menos el mínimo requerido por la empresa acreedora.
c. Liquidan su adeudo en no más del tiempo determinado por la empresa.

2. Un mal cliente se refiere a aquel deudor que causa pérdidas económicas a la compañía. Este tipo de clientes no pagaron su cuenta, aun después de aplicarles técnicas de cobranzas. Comúnmente es considerado mal cliente después de 90 días de mora.

Como este estudio está basado en el diseño de un modelo estadístico para determinar el riesgo de pago de los clientes de una entidad financiera, es necesario tener claro qué es un modelo estadístico.

2.1. Modelo Estadístico

Se construye con información propia. Se puede construir de manera específica para distintos segmentos de la población. Se adquiere conocimiento y experiencia sobre su población, y habilidad en el diseño e interpretación de los resultados. Se conserva la confidencialidad de la información Núñez, G. (2011).

Teniendo en cuenta que la variable respuesta es dicotómica, pues clasifica riesgo alto (1) o bajo (0) de crédito de los clientes de una entidad financiera, el modelo que mejor puede explicar este problema es el modelo de regresión logístico, por esto, es importante definir qué es Regresión Logística, pero antes es necesario definir qué es probabilidad binomial:

2.2. Probabilidad Binomial

Sea p la probabilidad de éxito y q, la probabilidad de fracaso, de modo que p+q=1. Supóngase que un experimento consiste de N pruebas, cada prueba efectuada esencialmente bajo las mismas condiciones; podemos suponer entonces que el resultado de cada prueba es independiente de los otros resultados [Baisnab, A. P, & Jas, A. B. M (1993)].

2.3. Regresión Logística

El modelo de regresión logística surge cuando se quiere estimar la probabilidad de un evento dicotómico de Si (1) o No (0), en este caso riesgo Bajo (0) o riesgo Alto (1) (en función de un conjunto de variables predictoras comúnmente llamados factores de riesgo, que pueden ser discretas o continuas, categóricas (nominales u ordinales), cualitativa o cuantitativa.

La regresión logística analiza datos distribuidos binomialmente de la forma:

$$Y_i \sim B(p_i, n_i), para\ i = 1, \dots, m,$$

donde los números de ensayos Bernoulli n_i son conocidos y las probabilidades de éxito p_i son desconocidas.

Los logits de las probabilidades binomiales desconocidas son modeladas como una función lineal de los X_i.

El modelo es entonces obtenido a base de lo que cada ensayo (valor de i) y el conjunto de variables explicativas/independientes puedan informar acerca de la probabilidad final. Estas variables explicativas pueden pensarse como un vector X_i k-dimensional y el modelo toma entonces la forma:

$$logit(p_i) = ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

2.4. Credit Scoring

El credit scoring es una exitosa colección de técnicas estadísticas que se han utilizado para otorgar créditos en la industria del crédito [Simbaqueba (2004)].
La colección de técnicas que conforman el scoring tiene como propósito principal generar un puntaje de riesgo a las solicitudes de crédito o a cuentas ya existentes.

3. MARCO METODOLOGICO

Dada la dificultad de vinculación con el sector empresarial para utilizar datos reales, se presenta un caso de estudio con datos simulados de una base semilla sobre créditos de una entidad financiera, que pretende generar un modelo que clasifique el riesgo en Alto o Bajo de los clientes.

El desarrollo de la pregunta problema se abordó a través de un modelo estadístico de regresión logística (credit scoring), las variables de estudio son: datos de crédito (el historial crediticio de los clientes y su comportamiento en cuanto a la morosidad de pagos) y datos demográficos (edad, sexo, ingresos, situación laboral, etc).

Para el desarrollo del estudio, se requirió una base de datos con información demográfica, historial de préstamos y comportamiento en cuanto a la morosidad de pagos de los clientes de una entidad financiera. Se requirió de igual manera soportes teóricos que aporten en el diseño del modelo estadístico. Además, se empleó el uso de un equipo que cuenta con el software estadístico R.

En primera instancia, se realizó una revisión teórica, para establecer una definición de los conceptos requeridos para el estudio, de igual manera se revisó la información contenida en la base de datos "semilla". Una vez revisada la información, se simulo una base en el software R a través de una distribución normal, se tomó una parte de la base simulada como entrenamiento y la restante como base de prueba para generar el modelo, por lo que la variable riesgo fue generada desde un principio para luego contrastarla con el modelo clasificador-pronosticador, una vez que se estimaron las probabilidades o riesgo de morosidad. Luego los datos se depuraron excluyendo las variables con exceso de campos sin respuesta o respuestas múltiples. Posteriormente, se hace un ensayo de modelo logístico, el cual define las variables más significativas, las cuales se emplearon en el diseño del modelo estadístico. Ya determinadas las variables, se procedió a

definir el modelo a través de una herramienta estadística. Finalmente, se implementó el modelo diseñado y se concluyó sobre el riesgo de pago de los clientes a partir de los resultados obtenidos.

