MASKANA, Vol. 9, No. 1, 89–103, 2018
https://publicaciones.ucuenca.edu.ec/ojs/index.php/maskana/article/view/1857
doi: 10.18537/mskn.09.01.09

Optimización de contactos telefónicos efectivos en gestión de cobranzas mediante un modelo de mejor horario de llamada, usando regresión multinomial

Adriana Uquillas-Andrade  ***, Andrés Carrera***

Departamento de Matemática, Facultad de Ciencias, Escuela Politécnica Nacional, Ladrón de Guevara E11253, POBox 17-01-2759, Quito, Ecuador.

Autor para correspondencia: Adriana Uquillas-Andrade, adriana.uquillas@epn.edu.ec

Fecha de recepción: 13 de junio 2017 - Fecha de aceptación: 5 de marzo 2018

RESUMEN

Los Centros de llamadas (en inglés *Call Centers*) representan una industria consolidada a nivel mundial y una de sus actividades es la gestión de cobranzas. El presente trabajo propone un modelo estadístico predictivo para aumentar la probabilidad de contactabilidad telefónica en la gestión de cobranzas a través del mejor horario de llamada. Esto lleva directamente a considerar más de dos posibilidades, es decir, nos enfrentamos a un problema de respuesta multicategórica por lo que se especifica un modelo multinomial. Los datos de corte transversal utilizados en el análisis empírico provienen de una empresa de cobranza de gran escala situada en Ecuador. Los individuos, objeto de este análisis, son prestatarios que se encontraban en mora de productos de crédito de consumo y de microcrédito. El estudio incluye el análisis de aproximadamente 6,000 individuos y el tratamiento de 139 variables explicativas, recogidas en un período histórico entre enero y septiembre de 2016. Los resultados sugieren que información histórica de contactabilidad, día de la semana, características del contrato moroso y la propensión de pago (dada por la razón del saldo en atraso entre el corto plazo y largo plazo), son determinantes de un contacto telefónico efectivo.

Palabras clave: Regresión multinomial, inteligencia de negocios, análisis de negocios, big data, centro de llamadas, contactabilidad.

ABSTRACT

Call Centers represent worldwide a consolidated industry and one of its activities is the management of collections. The present work proposes a predictive statistical model to increase the probability of phone contactability in the collection management through the best call schedule. This leads directly to consider more than two possibilities, that is, we are faced with a multicategorical response problem, so a multinomial model is specified. The cross-sectional data used in the empirical analysis comes from a large-scale collection company located in Ecuador. The individuals, object of this analysis, are borrowers who were in arrears in products of consumer credit and microcredit. The study includes the analysis of approximately 6000 individuals and the treatment of 139 explanatory variables collected between January and September 2016. The results suggest that historical contact information, day of the week, characteristics of debtors in arrears and propensity to pay (given by the ratio between arrears in short-term and in long-term) are determinants of an effective contact by phone.

Keywords: Multinomial regression, business intelligence, business analysis, big data, call center, contactability.

1. INTRODUCCIÓN

La gestión de cobranza es una de las actividades más solicitadas por los clientes de un *Call Center* ya que es una etapa fundamental en la administración de créditos masivos, por tanto, si no se cuenta con las herramientas que permitan un proceso efectivo y ágil se pueden generar desincentivos a los clientes deudores con relación al pago de sus obligaciones.

Debido a la gran competitividad, los *Call Center* enfrentan un desafío común y constante, cobrar más rápido y mejor sin gastar más, es decir existe una necesidad de generar estrategias que den mejores resultados en la cobranza y que permitan adelantarse al resto de acreedores. Por lo tanto, un punto clave está en dar énfasis en la gestión de datos y las tecnologías de información para de esta manera elaborar estrategias y adoptar medidas para promover los intereses de la organización. Recientemente, los avances en ciencia y tecnología han dejado disponibles grandes conjuntos de datos y, por lo tanto, requieren de análisis estadísticos de punta para describir comportamientos en aplicaciones que son grandes (i.e. de terabytes a exabytes) y complejas (i.e. de sensores a datos de redes sociales), y que por tanto requieren almacenamiento avanzado de datos, gestión centralizada e inclusive tecnologías de visualización.

Usualmente, las estrategias en gestión de cobranza se basan en segmentación y en utilización de canales de contacto como llamadas telefónicas, visitas de campo, envío de SMS o mails con el fin de obtener una respuesta positiva por parte del cliente para el cumplimiento de sus obligaciones. Dado que la gestión telefónica manual tiene un costo medio, impacto e interacción altos, la gestión de cobranza se realiza con mayor prioridad por este canal.

Una vez que un *Call Center* ya posee lo último en tecnología, cabe hacerse la siguiente pregunta: ¿Qué queda por hacer para diferenciarse de otros *Call Center* que poseen igual nivel tecnológico? La respuesta está en aplicar Inteligencia de Negocios. Por ejemplo, intuitivamente, es natural aplicar la lógica de llamar a un cliente a la misma hora que ha sido contactado en otras ocasiones, sin embargo ¿cómo aplicarlo en una lista de 100,000 clientes?, ¿Qué pasa con aquellos que nunca han sido contactados previamente? Las respuestas a estas interrogantes están basadas en minería de datos para análisis de asociación, segmentación y agrupación de datos, análisis de clasificación y regresión, detección de anomalías y modelos predictivos, herramientas que son capaces de provocar mejoras importantes en la contactabilidad y en los resultados de negocios.

El objetivo principal de este estudio es revelar los factores determinantes del mejor horario de llamada para clientes que están pasando por un proceso de cobranza, para de esta manera contribuir en el desarrollo de estrategias de llamadas para la administración de los portafolios de cobranza que permita optimizar los recursos y mejorar la recuperación. Como objetivos específicos se plantean: 1) analizar las principales peculiaridades de la gestión de cobranzas; 2) gestionar óptimamente la gran cantidad de datos disponibles; 3) definir lo que es un contacto telefónico efectivo; 4) desarrollar un modelo econométrico multicategórico de mejor horario de llamada; y 5) evaluar el poder de discriminación del modelo estadístico.

Para enfrentar problemas que requieren una respuesta cualitativa binaria, es bastante común el uso de modelos de regresión de tipo *logit* o *probit*, y con estos se han popularizado las medidas de KS, área bajo la curva ROC y el índice GINI que miden la calidad de discriminación y poder de predicción de los modelos de regresión. Este artículo plantea la siguiente hipótesis: es posible extender las medidas de separación o divergencia, medidas de asociación y de calidad de discriminación usados para modelos binarios a un modelo multinomial. Además, se responde a la pregunta de si, con la información disponible, es posible revelar qué factores influyen en la probabilidad de contacto telefónico efectivo de un cliente.

