"DISEÑO E IMPLEMENTACION DE UN SISTEMA DE PREDICCION DEL RIESGO CREDITICIO DE LAS SOLICITUDES DE INSTALACION DE UNA EMPRESA DE DISTRIBUCION ELECTRICA"

AUTORES:

Elizabeth Agila Soto[1], José Dávila Llerena[2], Maria Ramírez Macancela[3], Juan Alvarado[4]

[1] Egresada de Licenciada en Sistemas de Información 2006.
[2] Egresada de Licenciado en Sistemas de Información 2006.
[3] Egresada de Licenciada en Sistemas de Información 2006.
[4]Director de Tópico, Ingeniero en Electricidad y Computación, especialidad en computación 16 Noviembre 1990 ESPOL, Profesor de LSI-FIEC ESPOL.

RESUMEN

Versión español:

El objetivo principal de este trabajo ha sido el de obtener una herramienta para estimar la probabilidad de riesgo de morosidad de un cliente, a fin de observar y predecir su capacidad de pago ante las variaciones de parámetros que configuran sus políticas operacionales, de forma que se pueda evaluar bajo distintas consideraciones antes de conceder un medidor de energía eléctrica o crédito. El sistema permite evaluar las variables típicas ó esenciales que determinan la capacidad de pago de un cliente como por ejemplo, ingresos netos, cargas familiares, indicadores socio-económicos, etc.

English version:

The main objective of this work has been the one of obtaining a tool to estimate the probability of risk of slowness of a client, in order to observe and to predict your capacity of payment before the variations of parameters that its operational politicians configure, so that being can evaluate under different considerations before to give a energy electric meter ó credit. The system allows evaluate the variables typical or essential that determine the capacity of payment of client for example net revenue, relative charge, socio-economics indicators, etc.

INTRODUCCIÓN

El proceso de globalización exige hacer cambios radicales que aseguren la supervivencia de las organizaciones y su liderazgo en el mercado, por medio de la adquisición de competencias tales como, reducción de costos, alta calidad, manejo adecuado y oportuno de la información.

Los mismos que nos permitan la toma de decisiones acertadas. Estas decisiones deben sustentarse en sistemas de información de alta calidad, confiables y seguros que ofrezcan un excelente tiempo de respuesta.

El riesgo de crédito ha sido tradicionalmente la incertidumbre más significativa que las entidades comerciales y financieras asumen como consecuencia de su actividad.

Los efectos de la posible insolvencia de sus clientes justifican la necesidad de desarrollar herramientas de evaluación de la capacidad para afrontar sus deudas.

Aquí, surge la construcción de un modelo de minería de datos como una alternativa para superar los inconvenientes antes mencionados y otros en particular.

El estudio que se presenta en el desarrollo de esta tesis tiene como objetivo construir un modelo de minería de datos para evaluar la probabilidad de capacidad de pago de un cliente nuevo.

CONTENIDO

1. DESCRIPCION DEL NEGOCIO

Determinar la probabilidad de pago de un cliente nuevo para una empresa comercializadora de energía eléctrica, basado en un histórico de clientes, con datos tales como: número de cargas familiares, nivel socio-económico, edad, sector, sueldo, promedio de consumo, y otra información relevante que nos ayudaran a encontrar una mejor predicción del modelo a desarrollar.

Misión, satisfacer la demanda, negocios afines y fuentes no convencionales de energía eléctrica, de manera ágil, confiable y continua a todos los clientes en el área servida. Preservar el medio ambiente y contribuir al desarrollo socioeconómico del país con un recurso humano comprometido y altamente calificado. Estos servicios se ofrecerán con tarifas competitivas, mediante la incorporación de tecnología de punta reafirmando nuestros principios y valores corporativos.

2. METODOLOGIA DE MINERIA DE DATOS.

El descubrimiento de conocimiento en base de datos (KDD) combina las técnicas tradicionales con numerosos recursos desarrollados en el área de la inteligencia artificial. En estas aplicaciones el término "Minería de Datos" (Data mining) ha tenido más aceptación.

En algunos casos las herramientas provenientes de la inteligencia artificial son nuevas, no del todo comprendidas y carentes de un soporte teórico formal. Pero en este caso el objetivo es tan valioso, que los resultados prácticos han rebasado a la elegancia académica.