4. RESULTADOS

4.1. DESCRIPCION DE LA BASE DE DATOS SIMULADA

Se simula una base de datos de 5000 observaciones con 10 variables cuantitativas y cualitativas y una variable adicional de identificación de la observación.

X: Identificación de la observación

Min	Max
1	5000

Tabla 1

AGE: Edad

La edad de los clientes de la entidad financiera esta entre 18 a 67 años.

Min	Median	Max
18.00	33.00	67.00

Tabla 2

INCOME: Ingresos

Los ingresos de los clientes de la entidad financiera esta entre 15000 a 59200 diarios.

Min	Median	Max
15000	26110	59200

Tabla 3

GENDER_R: Género, donde 1 es mujer y 2 es hombre.

1	2
2564	2436

Tabla 4

MARITAL_R: Estado civil, donde 1 es divorciado (a), 2 es casado (a) y 3 es soltero (a).

1	2	3
1032	2691	1277

Tabla 5

NUMKIDS: Número de hijos.

0	1	2	3	4	5
878	1596	1589	725	195	17

Tabla 6

NUMCARDS: Numero de tarjetas.

0	1	2	3	4	5	6	7	8	9
483	909	1110	1005	758	466	203	49	13	4

Tabla 7

PAYMENT_R: Forma de pago, donde 1 es mensual y 2 es semanal.

1	2
2429	2571

Tabla 8

MORTGAGE_R: Hipoteca, donde 1 es no y 2 es sí.

1	2
1330	3670

Tabla 9

LOANS: Numero de préstamos con la entidad.

0	1	2	3	4
673	2073	1839	402	13

Tabla 10

RISK_R: Riesgo, donde 0 es bajo y 1 es alto.

0	1
1149	3851

Tabla 11

RESUMEN DE LA BASE DE DATOS

```
'data.frame':  5000 obs. of  10 variables:
 $ AGE       : int  25 32 39 34 30 35 33 42 41 18 ...
 $ INCOME    : int  49991 22209 23655 37143 15953 23059 41715 15492 41881 2687
8 ...
 $ GENDER_R  : Factor w/ 2 levels "1","2": 2 2 1 2 1 2 2 1 2 1 ...
 $ MARITAL_R : Factor w/ 3 levels "1","2","3": 2 2 1 3 2 1 3 2 3 3 ...
 $ NUMKIDS   : int  0 3 3 1 1 1 0 2 0 1 ...
 $ NUMCARDS  : int  0 2 5 0 4 2 0 4 1 1 ...
 $ PAYMENT_R : Factor w/ 2 levels "1","2": 1 2 2 1 2 2 1 2 1 1 ...
 $ MORTGAGE_R: Factor w/ 2 levels "1","2": 1 2 2 2 1 2 2 1 2 1 ...
 $ LOANS     : Factor w/ 5 levels "0","1","2","3",..: 2 2 3 3 2 3 2 3 1 3 ...
 $ RISK_R    : Factor w/ 2 levels "0","1": 0 1 1 1 1 1 0 1 1 1 ...
```

Tabla 12

4.2. MODELO BASE

Se genera el modelo logístico con la base de datos de entrenamiento y se observa que variables son más importantes, para esto se verifica en la tabla de coeficientes del resumen del modelo.

Este resultado muestra que las variables número de tarjetas y si tienen hipoteca no son tan importantes.

Modelo base:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.606e+00	3.663e-01	7.116	1.11e-12	***
AGE	-2.106e-02	6.256e-03	-3.366	0.000763	***
INCOME	-6.292e-05	6.493e-06	-9.690	< 2e-16	***
GENDER_R2	1.381e-01	9.212e-02	1.499	0.133911	
MARITAL_R2	4.406e-01	1.404e-01	3.138	0.001699	**
MARITAL_R3	4.531e-01	1.804e-01	2.511	0.012026	*
NUMKIDS	2.029e-01	5.883e-02	3.449	0.000564	***
NUMCARDS	-5.367e-02	3.597e-02	-1.492	0.135639	
PAYMENT_R2	1.754e-01	9.966e-02	1.760	0.078384	.
MORTGAGE_R2	-7.104e-02	1.096e-01	-0.648	0.516732	
LOANS1	3.210e-01	1.352e-01	2.375	0.017539	*
LOANS2	8.238e-01	1.676e-01	4.914	8.92e-07	***
LOANS3	1.065e+00	2.680e-01	3.973	7.09e-05	***
LOANS4	1.368e+01	3.009e+02	0.045	0.963735	