La modelización de programación de llamadas, por sí ya es un reto en todo el mundo, otro de los desafíos en este trabajo es la gestión de bases de datos de larga data (recopilación, extracción y análisis de datos). Los datos fueron estructurados y recopilados a través de diversos sistemas de gestión de bases relacionales de una empresa de cobranza de grande escala. Literatura especializada en *Inteligencia de Negocios* y *Big Data* tiene como objetivo desarrollar varias técnicas analíticas tales como minería de reglas de asociación, segmentación y agrupación de bases de datos, detección de anomalías y minería de gráficos. Estas aplicaciones muestran cómo la investigación académica de alta calidad puede abordar

problemas del mundo real y aportar soluciones que sean relevantes y duraderas. Sin embargo, Park, Huh, Oh, & Han (2012) sostienen que los sistemas de inteligencia de negocios son de valor limitado cuando tratan con datos inexactos y datos poco fiables.

Con relación a modelos de contactabilidad, la literatura es escasa. Cunningham, Martin, & Brick (2003) comparan diferentes algoritmos de llamadas con el objetivo de aumentar la probabilidad de llegar al hogar y determinar si el número es residencial o no. Los hallazgos mostraron que patrones de llamada por período de tiempo (combinaciones de llamadas de día / noche / fin de semana) tenían una mayor probabilidad de contactar a los hogares. En particular, las primeras cuatro llamadas tenían que tener un día, un fin de semana y dos intentos de noche. El experimento demuestró claramente que una variedad de períodos de tiempo en llamadas anticipadas reduce el número total de intentos de llamada, necesarios para contactar a los hogares y eliminar números no residenciales. Kreuter & Müller (2015) proponen el uso de encuestas de panel en lugar de encuestas transversales, pues mencionan que las mismas pueden utilizar información de comportamientos anteriores para mejorar los algoritmos de programación de llamadas. Los estudios observacionales anteriores mostraron el beneficio de llamar en momentos en que se había tenido éxito en el pasado. Los resultados de una encuesta nacional a gran escala en Alemania muestran ganancias modestas de eficiencia, medidas en número de intentos de llamada necesarios hasta el primer contacto, pero sin ganancias en la eficiencia para obtener cooperación.

Conforme se mencionó anteriormente, la mayoría de los *Call Center* especializados en cobranzas o ventas utilizan varios canales de contacto como teléfono, mensajes de texto, mails, envío de cartas, mensajes en redes sociales, visitas domiciliarias, y estos canales puede tener más de un tipo como teléfono convencional de trabajo o casa, teléfono celular, etc, o varios tipos de direcciones de correo electrónico, si es personal o del trabajo. Es por esta razón que Bayrak *et al.* (2013), proponen un método para producir un *score* que indique el mejor momento para usar un canal para un cliente específico, este *score* puede basarse en datos históricos de los clientes y el método puede ser lo suficientemente flexible para manejar variables de distintos niveles de disponibilidad de datos. La escala de "*mejor tiempo*" puede depender del canal utilizado para contactar al cliente. Para las llamadas telefónicas, la escala de tiempo relevante puede ser "*hora del día*", por ejemplo, descrita por las franjas horarias de una hora. Por otro lado, en el caso de los envíos de cartas, es posible que la escala de tiempo pertinente sea "*día de la semana*" y que para los correos electrónicos sea "*día de la semana*" en combinación con una medida más aproximada de "*hora del día*", como por la mañana / tarde / noche.

Por otro lado, Durrant, D'Arrigo, & Steele (2011) investigaron los mejores tiempos de contacto para diferentes tipos de hogares y la influencia de un encuestador en establecer contacto. En aquel trabajo se indica que los recientes desarrollos en el proceso de recopilación de datos de una encuesta han llevado a la recolección de los llamados procesos de campo, que amplían considerablemente la información básica sobre las llamadas a los encuestadores. Este artículo desarrolla un modelo de respuesta múltiple basado en datos de registro de llamadas del encuestador para predecir la probabilidad de contacto en cada llamada.

Queda claro que estudios previos acerca de los determinantes del mejor horario de llamada son realmente incipientes, trabajos que abordan cuestiones relacionadas al mejor momento para usar un canal con los clientes, o algoritmos para programación de llamadas, son los de Kreuter & Müller (2015) y Bayrak *et al.* (2013). Por tanto, el presente trabajo constituye un aporte técnico necesario, consistente en proponer un nuevo modelo estadístico de mejor horario de llamada a los clientes, para aumentar la contactabilidad telefónica en la gestión de cobranzas. Esto lleva directamente a considerar más de dos posibilidades, es decir, nos enfrentamos a un problema de respuesta multi categórica, por lo que se especifica un modelo multinomial. Los datos de corte transversal utilizados en el análisis empírico provienen de una empresa de cobranza de gran escala situada en Ecuador. Los individuos, objeto de este análisis, son prestatarios ecuatorianos que se encontraban en mora en productos de crédito de consumo y de microcrédito. Se obtuvo información de enero a septiembre del 2016, se modelizaron a los individuos que se los gestionó telefónicamente en los meses de julio y agosto del 2016, a estos meses se los llama puntos de observación. Se recopiló información histórica de los últimos 6 meses antes de los puntos de observación. Los resultados sugieren que información histórica de contactabilidad, día de la semana, características del contrato moroso y la propensión de pago (dada por la razón del saldo en atraso entre el corto plazo y largo plazo), son determinantes de un contacto telefónico efectivo.

2. MATERIALES Y MÉTODOS

2.1. Modelo estadístico

Los modelos multinomiales se analizan eligiendo una categoría como referencia de la variable dependiente o de respuesta y se modelan varias ecuaciones simultáneamente, una para cada una de las restantes categorías respecto a la de referencia. Se considera una variable de respuesta Y con más de dos categorías de respuesta, denotadas por $Y_1, Y_2, \dots, Y_k$. Se pretende explicar la probabilidad de cada categoría de respuesta en función de un conjunto de covariables $X = \{x_1, x_2, \dots, x_n\}$ observadas.

Cuando la variable de respuesta es politómica, la distribución de Bernouilli se convierte en una distribución multinomial así que para obtener un modelo lineal se obtendrán $\binom{k}{2}$ transformaciones logit, pero para construir el modelo logit de respuesta multinomial bastará con considerar (k-1) transformaciones logit básicas, definidas con respecto a una categoría de referencia. Tomando como categoría de referencia la última Y_k, las transformaciones logit generalizadas se definen en la ecuación (1):

$$L_j(x) = ln\left[\frac{p_j(x)}{p_k(x)}\right], \forall j = 1, \dots, k \tag{1}$$

siendo $L_j(x)$ el logaritmo de la ventaja de respuesta Y_j dado que las observaciones de las variables independientes caen en la categoría Y_j o en la Y_k.