• *Identificando el problema del negocio.* El Ciclo virtuoso de data mining empieza identificando las oportunidades del negocio. Desafortunadamente, ha demasiados estadísticos buenos y analistas competentes cuyo trabajo es esencialmente desperdiciado porque están resolviendo problemas que no ayudan a los negocios. Una buena minería de datos quiere evitar esta situación.

• *Aplicar Minería de Datos para transformar los datos en Información para la toma de decisiones*. Data Mining, la extracción de información oculta y predecible de grandes bases de datos, es una poderosa tecnología nueva con gran potencial para ayudar a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse).

• *Actuar sobre la información*, Una vez que se ha descubierto patrones escondidos en la información, es momento de actuar sobre ésta. Con los resultados obtenidos los tomadores de decisiones podrán tener respuestas anticipadas de situaciones riesgosas.

• *Medir los resultados*, Para medir los resultados se puede aplicar la minería de datos sobre información de un año anterior de la cual ya se sabe cual fue el comportamiento posterior, luego comparar los resultados obtenidos con la minería de datos con la información actual y determinar si las respuestas son muy parecidas. De ésta manera se prueba el modelo antes de utilizarlo en una situación de riesgo.

3. DESCRIPCION DE LOS DATOS

En este capitulo se proveerá una vista general de cómo son los datos, cual es la información relevante a utilizar, información externa o complementaria como indicadores socio-económicos, que ayuden a mejorar la predicción en el modelo de minería a utilizar.

• *Variables con información relevante,* Llamamos variables con información relevante a aquellas variables que contienen información propia del cliente, tales como el código de la provincia, cantón, parroquia y sector que son variables que nos ayudaran a determinar el estatus socioeconómico que el cliente posee, cuya información nos facilitara un mejor análisis de poder pago.

• *Variables de indicadores socio-económicos,* La información y el análisis socioeconómicos son una poderosa herramienta para evaluar los datos relevantes que vamos a evaluar para obtener un resultado muy cercano a la realidad. Los indicadores sociales, son medidas pensadas para reflejar cómo viven las personas. Otra característica común a los indicadores es su intención de resumir un gran cúmulo de datos cuantitativos o estadísticas.

• *Variables de Evaluación,* esta variables las comprende la variable dicotómica que determina si el cliente es bueno o malo, Y=0 ó Y=1 y la variable de probabilidad que determina la capacidad de pago de un cliente.

4. PREPROCESAMIENTO DE LOS DATOS.

Este capitulo tiene como Objetivo general de predecir resultados y/o descubrir relaciones en los datos. Esto puede ser descriptivo, y podemos descubrir patrones que describen los datos o predictivo para pronosticar el comportamiento del modelo basado en los datos disponibles.

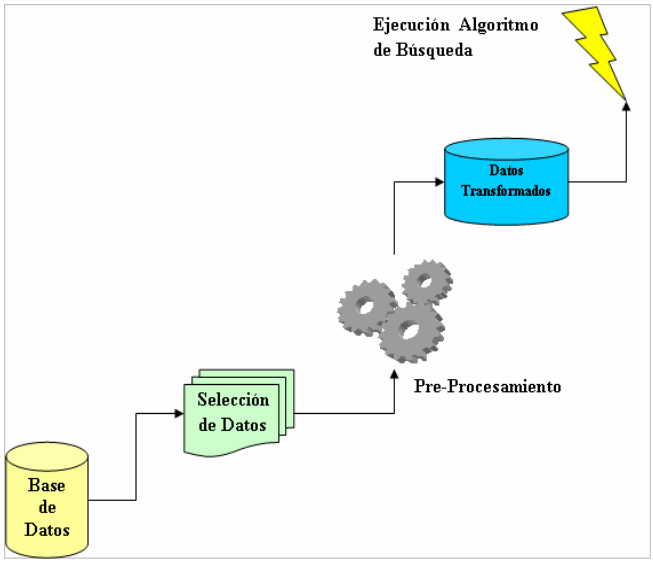

Figura 1. Proceso de predicción de riesgo crediticio

• *Selección de los datos*, a partir del proceso de conocimiento de los datos se procede a la selección de los datos con mayor relevancia para la predicción del modelo a desarrollar, en el caso de predicción de posibles clientes morosos, los datos mas importantes utilizados son lo que tienen que ver con las variables socio-económicas: edad, estado civil, número de hijos, nivel de ingresos, categoría laboral, historial de pagos, etc. En este caso, su principal aplicación se centra en créditos.