Tabla 13

Por lo anterior se corre el modelo sin las dos variables que son poco significativas.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.610e+00	3.479e-01	7.502	6.30e-14	***
AGE	-2.434e-02	5.860e-03	-4.154	3.27e-05	***
INCOME	-6.277e-05	6.479e-06	-9.687	< 2e-16	***
MARITAL_R2	4.619e-01	1.397e-01	3.307	0.000942	***
MARITAL_R3	4.843e-01	1.788e-01	2.709	0.006755	**
NUMKIDS	1.889e-01	5.787e-02	3.264	0.001098	**
PAYMENT_R2	1.599e-01	9.842e-02	1.624	0.104338	
LOANS1	3.160e-01	1.345e-01	2.351	0.018748	*
LOANS2	7.924e-01	1.637e-01	4.841	1.29e-06	***
LOANS3	9.987e-01	2.615e-01	3.819	0.000134	***
LOANS4	1.361e+01	3.012e+02	0.045	0.963947	

Tabla 14

Es necesario tener claras algunas definiciones, para realizar la verificación del modelo.

En Deride Silva, Julio (2010) define que es sensibilidad especificidad:

> *Sensibilidad:* Es la probabilidad de clasificar correctamente a un individuo cuyo estado real es definido como positivo, respecto a la condición de prueba.
>
> *Especificidad:* Es la probabilidad de clasificar correctamente a un individuo cuyo estado real es definido como negativo, respecto a la condición de prueba.

Dada una muestra, podemos estimar las probabilidades anteriores de la siguiente forma:

Sensibilidad = $\frac{número\ de\ verdaderos\ positivos}{número\ de\ positivos\ reales}$ = FVP

Especificidad = $\frac{número\ de\ verdaderos\ negativos}{número\ de\ negativos\ reales}$= FVN,

(FVP: fracción de verdaderos positivos y FVN: fracción de verdaderos negativos).

4.2.1. ENSAYO DEL MODELO DE REGESION LOGISTICO CON LA VARIABLE RIESGO

Ensayo del modelo con la variable riesgo:

Modelo / Base	0	1
0	48	291
1	41	1110

Tabla 15

Exactitud	77%
Sensibilidad	14%
especificidad	96%

A pesar de que la exactitud y la especificidad del modelo manejan un porcentaje mayor al 60%, no clasifica adecuadamente los clientes, es decir, que los clientes que están clasificados en la base simulada con bajo riego, en el modelo se están clasificando de la siguiente manera: 48 con bajo riesgo y 291 con alto riesgo, por lo que se pierden muchos clientes buenos.

Por la anterior es necesario aplicar el método de balanceo de muestras para encontrar un mejor modelo.

4.3. MUESTRA BASE ENTRENAMIENTO Y BASE DE PRUEBA

Se seleccionó la base de entrenamiento aleatoriamente, este corresponde al 60% del total de la base, para ello se crea una nueva variable la cual es X y sirve para identificar cada observación, la base se dividió en: 60% de entrenamiento y 40% de prueba.

División de base:

ENTRENAMIENTO	PRUEBA
3000	2000

Tabla 16

4.4. BALANCEO DE LA MUESTRA ENTRENAMIENTO

Dado que de la muestra de entrenamientos se divide en:

RISK_R- ENTRENAMIENTO: Riesgo, donde 0 es bajo y 1 es alto:

0	1
682	2318

Tabla 17

Se utilizó la técnica de oversampling (sobremuestreo) el cual consiste, en utilizar un muestreo simple para igualar la cantidad de datos de las dos variables en la base de entrenamiento, con lo que se obtiene:

RISK_R- ENTRENAMIENTO-2: Riesgo, donde 0 es bajo y 1 es alto:

Oversampling:

0	1
2318	2318

Tabla 18

4.5. MODELO DE ENTRENAMIENTO DE REGRESION LOGISTICA

Se genera el modelo logístico con la nueva base de datos de entrenamiento y se observa que variables son más importantes, para esto se verifica en la tabla de coeficientes del resumen del modelo.

Este resultado muestra que las variables número de tarjetas y forma de pago no son tan importantes.