El modelo lineal para cada una de las transformaciones *logit* generalizadas, para n variables explicativas, ecuación (2):

$$L_j(x) = \sum_{s=0}^{n} b_{sj}x_s = xb_j\,, \forall j = 1, \dots k-1 \tag{2}$$

Para cada vector de valores observados de las variables explicativas $x = (x_0, x_1, \dots, x_n)'$ con $x_0 = 1$ y $b_j = (b_{0j}, b_{1j}, \dots, b_{nj})'$ el valor de parámetros asociados a la categoría Y_j.

Para las probabilidades de respuesta, se puede escribir el modelo de la manera presentada en las ecuaciones (3) y (4):

$$p_j(x) = \frac{\exp\left(\sum_{s=0}^{n} b_{sj}x_s\right)}{1 + \sum_{j=1}^{k-1}\exp\left(\sum_{s=0}^{n} b_{sj}x_s\right)}\,, \forall j = 1, \dots, k-1 \tag{3}$$

$$p_k(x) = \frac{1}{1 + \sum_{j=1}^{k-1}\exp(\sum_{s=0}^{n} b_{sj}x_s)} \tag{4}$$

donde b_{sj} es el coeficiente estimado de la variable x_s asociado a la categoría *j*.

Para estimar los coeficientes del modelo de regresión multinomial, se usa el método de máxima verosimilitud. Los cálculos para las estimaciones de los coeficientes de la regresión logística multinomial no son directos, por lo que es necesario usar métodos iterativos como el método de Newton-Rapson. Usando estos métodos se obtienen los coeficientes y sus errores estándar.

2.2. Selección de variables

Las medidas de separación o divergencia ayudan a conocer el poder predictivo de las variables numéricas continuas. Para modelos de respuesta binaria, es común usar la prueba de Kolmogorov-Smirnov para seleccionar las mejores variables explicativas de acuerdo con el valor que tenga este estadístico, es decir, a mayor valor del estadístico mayor poder de predicción de la variable.

Al estadístico KS se define de acuerdo con la ecuación (5):

$$KS = \sup_x |F_X - G_X| \tag{5}$$

donde F_X representa la función de distribución acumulada empírica para la población 1 y G_X representa la función de distribución acumulada empírica 2. El KS corresponde a la distancia vertical máxima entre

los gráficos de F_X y G_X sobre la amplitud de los posibles valores de x. De esta forma, la prueba KS contrasta la hipótesis de que si las dos distribuciones son idénticas o no.

En este caso, la variable respuesta tiene más de dos categorías, por tanto, surge la necesidad de extender este concepto para el caso multinomial. Utilizando el trabajo realizado por Loftus *et al*. (2015), se describe la generalización del estadístico KS para más de dos muestras, al que se lo denotará por KSM (Kolmogorov-Smirnov Measure).

Tomando en cuenta que el estadístico KS se ajusta a la definición de una métrica y que sus valores están entre 0 y 1, se define KSM como la suma ponderada de los valores de KS de todas las $\binom{k}{2}$ combinaciones por variable. Los pesos se toman proporcionales al tamaño total de la muestra, de modo que la medida KSM está dada por la ecuación (6):

$$KSM_s = \sum_{k=1}^{K} \sum_{k \neq k'} \frac{N_k + N_{k'}}{N(K-1)} KS_s(k, k') \tag{6}$$

donde $KS_s(k, k')$ es el valor del estadístico KS al comparar las distribuciones de una variable x_s cuando la variable de respuesta es $k\ o\ k'$ y N es el tamaño total de la muestra. Como la suma de los pesos es 1, el estadístico KSM está en el intervalo [0,1] y tiene la misma interpretación que el estadístico KS, valores cercanos a uno indican una mayor diferencia en las distribuciones de x_s cuando la variable de respuesta tiene múltiples categorías.

Por otro lado, las medidas de asociación son indicadores que miden el poder predictivo de las variables categóricas consideradas importantes para formar parte del modelo. El presente trabajo utiliza el estadístico chi-cuadrado para estudiar la dependencia entre la variable dependiente politómica y las variables explicativas categóricas, sean estas binarias o politómicas.

Tabla 1. Estructura de una tabla de contingencia para medir la asociación entre la variable Y y las variables $X_{i.}$

Y/X	X_1	X_2	…	X_j	…	X_p	Totales
Y_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1p}	$n_{1.}$
Y_2	n_{21}	n_{22}	…	n_{2j}	…	n_{1p}	$n_{2.}$
.							.
.							.
.							.
Y_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ip}	$n_{i.}$
.							.
.							.
.							.
Y_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{kp}	$n_{k.}$
Totales	$n_{.1}$	$n_{.1}$	…	$n_{.j}$	…	$n_{.p}$	n

De la Tabla 1 se obtienen las ecuaciones (7) y (8):

$$n_{i.} = \sum_{j=1}^{p} n_{ij} \qquad \forall i = 1, \dots, k \tag{7}$$

$$n_{.j} = \sum_{i=1}^{k} n_{ij} \qquad \forall j = 1, \dots, p \tag{8}$$

La prueba Chi-cuadrado contrasta la hipótesis nula de independencia de las variables X e Y versus la hipótesis alternativa de existencia de asociación entre estas variables a un determinado nivel de significación α, en base a la información recogida en la tabla de contingencia (Tabla 1).

Se define el valor n'_{ij} como la frecuencia esperada que correspondería al par de categorías (Y_i, X_j) y está dado por la ecuación (9):

$$n'_{ij} = \frac{n_{i.} n_{.j}}{n} \quad \forall i = \{1,2,\dots,k\}; j = \{1,2,\dots,p\} \tag{9}$$

El valor del estadístico asociado a la prueba puede ser calculado por la ecuación (10):

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} \frac{(n'_{ij} - n_{ij})^2}{n'_{ij}} \tag{10}$$

En general, en tablas de k x p se utiliza el coeficiente de contingencia de Pearson definido por la ecuación (11).

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} \tag{11}$$

Este coeficiente varía entre $0 \leq C \leq \sqrt{\frac{q-1}{q}} < 1$ donde $q = mín\{k, p\}$. Valores cercanos a 0 indicarán independencia entre las variables y valores cercanos a 1 indicarán que existe relación entre las mismas.

2.3. Análisis de multicolinealidad

Se define a la multicolinealidad como el problema de que una variable explicativa en el modelo de regresión sea una combinación lineal de las demás, es decir, que dos o más variables estén linealmente correlacionadas. Las consecuencias de multicolinealidad en una regresión son los altos errores estándar e incluso la imposibilidad de cualquier estimación.