• *Pre-procesamiento*, en esta etapa se procede con la limpieza de los datos, detección de datos aberrantes. Análisis por medio de gráficos de distribución de frecuencias, con la ayuda de programas como WEKA o SPSS para el análisis de la muestra, evaluando que sea completamente aleatoria y que no contenga ningún sesgo definido.

• *Normalización*, este proceso se aplica para aquellas variables con valores o rangos muy elevados, se estandariza dentro de un rango, se aplica a cada valor: $(Xi– \mu) / \sigma$.

• *Asociación de Registros con Indicadores Socio-Económicos*, son consideramos de elevada importancia los indicadores socio-económicos de los clientes nos ayudara a encontrar una probabilidad más exacta de si el cliente caerá o no en mora.

• *Conociendo los datos*, para conocer como esta la muestra seleccionada podemos utilizar distribuciones de frecuencias o histogramas, gráficos que nos ayudan a visualizar el comportamiento de los datos.

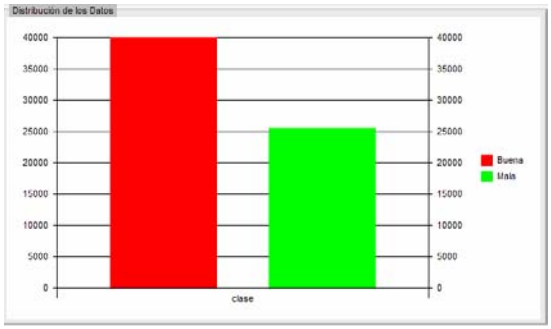

Figura 2. Distribución de Frecuencias

5. SELECCIÓN Y EVALUACION DEL MODELO

• *Selección del modelo*, entre las metodologías de medición del riesgo crediticio destacan los modelos scoring, basados en técnicas estadísticas. En la valoración empírica de la morosidad se puede aplicar la técnica de regresión logística, que permite determinar los factores que influyen en el comportamiento de pago de los clientes

• *El riesgo crediticio y su estimación*, supone la posibilidad de sufrir una pérdida como consecuencia del incumplimiento de la contrapartida al asumir lo acordado, bien sea por incapacidad de ésta o por falta de disposición, en tiempo o en forma. El ejemplo más sencillo es la no recuperación de las cuotas e intereses de un convenio de pago concedido. La existencia de riesgo crediticio depende entonces de la solvencia y compromiso del deudor, pero su magnitud está relacionada con el volumen de la operación.

• *Valoración de la probabilidad de insolvencia*, son aspectos a analizar ante la petición de un crédito: carácter u honradez del deudor (poder de pago), capacidad de generación de fondos (ingresos), cargas familiares y condiciones del entorno (índices socioeconómicos).

• *El método scoring*, trata de un sistema de calificación de créditos para automatizar la toma de decisiones antes de conceder o no un la concesión de un crédito, ya sea a una persona física o jurídica. La validez del método depende en buena medida de la calidad de la información disponible. Cuando el análisis se refiere a personas jurídicas, se parte de los estados contables, a partir de los cuales se elaboran ratios de liquidez, autofinanciación, rentabilidad económica, rotación, volumen de activo.

• *Análisis empírico*, con el fin de analizar y valorar la morosidad como forma de manifestación del riesgo de crédito, desde una perspectiva empírica, hemos realizado un estudio, aplicando la técnica de regresión logística a una muestra de clientes, para seleccionar, a partir de la misma, aquellos factores que tienen mayor influencia en el comportamiento de pago de los clientes y, por consiguiente, en su posible insolvencia. Ello implica la consideración de aspectos tales como la selección de las variables a incluir en el modelo y de la

muestra objeto de estudio, así como la presentación de los principales resultados obtenidos.