Tabla de coeficientes del modelo logístico de entrenamiento.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.407e+00  2.529e-01   5.564 2.64e-08 ***
AGE         -2.978e-02  4.495e-03  -6.627 3.43e-11 ***
INCOME      -5.900e-05  4.731e-06 -12.471  < 2e-16 ***
GENDER_R2    1.587e-01  6.363e-02   2.494 0.012629 *
MARITAL_R2   4.347e-01  9.742e-02   4.462 8.13e-06 ***
MARITAL_R3   3.985e-01  1.237e-01   3.222 0.001273 **
NUMKIDS      1.769e-01  3.933e-02   4.497 6.89e-06 ***
NUMCARDS     1.498e-02  2.545e-02   0.589 0.556034
PAYMENT_R2   2.472e-01  6.792e-02   3.640 0.000272 ***
MORTGAGE_R2 -6.617e-02  7.448e-02  -0.888 0.374294
LOANS1       3.368e-01  9.792e-02   3.440 0.000582 ***
LOANS2       8.275e-01  1.181e-01   7.004 2.49e-12 ***
LOANS3       9.321e-01  1.783e-01   5.227 1.72e-07 ***
LOANS4       1.394e+01  1.637e+02   0.085 0.932150
```

Tabla 19

Por lo que se genera un nuevo modelo logístico con las variables importantes.

Tabla de coeficientes del nuevo modelo logístico.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.407e+00  2.529e-01   5.564 2.64e-08 ***
AGE         -2.978e-02  4.495e-03  -6.627 3.43e-11 ***
INCOME      -5.900e-05  4.731e-06 -12.471  < 2e-16 ***
GENDER_R2    1.587e-01  6.363e-02   2.494 0.012629 *
MARITAL_R2   4.347e-01  9.742e-02   4.462 8.13e-06 ***
MARITAL_R3   3.985e-01  1.237e-01   3.222 0.001273 **
NUMKIDS      1.769e-01  3.933e-02   4.497 6.89e-06 ***
PAYMENT_R2   2.472e-01  6.792e-02   3.640 0.000272 ***
LOANS1       3.368e-01  9.792e-02   3.440 0.000582 ***
LOANS2       8.275e-01  1.181e-01   7.004 2.49e-12 ***
LOANS3       9.321e-01  1.783e-01   5.227 1.72e-07 ***
```