Para estudiar el problema de multicolinealidad se utiliza el índice de condicionamiento (IC), definido por la ecuación (12):

$$IC = \sqrt{\frac{\lambda_{máx}}{\lambda_{mín}}} \tag{12}$$

donde $\lambda_{máx}$ y $\lambda_{mín}$ son los valores propios máximo y mínimo respectivamente, de la matriz de correlaciones de las variables explicativas. Si $IC < 10$, no hay presencia de multicolinealidad; si $10 \leq IC \leq 15$ existe multicolinealidad moderada, y si $IC > 15$ existe multicolinealidad fuerte (Milone, 2009).

2.4. Medidas de poder de discriminación

Para el modelo de regresión multinomial, donde se tendrán k vectores de probabilidades estimadas, se utilizará la medida de KS extendida para el caso multinomial KSM (explicada en detalle en la sección anterior). Según Anderson (2007), para valores de KS inferiores a 0.2 debe cuestionarse el poder de discriminación del modelo. Por otro lado, valores superiores a 0.7 podrían traer cuestionamientos de que el modelo es demasiado bueno para ser verdad. En la industria bancaria un KS con valor de 0.5 se consideran como un *benchmark.*

El área bajo el ROC (AUROC: area under the curve ROC) se ha convertido en un criterio de evaluación de desempeño estándar en problemas de reconocimiento de patrones de dos clases. Los trabajos de Hand & Till (2001) y Landgrebe & Duin (2006) extienden la medida AUC para el caso multinomial o multiclase y se la nombra VUS (*volumen under the surface*).

Las covariables x son clasificadas dentro de las categorías $Y_1, Y_2, \dots, Y_k$ de la variable dependiente Y. Cada categoría tiene una distribución condicional $g(x|Y_j)$ y una probabilidad $p(Y_j)$. La asignación de las categorías se basa en la regla de Bayes, la cual asigna para cada individuo su probabilidad más alta, de acuerdo a la ecuación (13):

$$p(Y_j|x) = \frac{p(Y_j)g(x|Y_j)}{p(Y_1)g(x|Y_1) + p(Y_2)g(x|Y_2) + \cdots + p(Y_k)g(x|Y_k)} \quad (13)$$

Luego, de acuerdo a la ecuación (14), para cada individuo se tomará:

$$argmáx_{j=1}^{k} p(Y_j|x) \quad (14)$$

En la práctica se desconocen las distribuciones condicionales de las categorías, estas se estiman típicamente a partir de ejemplos representativos que se supone que se extraen aleatoriamente de la distribución verdadera, y se pueden usar en el mismo marco. Entonces, usando las probabilidades estimadas del modelo de regresión multinomial, cada categoría tendrá una probabilidad de ocurrencia $p_j(x)$ y, de acuerdo con las ecuaciones (13) y (14), a cada individuo le corresponderá $\text{máx}_{j=1}^{k} p_j(x)$.

Las clasificaciones se analizan a detalle por medio de la matriz de confusión (Tabla 2) de dimensión $k \times k$, donde los elementos de la diagonal representan las clasificaciones correctas en cada categoría, y, los elementos fuera de la diagonal, los errores relacionados con cada categoría. El caso de dos categorías es muy conocido, con dos elementos fuera de las diagonales r_{12} y r_{21}, popularmente conocidos como falsos negativos y falsos positivos, respectivamente, y dos elementos diagonales r_{11} y r_{22}, las verdaderas tasas positivas (sensibilidad) y verdaderas negativas, respectivamente (especificidad). En este caso se obtiene un gráfico de sensibilidad vs especificidad (Fig. 1) donde, si el área bajo la curva tiene un valor de 0.0, implica que las predicciones del modelo son perfectamente erróneas, un valor de 0.5 indica que realiza una predicción aleatoria y un valor de 1 implica que el modelo realiza una predicción perfecta; un valor superior a 0.7 es considerado adecuado en el caso de dos categorías (Anderson, 2007). Esta área se conoce como el área bajo el ROC (AUROC) y puede ser escrita como la ecuación (15):

$$AUC = \int r_{22} dr_{11} \quad (15)$$

Tabla 2. Estructura de una Matriz de Confusión utilizada para medir la sensibilidad y especificidad de los pronósticos.

Real/Pronóstico	Y_1	Y_2	...	Y_k
Y_1	r_{11}	r_{12}	$\cdots$	r_{1k}
Y_2	r_{21}	r_{22}	$\cdots$	r_{2k}
.	.	.	$\cdots$	.
.	.	.	$\cdots$	.
.	.	.	$\cdots$	.
Y_k	r_{k1}	r_{k2}	$\cdots$	r_{kk}

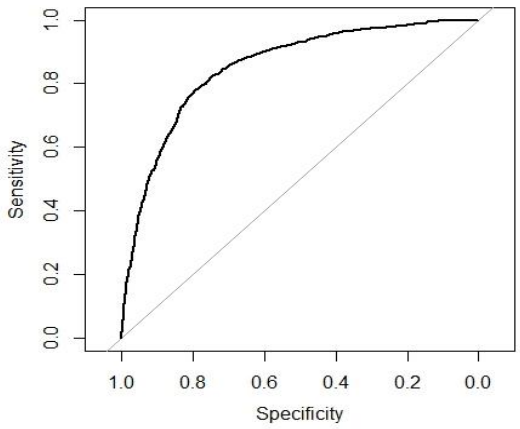

Figura 1. Curva ROC: Gráfico de sensibilidad vs especificidad.

El coeficiente GINI, está relacionado con la AUROC mediante la igualdad presentada en la ecuación (16), para el caso de dos categorías.

$$GINI = 2AUROC - 1 \quad (16)$$

La extensión del AUC al caso multinomial, lleva al cálculo del volumen bajo la superficie del hiperplano ROC. En este caso, se considera solamente las dimensiones ROC correspondientes a los elementos diagonales de la matriz de confusión. El VUS (*Volume Under the Surface*) simplificado se puede escribir de acuerdo a la ecuación (17):

$$VUS = \int \dots \int \int r_{11} dr_{22} dr_{33} \dots dr_{kk} \tag{17}$$

Esta medida permite evaluar la clasificación sobre todos los puntos responsables de las dimensiones ROC correspondientes a los elementos de la diagonal de la matriz de confusión.

2.5. Fuentes de datos y variables.

Considerando la información disponible en la base de datos, apoyándose en investigaciones anteriores y en la literatura revisada, se incluye en la especificación del modelo variables explicativas de acuerdo con los siguientes grandes grupos de información: variables de comportamiento crediticio, variables de gestión telefónica, informaciones de contrato de crédito y variables sociodemográficas. Variables importantes para este estudio tales como profesión y estado civil no fueron tomadas en cuenta por escasez de información y por dudas acerca de su calidad.