• *Selección de Variables*, el fenómeno a explicar es la morosidad o insolvencia de los clientes, por lo que se tratará de determinar si se trata de un cliente moroso o no, siendo éstas las dos modalidades de la variable dependiente o respuesta que modela el fenómeno de la morosidad. En concreto, en el análisis de regresión logística, las modalidades de la variable se interpretan como la ocurrencia y no ocurrencia del acontecimiento que se analiza, que se codifican con los valores uno y cero, respectivamente. Las variables independientes seleccionadas para su posible inclusión en el modelo se han agrupado en dos bloques: variables relativas a la operación de crédito (consumo, infracciones, meses deuda, numero de veces cortado, numero de veces que ha sido multado, meses de deuda y valor vencido) y variables relativas al perfil del solicitante (edad, estado civil, cargas familiares, ingresos, residencia, propiedad de vivienda).

• *Muestra de Clientes*, la muestra de clientes se ha seleccionado tomando una muestra aleatoria de clientes, resultando un total de 25543 clientes morosos y 39992 no morosos. Como se observa, la muestra seleccionada recoge una proporción elevada de clientes morosos en comparación con los no morosos, lo que nos obliga a tomar con precaución de los resultados obtenidos.

• *Análisis de resultados*, una vez estimado el modelo, el siguiente paso es valorar la significación individual de cada coeficiente, aspecto que se ha considerado para la inclusión de las variables en el mismo. Asimismo, se evalúa la bondad de ajuste del modelo, obteniéndose el estadístico de razón de verosimilitud.

• *Evaluación del Modelo*, no cabe ninguna duda que la regresión logística es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en investigación clínica y epidemiología, de ahí su amplia utilización. El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías (politómico).

• *Los coeficientes del modelo logístico como cuantificadores de riesgo,* una de las características que hacen tan interesante la regresión logística es la relación que éstos guardan con un parámetro de cuantificación de riesgo conocido en la literatura como "odds ratio" (aunque puede tener traducción al castellano, renunciamos a ello para evitar confusión ya que siempre se utiliza la terminología inglesa). El odds asociado a un suceso es el cociente entre la probabilidad de que ocurra frente a la probabilidad de que no ocurre.

$$Prob(Y) = \frac{1}{1 - e^{(\beta_0 + \beta_1 X_1 + \ldots\ldots \beta_n X_n)}}$$

6. DISEÑO E IMPLEMENTACION

Las Base de Datos, fue creada en SQL Server 2000, en ella se almacenan todos los datos requeridos para el procesamiento del modelo desarrollado, así como también los resultados de la misma.

La aplicación, es decir la interfaz de usuario fue desarrollada en Visual Basic 6.0, este componente comprende la interacción de todos sus componentes, hace de intérprete entre el software de predicción y el usuario y a su vez el manejo de información en la base de datos

La aplicación desarrollada permitirá conocer el riesgo crediticio de aceptar un cliente nuevo. La misma que permitirá la carga inicial de los datos, el pre-procesamiento, la aplicación del modelo de regresión logística en base a datos históricos.

7. CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

Una vez identificado el problema del negocio, la alta morosidad de los clientes, se procedió a seleccionar los datos que serían de mayor relevancia para el caso analizado. Los datos más importantes utilizados son lo que tiene que ver con las variables socio-económicas: edad, estado civil, número de cargas familiares, nivel de ingresos y pagos vencidos.

En la etapa del pre-procesamiento se procedió con la limpieza de los datos, detección de datos aberrantes y valores nulos. Este análisis se lo realizó por medio de gráficos de distribución de frecuencias, con la ayuda de programas como WEKA o SPSS para el análisis de la muestra, evaluando que sea completamente aleatoria y que no contenga ningún sesgo definido.

Para evitar desbordamientos en la evaluación del modelo se ha normalizó las variables de sueldo y valor vencido.

La muestra de clientes se ha seleccionado tomando una muestra aleatoria de clientes, resultando un total de 25543 clientes morosos y 39992 no morosos. Como se observa, la muestra seleccionada recoge una proporción elevada de clientes morosos en comparación con los no morosos, lo que nos obliga a tomar con precaución de los resultados obtenidos, ya que pueden no ser extrapolables a la población de clientes de la que se ha extraído la muestra.

El modelo seleccionado para nuestro caso fue el de regresión logística con variable dependiente de tipo dicotómica. Esta variable es la variable “clase”, la misma que identifica si un cliente paga o no paga. Lo que se obtuvo a partir de esta variable y el resto de variables involucradas, es la probabilidad de que un cliente nuevo pague o no pague luego de registrar sus datos y aplicar el algoritmo de regresión logística. De ésta manera tendremos que si la probabilidad de que pague o no pague es mayor a 0.50, el futuro cliente tiene mas de la mitad de la probabilidad de que cumpla con sus pagos, mas si la

probabilidad obtenida es menor a 0.50, se recomienda evaluar si se acepta al futuro cliente, ya que existe una alta probabilidad de que incumpla sus pagos.