Tabla 20

4.6. MODELO DE REGRESION LOGISTICO FINAL

m2<-RISK_R ~ AGE + INCOME + GENDER_R + MARITAL_R + NUMKIDS + PAYMENT_R + LOANS

Tabla de coeficientes del resumen del modelo logístico final.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.347e+00  2.437e-01   5.529 3.23e-08 ***
AGE         -2.874e-02  4.208e-03  -6.830 8.50e-12 ***
INCOME      -5.911e-05  4.728e-06 -12.500  < 2e-16 ***
GENDER_R2    1.560e-01  6.355e-02   2.455 0.014097 *
MARITAL_R2   4.347e-01  9.658e-02   4.501 6.76e-06 ***
MARITAL_R3   3.944e-01  1.220e-01   3.232 0.001228 **
NUMKIDS      1.803e-01  3.899e-02   4.623 3.78e-06 ***
PAYMENT_R2   2.566e-01  6.718e-02   3.819 0.000134 ***
LOANS1       3.406e-01  9.773e-02   3.485 0.000492 ***
LOANS2       8.451e-01  1.164e-01   7.263 3.80e-13 ***
LOANS3       9.662e-01  1.743e-01   5.544 2.95e-08 ***
LOANS4       1.401e+01  1.639e+02   0.086 0.931848
```

Tabla 21

MODELO FINAL

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

X_1: AGE

X_2: INCOME

X_3: GENDER_R2

X_4: MARITAL_R2

X_5: MARITAL_R3

X_6: NUMKIDS

X_7: PAYMENT_R2

X_8: LOANS1

X_9: LOANS2

X_{10}: LOANS3

X_{11}: LOANS4

Donde, $\beta_0, \beta_1, \ldots, \beta_k,$ son los interceptos de las variables, y $X_1, X_2, \ldots, X_K,$ son los valores de las variables en una determinada observación.

Entonces el modelo que clasifica el riesgo es:

$logit$ =1,347 - 0,02874*X_1 - 0,00005911*X_2 + 0,156*X_3 + 0,4347*X_4 + 0,3944*X_5 + 0,1803*X_6 + 0,2566*X_7 + 0,3406*X_8 + 0,8451*X_9 + 0,9662*X_{10} + 1,401*X_{11}

Para verificar si el modelo es Bueno gráficamente y teniendo los valores de especificidad y sensibilidad se halla la curva Roc.

> *Curva Roc:* Para cada valor de c se encuentra un par (Sensibilidad(c), 1−Especificidad(c)), los cuales definen una curva. Esta curva es la llamada Curva ROC. Para estimaciones muestrales, la curva ROC se construirá variando el nivel de corte de la variable en estudio y estimando con los pares (FVP, FFP).

El análisis de la curva ROC se hace mediante la comparación del área bajo la curva. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin valor diagnóstico. Gráficamente puede suceder alguno de estos casos:

Casos curva Roc:

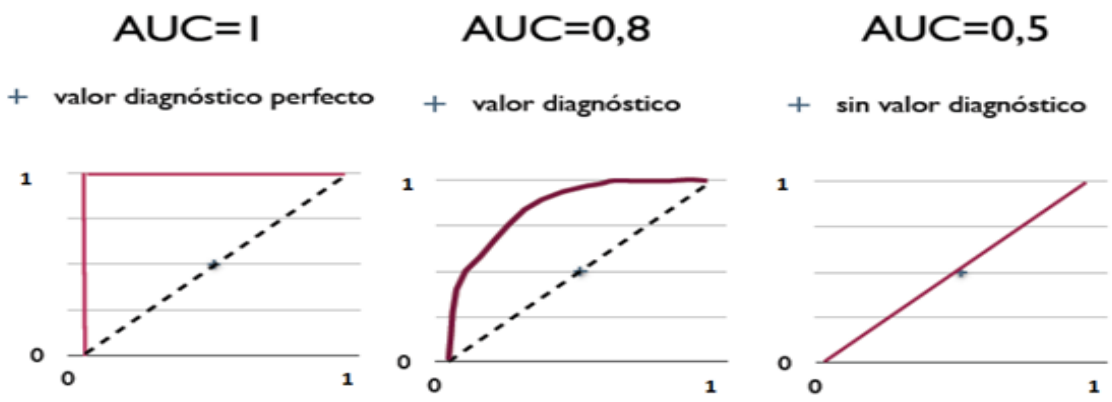

Ilustración 1

4.7. ENSAYO DEL MODELO DE REGRESION LOGISTICO FINAL CON MUESTRA DE PRUEBA

Se ensaya el modelo con la base de datos de prueba.

Base \ Modelo	0	1
0	283	163
1	549	1005

Tabla 22

Exactitud 64%
Sensibilidad 63%
Especificidad 65%

4.8. CALCULO DE CURVA ROC

Curva Roc del Modelo:

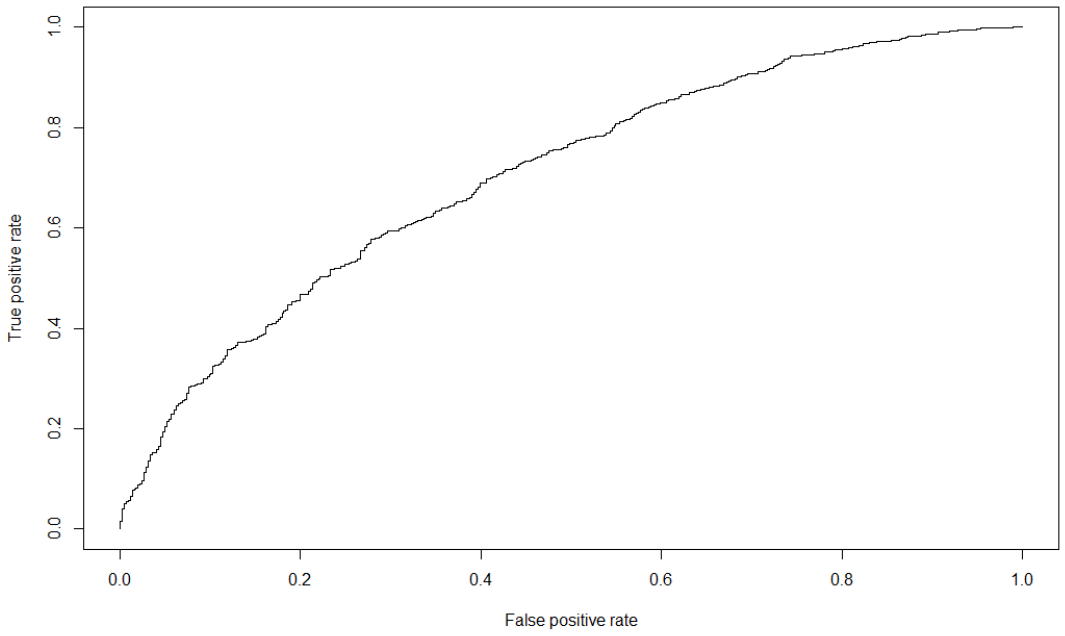

Ilustración 2

Existe un 64% de probabilidad de que el estudio de crédito realizado a un cliente sea más correcto.

Gráficamente se observa que el área bajo la curva es mayor al 0,5, con lo cual se dice que el modelo explica en un buen porcentaje la clasificación del riesgo al otorgar un crédito a un cliente o nuevo cliente.

5. DISCUSIÓN

Prueba del modelo con un caso de la base simulada.

Aleatoriamente se toma la identificación 4221.

	AGE	INCOME	GENDER_R	MARITAL_R	NUMKIDS	NUMCARDS	PAYMENT_R	MORTGAGE_R	LOANS	RISK_R
4221	27	15001	2	2	0	2	2	2	1	1

Tabla 23

$logit$ =1,347 - 0,02874*X_1 - 0,00005911*X_2 + 0,156*X_3 + 0,4347*X_4 + 0,3944*X_5 + 0,1803*X_6 + 0,2566*X_7 + 0,3406*X_8 + 0,8451*X_9 + 0,9662*X_{10} + 1,401*X_{11}

$logit$ =1,347 - 0,02874*27 - 0,00005911*15001 + 0,156*1 + 0,4347*1 + 0,3944*0 + 0,1803*0 + 0,2566* 1 + 0,3406*1 + 0,8451*0 + 0,9662*0 + 1,401*0