Se recopiló información histórica de contactabilidad de los últimos 6 meses antes de los puntos de observación. De acuerdo con Kreuter & Müller (2015), la información histórica de contactabilidad es relevante en este tipo de trabajo. Además, a partir de las informaciones brutas de los grandes grupos de información antes indicados, se construyeron variables transformadas con el fin de incorporar nociones de comportamiento temporal y dar dinamismo al modelo. Esto permitió pasar de un conjunto de aproximadamente 45 variables a la disponibilidad de 139 variables explicativas (5 variables categóricas y 134 variables numéricas continuas).

A partir de esta información, se toman dos muestras aleatorias, una para modelamiento que consta del 60% de la población total y otra de validación que consta del 40%. Se definieron ventanas de tiempo de comportamiento y de desempeño (Fig. 2). En la ventana de comportamiento se construyen las variables históricas que proporcionan dinamismo temporal al modelo y en la ventana de desempeño se define la variable dependiente.

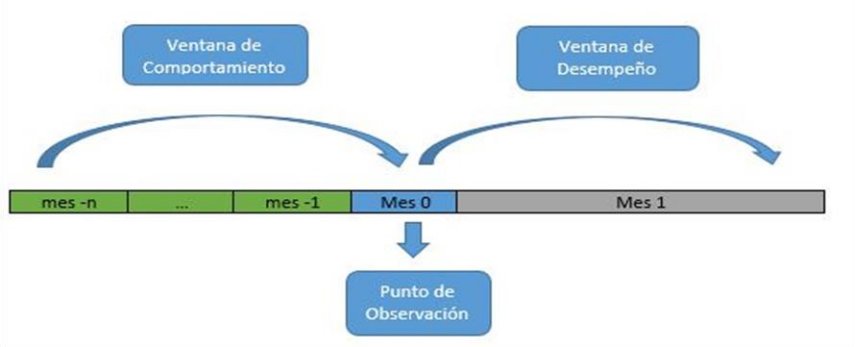

Figura 2. Ventanas de tiempo de comportamiento histórico y de desempeño.

2.6. Variable dependiente.

La variable dependiente Y es una variable cualitativa donde cada categoría es un horario del día en el cual se puede contactar telefónicamente a un cliente. Para definir estas categorías se analizó la información de gestiones telefónicas de enero 2016 a septiembre 2016, se consideraron las conexiones efectivas y llamadas telefónicas realizadas en cada hora del día durante el mes.

Alineados con el negocio y la gestión actual de cobranza de la empresa proveedora de la información, fueron establecidas las siguientes cuatro categorías de horario de llamadas: (a) 7:00 a 9:00; (b) 9:00 a13:00; (c) 13:00 a 16:00; (d) 16:00 a 21:00. En la Figura 3 se muestra el patrón de conexiones efectivas donde, para los horarios 7:00-9:00 y 13:00-16:00, los porcentajes de conexión efectiva son más altos que para los horarios de 9:00-13:00 y 16:00-21:00.

Se define como pce_j al porcentaje de conexión efectiva en el horario *j*, *j* = {1, 2, 3, 4}, de acuerdo con la ecuación (18).

$$pce_j = 100\frac{\#contactos\ efectivos_j}{\#llamadas\ totales_j} \tag{18}$$

Se etiqueta como contactado en el horario j a todos los individuos cuyo valor máximo de porcentaje de conexión efectiva corresponde al horario j, si el valor máximo de porcentaje de conexión efectiva es 0 se los etiqueta como no contactado en ningún horario (NC) y la variable *Y* toma el valor de 0. Por lo tanto, *Y* se define de acuerdo con la ecuación (19):

$$Y = \begin{cases} j & si\ máx_{j\in\{1,2,3,4\}} pce_j \neq 0 \\ 0 & caso\ contrario \end{cases} \tag{19}$$

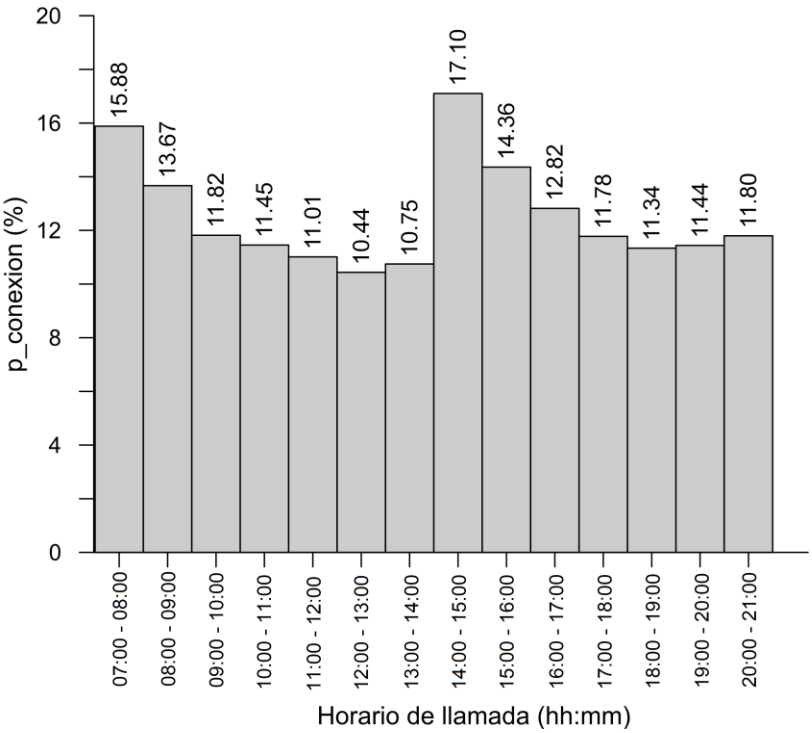

Figura 3. Porcentaje de Conexiones Efectivas agosto 2016.

3. RESULTADOS

Los datos fueron estructurados y recopilados a través de diversos sistemas de gestión de bases de datos relacionales. Se obtuvo información de enero a septiembre del 2016, se modelizaron a los individuos que se los gestionó telefónicamente en julio y agosto del 2016, a estos meses se los llama puntos de observación.

Para el filtrado de las variables numéricas continuas, se utilizó la medida KSM. En la Figura 4 se puede ver que a partir de la variable 115 la medida del KSM tiende a 0, por tanto, estas variables son las que se pueden descartar con seguridad. De las variables restantes, se realiza un análisis de correlación cruzada y se descartan variables explicativas con correlaciones cruzadas mayores al 70%, manteniendo aquellas que presentan mayor KSM, así, el conjunto de variables numéricas continuas candidatas para el modelo se reduce a 45.

Con relación a las variables categóricas se utiliza el coeficiente de contingencia de Pearson (CCP), definido a detalle en la sección 2. En la Figura 5 se presenta el gráfico del CCP por variable, en este caso se dispone de una cantidad baja de variables y, de estas, dos se pueden descartar con seguridad.