El modelo de regresión logística utilizado para determinar la probabilidad que existe de que un cliente nuevo pague o no pague, basados en datos históricos de clientes actuales, ha sido probado con datos anteriores, luego verificado con datos posteriores de los clientes mas recientes para así determinar la probabilidad de acierto que tiene el modelo utilizado.

Así mismo en las pruebas realizadas al algoritmo en la cual se excluyó algunas variables, para determinar la importancia que tenía cada una de ellas en el modelo, se determinó que son de vital importancia la variable sueldo, pues al omitirla en la ejecución del algoritmo, daba resultados que guardaban poca coherencia con la realidad.

De igual manera se encontró que la variable de cargas familiares está estrechamente relacionada con la variable de sueldo y a su vez ambas con la variable clase, ya que resultaba muy coherente la relación que existía si un cliente percibía un sueldo relativamente aceptable y con un número de cargas familiares no mayor a tres, por consiguiente se observó que la variable clase contenía “si paga” y a su vez la probabilidad obtenida era mayor a 0.50.

Además se pudo determinar que las variables demográficas (provincia, cantón, parroquia y barrio) asociadas con los indicadores socioeconómicos de empleo, pobreza y servicio eléctrico, si bien es cierto colaboraban para determinar el nivel socio-económico de un cliente nuevo, no era de ninguna forma determinante si el futuro cliente pagaría o no, ya que resultaba muy discriminante solo basar el estudio de si el cliente pagaría o no, solo por el hecho de vivir en determinado zona. Por este motivo es necesario incluir en el modelo todas las variables detalladas anteriormente.

Con los resultados obtenidos con la aplicación desarrollada se colabora en gran medida con una fuerte herramienta para los tomadores de decisiones que tienen que ver con el área de otorgación de crédito a un nuevo cliente.

RECOMENDACIONES

Para que este tipo de sistemas de sistemas siga adelante se debe estar constantemente evaluando los resultados así como también haciendo los respectivos ajustes del modelo ya sea con la selección de variables, parámetros de los algoritmos de búsqueda o la misma aplicación de que algoritmo utilizar.

Por ejemplo sería importante si en el futuro se pudiera intercambiar información con empresas proveedoras del servicio de agua potable, así como las empresas proveedoras del servicio telefónico, en el aspecto de morosidad de sus clientes. Estas variables podrían ayudar bastante al momento de evaluar a un futuro cliente. Y sobre todo si la información obtenida es del mismo cliente

solicitante, ya que sería un historial de pago específico. Sería un caso similar a la forma en que se evalúa a los clientes en el sistema financiero, los cuales tienen acceso al sistema de la Central de Riesgos para determinar la capacidad de pago de un futuro cliente.

Cabe indicar que si bien el modelo funciona utilizando los parámetros más importantes, es necesario que continuamente se evalúe el modelo utilizando nuevos parámetros que podrían optimizar el modelo, siempre evitando recargar el número total de parámetros. Además no se debe olvidar que la herramienta es muy útil al momento de aceptar a un cliente nuevo, pero de igual forma el análisis humano no deja de ser determinante al momento de tomar la decisión final, siendo la aplicación una herramienta de importante ayuda.

BIBLIOGRAFIA

1. Los Fundamentos de Datamining: www.monografias.com
2. Escarbando en las Bases de Datos con técnicas de Minería de Datos: www.5campus.com
3. La Regresión Logística: www.alceingenieria.com
4. La Regresión Logística: www.ilustrados.com
5. EMELGUR (EMPRESA ELECTRICA REGIONAL GUAYAS – LOS RIOS
6. Data Mining Practical Machine Learning Tools and Techniques,Second Edition. Ian H. Witten; Department of Computer Science, University of Waikato; Eibe Frank, Department of Computer Science, University of Waikato.
7. Wiley,.Data.Mining.Techniques.for.Marketing.Sales.and.Customer.Suppo rt.(2004),.2Ed. Michael J. A. Berry, Gordon S. Linoff.

Ing. Juan Alvarado
Director de Tesis