$logit = 0{,}87$

Como el valor del $logit$ se aproxima a 1, se verifica la información de la base simulada, es decir que el cliente tiene alto riesgo, es decir, no es bueno realizar préstamos a este cliente.

Ejemplo con cliente en potencia:

Un hombre con las siguientes características desea solicitar un crédito:

Ingresos diarios: 22000, Edad: 26 años, Estado civil: casado, Número de hijos: 0, Número de tarjetas: 3, Forma de pago: Mensual, Hipoteca: Si, Créditos con la entidad: 0

AGE	INCOME	GENDER_R	MARITAL_R	NUMKIDS	NUMCARDS	PAYMENT_R	MORTGAGE_R	LOANS
26	22000	2	2	0	3	1	2	0

Tabla 24

$logit$ =1,347 - 0,02874*X_1 - 0,00005911*X_2 + 0,156*X_3 + 0,4347*X_4 + 0,3944*X_5 + 0,1803*X_6 + 0,2566*X_7 + 0,3406*X_8 + 0,8451*X_9 + 0,9662*X_{10} + 1,401*X_{11}

$logit$ =1,347 - 0,02874*26 - 0,00005911*22000 + 0,156*1 + 0,4347*1 + 0,3944*0 + 0,1803*0 + 0,2566* 0 + 0,3406*0 + 0,8451*0 + 0,9662*0 + 1,401*0

$logit$ =- 0,11

Aunque este hombre no haya tenido créditos con la entidad, el modelo recomienda otorgar el crédito,

6. CONCLUSIONES

Se logró determinar las variables significativas para generar el modelo a través del modelo logit inicial, estas son: edad, ingresos diarios, estado civil, número de hijos, número de créditos con la entidad, mientras que las variables hipoteca y forma de pago resultaron no significativas.

Se mostró que un modelo de regresión logística para pronosticar el otorgamiento de créditos a clientes o futuros clientes se desempeña adecuadamente como modelo clasificador sobre una base de datos simulada; en consecuencia, se asume factible que este tipo de modelos sea adecuado sobre casos reales o similares.

La Curva Roc mostró que en general el modelo es un buen predictor, ya que el porcentaje de exactitud es mayor al 50%

La metodología aquí presentada puede ser usada en futuros trabajos de clasificación de riesgo de crédito, para entidades financieras con características similares al estudio realizado.

Se contrasto el modelo logístico con un caso aleatorio de la base simulada, y se obtuvo que el modelo se ajusta a esta. De igual manera se evaluó el modelo con un cliente en potencia.

7. LISTA DE REFERENCIAS

Baisnab, A. P, & Jas, A. B. M. (1993). *Elements of probability and statistics*. Tata McGraw-Hill Education.

González, R. "Cada día caen en cartera vencida unos 3 mil 305 préstamos al consumo", *La jornada*, 11 de diciembre de 2008. Recuperado el 28 de febrero de 2015 de: http://www.jornada.unam.mx/2008/12/11/index.php?section=economia&article=029n1eco

Miranda, 2013. C. A. L. MODELO PREDICTIVO DE RIESGO DE MOROSIDAD PARA CRÉDITOS BANCARIOS USANDO DATOS SIMULADOS.

Montgomery, C., Peck, E. & Vining, G. (2006). *Introducción al Análisis de Regresión Lineal*. (3) (3). México: Compañía Editorial Continental.

Navarro Castaño, D. (2003). *Temas de: Administración Financiera.* Universidad Nacional de Colombia Sede Manizales.

Nieto, S. (2010). *Crédito al Consumo: La Estadística aplicada a un problema de Riesgo Crediticio*. Tesis de maestría. Universidad Autónoma Metropolitana, México, D.F.

Núñez, G. (2011). Crédito al Consumo: La Estadística aplicada a un problema de Riesgo Crediticio. *Actuarios Trabajando: Revista Mexicana de Investigación Actuarial Aplicada*. 4(6), 1-109