Al contar con 48 variables explicativas se procede a seleccionar el mejor modelo mediante la técnica *Stepwise* (método paso a paso). Bendel & Afifi (1977) dan detalles sobre este método.

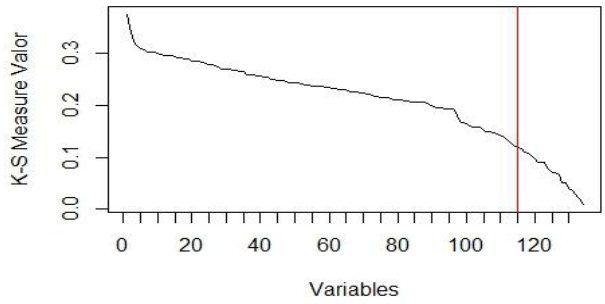

Figura 4. Medida KSM obtenida para cada variable en estudio

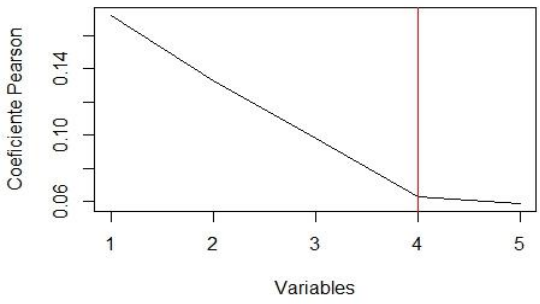

Figura 5. Coeficiente de Correlación de Pearson obtenido para cada variable en estudio.

El modelo de regresión logística multinomial propuesto está constituido por las siguientes variables explicativas:

- num_conex: Número de conexiones efectivas al punto de observación.
- p_conex: Porcentaje de conexión efectiva al punto de observación.
- DIAGESTION: IS (Inicio Semana): lunes y martes, MS (Mitad de Semana): miércoles y jueves y FD (Fin de Semana): viernes y sábado. Esta variable fue recategorizada a través de árboles de decisión con el objetivo de maximizar la explicación de la variable respuesta. Su codificación final fue: 0 si DIAGESTION=IS y 1 caso contrario.
- PRODUCTO: Variable categórica relacionada al tipo de contrato, recategorizada con árboles de decisión codificada con 1 si PRODUCTO=Rotativo y 0 caso contrario.
- max_porc_conex_6M: Máximo de los porcentajes de conexión efectiva en el largo plazo.
- num_conex_2M: Número de conexiones efectivas en el corto plazo.
- min_num_conex_4M: Mínimo del número de conexiones efectivas en el mediano plazo.
- rsaldo_inicial_2M_26: Relación de saldo en atraso entre el corto y largo plazo.

En las Tabla 3 se presentan los modelos estimados para las categorías de horario 07:00-09:00, 09:00-13:00, 13:00-16:00 y 16:00-21:00, respectivamente; también se muestra el coeficiente estimado, error estándar, significancia y los extremos del intervalo de confianza al 95%.

Los coeficientes estimados b_{sj} asociados a las categorías Y_j de la variable dependiente *Y*, se interpretan en términos de los cocientes de ventajas (en inglés *odds ratio*), calculados por $exp(b_{sj})$. Se interpreta cada una de las variables independientes entre los distintos horarios de contacto, tomando como referencia NC: no contactado en ningún horario. En la Tabla 4 se presentan los coeficientes estimados junto con los *odds ratio* para el horario de 07:00-9:00. Por ejemplo, en la Tabla 4 se sugiere que la ventaja de contactar en el horario de 07:00-9:00, frente a NC, es de 1.620 veces, a medida que el número de conexiones efectivas aumenta en una unidad, *ceteris paribus*. La ventaja de contactar en el horario de 07:00-9:00 entre semana (miércoles o jueves) o en fin de semana (viernes o sábado), frente a NC, en inicio de semana (lunes o martes) es de 2.425 veces *ceteris paribus*. La ventaja de contactar en el horario de 07:00-9:00 para cliente con producto rotativo, frente a no contactar en ningún horario en otro producto, es de 1.718 veces, *ceteris paribus*. La ventaja de contactar en el horario de 07:00-9:00, frente a no contactar en ningún horario, es de 1.008 veces, a medida que el máximo de los porcentajes de conexión efectiva en el largo plazo aumenta en una unidad, *ceteris paribus*. La ventaja de contactar en el horario de 07:00-9:00, frente a no contactar en ningún horario, es de 1.101 veces, a medida que el número de conexiones efectivas en el corto plazo aumenta en una unidad, *ceteris paribus*. La ventaja de contactar en el horario de 07:00-9:00, frente a no contactar en ningún horario, es de 1.521 veces, a medida que el mínimo de conexiones efectivas en el mediano plazo aumenta en una unidad, *ceteris paribus*. La ventaja de contactar en el horario de 7:00-9:00, frente a no contactar en ningún horario, es de 2.226 veces, a medida que la razón del saldo en atraso, en el corto y largo plazo, aumenta en una unidad, *ceteris paribus*.

Tabla 3. Modelo predictivo estimado para diferentes horarios de llamada.

Horario de llamada	Estadísticas del modelo predictivo	Variables Explicativas								
		Intercepto	*num_conex*	*p_conex*	*DIAGESTION*	*PRODUCTO*	*max_porc_conex_6M*	*num_conex_2M*	*min_num_conex_2M*	*Rsaldo_inicial_2M_6M*
07:00-9:00	Coeficiente	-6.1	0.483	0.027	0.886	0.541	0.008	0.096	0.419	0.8
	Std. Error	0.4	0.045	0.004	0.269	0.21	0.003	0.025	0.121	0.27
	Significancia	0	0	0	0.001	0.01	0.002	0	0.001	0.003
	Sup.	-6.884	0.394	0.019	0.358	0.13	0.003	0.047	0.182	0.272
	Inf.	-5.315	0.571	0.035	1.414	0.953	0.014	0.146	0.657	1.328
09:00-13:00	Coeficiente	-3.976	0.356	0.026	0.185	0.548	0.007	0.079	0.372	0.673
	Std. Error	0.261	0.041	0.003	0.141	0.148	0.002	0.02	0.103	0.228
	Significancia	0	0	0	0.189	0	0.001	0	0	0.003
	Sup.	-4.488	0.276	0.019	-0.091	0.257	0.003	0.039	0.169	0.225
	Inf.	-3.465	0.436	0.032	0.461	0.839	0.01	0.12	0.574	1.12
13:00-16:00	Coeficiente	-5.126	0.43	0.028	0.732	0.82	0.011	0.097	0.145	0.337
	Std. Error	0.343	0.042	0.004	0.194	0.182	0.002	0.022	0.114	0.289
	Significancia	0	0	0	0	0	0	0	0.205	0.243
	Sup.	-5.798	0.347	0.021	0.352	0.462	0.007	0.053	-0.079	-0.229
	Inf.	-4.453	0.513	0.035	1.113	1.177	0.015	0.14	0.369	0.904
16:00-21:00	Coeficiente	-4.131	0.37	0.026	0.265	0.874	0.004	0.091	0.4	0.584
	Std. Error	0.266	0.04	0.003	0.139	0.156	0.002	0.02	0.101	0.23
	Significancia	0	0	0	0.056	0	0.036	0	0	0.011
	Sup.	-4.653	0.292	0.02	-0.007	0.568	0	0.053	0.203	0.134
	Inf.	-3.608	0.449	0.033	0.538	1.18	0.008	0.13	0.597	1.035

Tabla 4. Coeficientes estimados junto con los *odds ratio* del modelo predictivo para el horario de 07:00-09:00.