Simbaqueba, Lilian. *¿Qué es el scoring? Una visión práctica de la gestión del riesgo de crédito*. Instituto del Riesgo Financiero, Bogotá, 2004.

8. ANEXOS

Base de datos simulados para el modelo de Regresión logístico (Credit scoring) .La siguiente tabla contiene la base de datos para 5000 clientes obtenida mediante simulación para generación del modelo. El significado de cada variable en esta tabla es como se definió en el capítulo 4, numeral 4.1. Descripción de la base de datos simulada.

Base simulada:

	AGE	INCOME	GENDER_R	MARITAL_R	NUMKIDS	NUMCARDS	PAYMENT_R	MORTGAGE_R	LOANS	RISK_R
1	25	49991	2	2	0	0	1	1	1	0
2	32	22209	2	2	3	2	2	2	1	1
3	39	23655	1	1	3	5	2	2	2	1
4	34	37143	2	3	1	0	1	2	2	1
5	30	15953	1	2	1	4	2	1	1	1
6	35	23059	2	1	1	2	2	2	2	1
7	33	41715	2	3	0	0	1	2	1	0
8	42	15492	1	2	2	4	2	1	2	1
9	41	41881	2	3	0	1	1	2	0	1
10	18	26878	1	3	1	1	1	1	2	1
11	34	35482	1	2	1	2	2	1	1	1
12	33	22532	2	1	2	3	2	2	2	1
13	28	34337	2	3	0	0	1	2	0	1
14	40	21085	2	1	3	4	1	1	3	0
15	46	28285	1	1	2	6	2	1	2	1
16	38	31293	2	2	2	3	2	1	1	1
17	28	26523	1	2	2	2	2	2	1	1
18	56	30700	1	1	4	9	2	2	3	0
19	32	26809	1	1	2	3	1	2	2	1
20	34	38019	1	2	1	3	2	2	1	0
21	30	29978	2	3	1	0	1	1	1	0
22	33	38381	1	3	0	0	1	2	0	1
23	27	24150	1	3	2	1	2	2	1	1
24	39	22534	1	1	3	5	1	2	2	1
25	32	22760	2	1	2	2	1	2	1	1
26	28	15391	1	2	1	1	2	2	2	1
27	19	22980	1	3	1	4	2	2	2	1
28	20	35260	1	3	0	0	1	2	1	0

29	30	25596	1	3	1	1	2	2	1	1
30	35	26363	1	2	1	4	2	1	2	1
31	36	21267	2	3	2	2	2	2	1	0
32	44	23705	1	2	1	1	1	2	2	1
33	39	33127	2	2	3	5	1	2	2	1
34	20	33738	1	2	1	0	1	2	1	1
35	40	54464	2	2	0	2	1	1	1	1
36	44	28984	2	2	3	5	1	2	3	1
37	36	21301	2	1	4	5	1	2	2	1
38	31	19778	2	2	2	2	2	2	1	1
39	35	30660	2	3	1	2	1	2	2	0
40	41	27642	2	1	3	4	2	2	1	1
41	19	33207	1	3	2	0	1	2	0	0
42	26	32118	2	3	1	0	1	2	0	0
43	28	31372	1	2	2	1	1	2	0	1
44	38	36873	1	2	1	2	1	1	2	0
45	42	31793	2	2	2	3	2	2	1	0
46	35	29379	2	2	1	3	2	2	2	1
47	46	33770	1	1	2	3	2	2	2	1
48	21	21318	1	2	0	1	2	2	1	0
49	24	32871	1	3	2	0	1	1	1	0
50	22	37433	1	3	0	2	1	2	0	0
51	29	26395	2	1	1	3	2	2	2	0
52	39	39056	1	2	2	4	2	1	2	0
53	38	29104	2	3	1	0	1	2	1	1
54	23	27740	2	1	2	5	2	2	2	1
55	39	25893	2	2	2	4	2	1	2	1
56	19	16526	1	2	1	0	1	2	1	1
57	23	24121	2	2	2	3	2	2	0	1
58	56	31744	1	2	3	7	1	1	1	1
59	39	22997	1	1	2	7	2	1	2	1
60	63	44840	2	1	3	6	2	2	1	0
61	21	24711	1	3	0	0	2	1	1	1
62	42	36390	1	2	2	4	2	2	2	1
63	40	25864	1	1	2	4	2	2	2	0
64	36	20224	1	1	4	6	2	2	2	1
65	26	18761	2	3	1	1	1	2	1	1
66	53	22185	2	2	2	5	1	2	2	1
67	18	19456	1	3	1	2	2	2	1	1
68	36	18438	2	2	3	4	2	2	2	1
69	28	18985	1	2	2	4	1	1	2	1
70	44	25720	1	3	0	3	2	2	0	1
71	30	19901	1	2	0	3	2	1	2	1

72	22	33024	2	3	0	0	1	1	0	0
73	40	25752	1	2	1	4	2	2	2	1
74	23	40693	1	2	2	1	2	2	2	1
75	23	20333	2	3	1	3	2	2	1	0
76	25	19000	1	2	1	3	2	2	1	1
77	43	34876	1	2	1	4	1	1	1	1
78	29	27506	2	2	2	2	1	2	1	1
79	27	25758	2	2	1	3	2	2	2	1
80	45	25072	1	1	1	2	2	2	2	1
81	35	39288	2	2	1	1	2	1	2	1
82	38	18794	1	2	3	4	2	2	2	1
83	29	31192	1	2	1	3	1	2	0	0
84	40	24577	2	2	2	2	1	2	2	0
85	21	37248	1	3	1	2	1	2	0	1
86	33	23855	1	2	2	4	2	1	1	1
87	44	23388	1	2	2	2	2	2	2	1
88	39	32088	1	2	1	3	2	2	1	1
89	28	27409	2	2	1	2	1	2	1	1
90	35	25190	2	3	3	4	2	2	3	1
91	23	21704	1	3	2	0	1	2	0	0
92	47	23461	2	2	3	1	2	2	2	1
93	33	23170	1	2	3	4	1	1	2	1
94	32	25700	1	3	1	2	1	2	1	1
95	33	34180	2	2	2	2	2	2	3	1
96	20	15566	2	3	0	2	1	1	1	1
97	38	36059	1	2	1	2	1	2	1	0
98	29	25243	2	3	1	3	2	2	0	0
99	29	24230	2	2	0	3	1	1	1	1
100	21	55897	2	3	2	2	1	2	1	1
…										
5000	34	32355	1	2	1	1	2	1	1	1

SCRIP EN R