07:00-09:00	Coeficiente	Odss
Intercepto	-6.100	0.002
num_conex	0.483	1.620
p_conex	0.027	1.027
DIAGESTION	0.886	2.425
PRODUCTO	0.541	1.172
*max_porc_conex_6M_*	0.008	1.008
num_conex_2M	0.096	1.101
min_num_conex_4M	0.419	1.521
Rsaldo_inicial_2M_6M	0.800	2.226

La interpretación del resto de coeficientes estimados para cada categoría se realiza de manera análoga. Además, el índice IC definido a detalle en la sección 2, tiene un valor de 2.686; por tanto, se puede concluir que el modelo no presenta problemas de multicolinealidad.

Poder de discriminación del modelo

En la Figura 6 se muestran los gráficos correspondientes al estadístico KS y a la curva ROC para cada horario de la variable respuesta para la muestra de validación, mientras que en la Tabla 5 se muestran los valores de las medidas de calidad de discriminación para cada horario de la variable respuesta. El valor de los estadísticos se encuentra dentro de los rangos mencionados anteriormente, por tanto, los resultados muestran una buena capacidad de discriminación entre un horario específico de llamada y el no llamar al cliente.

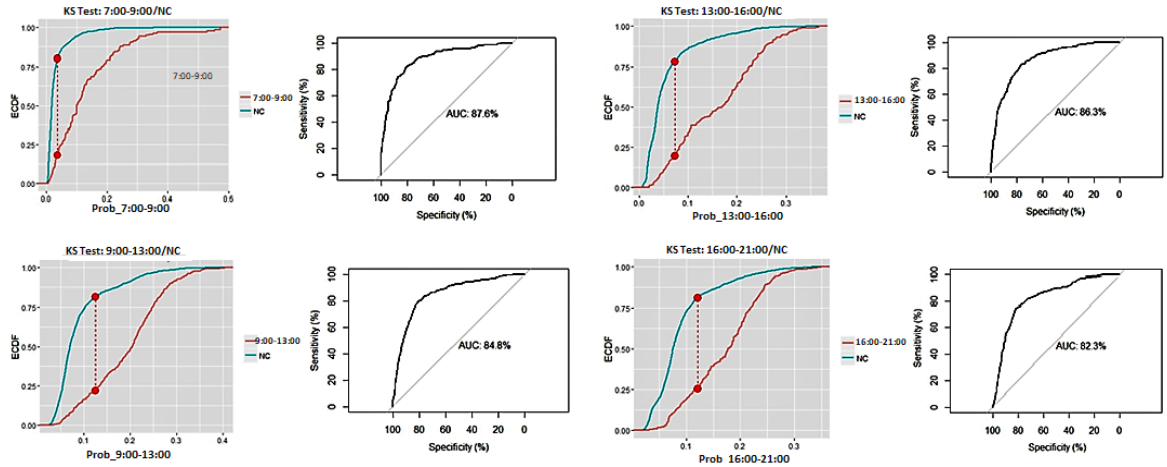

Figura 6. Gráficos correspondientes al estadístico KS y a la curva ROC para cada horario de la variable respuesta para la muestra de validación.

Tabla 5. Valores de las medidas de calidad de discriminación para cada horario de la variable respuesta.

Horario de llamada	Validación		
	KS	AUROC	GINI
07:00-09:00	0.63	0.88	0.77
09:00-13:00	0.60	0.85	0.70
13:00-16:00	0.60	0.87	0.73
16:00-21:00	0.57	0.82	0.65

Por otro lado, los indicadores globales (KSM y VUS) del modelo son KSM = 0.41 y VUS = 0.66. Para el cálculo de estos indicadores se ha tomado la máxima probabilidad de contactabilidad estimada para cada individuo, como el mejor horario de llamada, si bien estos indicadores son inferiores a los indicados en la Tabla 5, todavía presentan valores aceptables de discriminación. Este resultado puede deberse a que en el modelo no fue posible incluir variables importantes tales como estado civil, profesión, sector de residencia, números de teléfonos activos, si el teléfono es de casa, de una referencia personal u oficina, indicador si el cliente realizó una llamada al *Call Center* y tiempos de llamadas (Bayrak *et al.*, 2013).

Finalmente, la metodología más usada en la industria para tratar este tipo de modelos con variable de respuesta multicategórica, son los árboles de decisión. Usando esta técnica, fueron necesarias 15 variables explicativas para alcanzar similares niveles de desempeño a los obtenidos con la regresión multinomial. Casi el doble de las variables explicativas del modelo propuesto (total de 8 variables), lo que indica la eficiencia en discriminación del modelo de regresión multinomial, además de tener la ventaja de interpretación de los parámetros, característica que el árbol de decisión no posee. Se valida de esta forma que el método propuesto puede generar mejores resultados, satisfaciendo el principio de parsimonia.

En la Tabla 6 se muestran los resultados de los estadísticos KS, AUROC y GINI obtenidos con la técnica de Árbol de Decisión. Las medidas KSM y VUS fueron 0.41 y 0.65, respectivamente.

Tabla 6. Medidas de calidad de discriminación, obtenidas con la técnica de árbol de decisión.

Horario de llamada	Árbol de decisión con 15 variables		
	KS	AUROC	GINI
07:00-09:00	0.64	0.89	0.78
09:00-13:00	0.58	0.85	0.71
13:00-16:00	0.59	0.86	0.73
16:00-21:00	0.59	0.84	0.69

4. DISCUSIÓN

Los resultados sugieren que información histórica de contactabilidad, el día de la semana, características del contrato moroso y la propensión de pago, son determinantes de un contacto telefónico efectivo. En lo que se refiere a información histórica de contactabilidad y día de semana, como determinantes de contacto efectivo de una llamada, los resultados coinciden con los estudios realizados por Cunningham *et al.* (2003) y Bayrak *et al.* (2013). Adicionalmente, en este trabajo se incluyeron informaciones históricas crediticias del cliente, en específico, la relación histórica de saldo en atraso entre el corto y largo plazo es un factor determinante.