```
step(modelo,direction = "backward")
d<-sort(sample(nrow(datos), nrow(datos)*.6))

muestra<-datos[d,]
dim(muestra)
prop.table(table(muestra$RISK_R))

prueba<-datos[-d,]
dim(prueba)
prop.table(table(prueba$RISK_R))
table(muestra$RISK_R)

muestra_1<-muestra[muestra$RISK_R==1,]
muestra_2<-muestra[muestra$RISK_R==2,]
id_muestra_1<-muestra2<-sample(x = 1:nrow(muestra_1),size = nrow(muestra_2),re
place = T)

muestra_1<-muestra_1[id_muestra_1,]


muestra2<-rbind(muestra_1,muestra_2)
table(muestra2$RISK_R)

modelo<-glm(RISK_R~.,data = muestra2[,1:10],family = binomial)
summary(modelo)

paso<-step(modelo,direction = "backward")

m2<-RISK_R ~ AGE + INCOME + GENDER_R + MARITAL_R + NUMKIDS + PAYMENT_R + LOANS
modelo2<-glm(m2,data = muestra2[,1:10],family = binomial)
modelo2
summary(modelo2)

table(prueba$RISK_R,round(predict(object = modelo2,newdata = prueba,type = "re
sponse")))

pred<-predict(object = modelo2,type="response",prueba)
pred<-prediction(pred,prueba$RISK_R)
perf<-performance(pred,"tpr","fpr")
plot(perf)
```