Tomando en cuenta las medidas de KS, AUROC y GINI el mejor modelo, en términos de modelamiento es el horario de 7:00-9:00, seguido por el horario de 13:00-16:00. Siendo así, se podrían implementar mayores esfuerzos en estos horarios para aumentar la contactabilidad. Considerando los umbrales de discriminación establecidos por Anderson (2007), los resultados muestran que el poder de discriminación de los modelos es muy adecuado por lo que se concluye que con la información disponible es posible revelar qué factores influyen en la probabilidad de contacto telefónico efectivo de un cliente y se esperaría que su uso en la gestión de cobranzas provoque mejoras importantes en la tasa de contactabilidad y en los resultados de negocios puesto que los *Call Center* enfrentan un desafío común y constante, cobrar de manera eficiente, es decir, existe una necesidad de generar estrategias que den mejores resultados en la cobranza y que permitan adelantarse al resto de acreedores.

Aunque los modelos presentan un buen desempeño, es de relevancia para futuras investigaciones la inclusión de variables cadastrales tales como estado civil, profesión, sector de residencia, y de acuerdo con Bayrak *et al.* (2013), otras informaciones que dependen de la gestión de llamadas tales como números de teléfonos activos, indicador si el cliente realizó una llamada al *Call Center*, etc. La no inclusión de estas variables puede haber ocasionado que el desempeño del modelo global sea inferior al desempeño de cada categoría en el modelo multinomial.

En términos metodológicos, una oportunidad para desarrollos futuros con el objetivo de obtener mayor flexibilidad con las variables de los modelos es tratar las k-1 regresiones logísticas de manera independiente (donde k es el número de categorías de la variable repuesta). De este modo se pueden tratar las variables de manera más específica y así obtener variables diferentes en cada modelo logístico binomial. Esta metodología llevaría a construir otro tipo de modelo multinomial que podría compararse con los resultados de discriminación obtenidos en este artículo. Además, concentraciones en una sola categoría es común en la práctica (en este caso concentraciones en la categoría No Contactado). Para esto se han desarrollado modelos estadísticos que describen este fenómeno y permiten derivar conclusiones realistas y confiables a partir de las inferencias. Comparar los resultados de este trabajo con los modelos llamados Modelos Inflados en cero, debe ser de interés teórico-práctico.

Por otro lado, Kreuter & Müller (2015) proponen el uso de datos en panel, en lugar de datos en corte transversal, puesto que en esta estructura se puede utilizar información de comportamientos anteriores para mejorar los algoritmos de programación de llamadas. La estructura de datos en panel ofrece dos dimensiones, la dimensión transversal y la de series de tiempo, con lo que, para obtener ese tipo de información, debe acompañarse a los individuos a lo largo del tiempo y por tanto no es posible asumir que los individuos son independientemente distribuidos temporalmente (Wooldridge, 2012). En este trabajo se usan datos con corte transversal agrupados de manera independiente. Este muestreo

independiente en diferentes momentos de tiempo, consecuentemente, en este tipo de corte transversal los individuos tampoco están independientemente distribuidos a lo largo del tiempo. Según Wooldridge (2012), aun cuando con este tipo de muestro es posible obtener estimaciones más precisas y pruebas estadísticas con más poder, en este tipo de modelado suele hacerse evidente la presencia de un cambio estructural a través del tiempo. Esto significa que el impacto de los determinantes del mejor horario de llamada puede cambiar con el transcurso del tiempo y por tanto sería necesario incluir, por ejemplo, una variable temporal a los modelos. En este trabajo, es de esperar que este cambio estructural suceda, aunque con poca probabilidad, pues, aunque el registro histórico disponible fue de solamente de 6 meses, se comprobó la no necesidad de inclusión de una variable temporal. Además, para dar dinamismo al modelo, y de cierta manera incluir esa dimensión temporal faltante, se construyeron variables explicativas comportamentales recogidas en diferentes espacios temporales, lo cual evidencia que no solo con la estructura de datos en panel se puede añadir una dimensión temporal a los datos, sino también con el tratamiento y uso que se da a la información disponible.

5. CONCLUSIONES

La modelización de programación de llamadas es un reto. Estudios previos acerca de los determinantes del mejor horario de llamada son realmente incipientes (Durrant *et al.*, 2011; Kreuter & Müller, 2015; Bayrak *et al.* 2013), por lo que este trabajo constituye un aporte técnico importante, además de ser un modelo original, pues los autores desconocen que se haya desarrollado algo similar en Latinoamérica.

En este artículo, los autores proponen un modelo de regresión multinomial de mejor horario de llamada a los clientes, para aumentar la contactabilidad telefónica en la gestión de cobranzas. El estudio deja en evidencia la eficiencia en discriminación y la ventaja de interpretación de los parámetros del modelo planteado, frente a la técnica del *árbol de decisión*. Se valida de esta forma que el método propuesto puede generar mejores resultados, satisfaciendo el principio de parsimonia. La metodología planteada podría ampliarse fácilmente a otras situaciones de gestión e inteligencia de negocios tales como ventas, promociones, gestión de despacho, entre otros.

La gestión de bases de datos de larga data es justamente otro gran desafío de este trabajo (recopilación, extracción y análisis de datos). La importancia de trabajar con gran cantidad de datos no gira en torno a la cantidad de datos que se tiene, pero sí en torno al tratamiento y uso que se les da a los mismos. Al combinar una gran disponibilidad de datos con herramientas estadísticas de gran potencia, el modelo propuesto, y en general, la inteligencia y análisis de negocios, pueden llevar a reducciones de costos y tiempo, desarrollo de nuevos productos, estrategias, ofertas optimizadas y toma de decisiones inteligentes.

Finalmente, según Chen (2011), la próxima década promete ser desafiadora para la investigación y desarrollo de alto impacto en inteligencia de negocios y análisis, tanto para la industria como para el mundo académico. La comunidad empresarial y la industria ya han dado pasos importantes para adoptar inteligencia de negocios a sus necesidades. La comunidad de ciencias de datos enfrenta desafíos y oportunidades únicas para hacer impactos científicos y sociales relevantes y duraderos. Este trabajo representa un aporte para la comunidad de ciencias de datos.

REFERENCIAS

Anderson, R. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*. Oxford, UK: OUP.

Bayrak, H., Bulbul, A. A., Conser, E. T., Bergh, G. de, Dorai, C., Veen, A. (2013). *US20130060587A1*. United States. Retrieved from https://patents.google.com/patent/US20130060587A1/en

Bendel, R. B., Afifi, A. A. (1977). Comparison of stopping rules in forward 'stepwise' regression. *Journal of the American Statistical Association*, *72*(357), 46-53. https://doi.org/10.2307/2286904

Chen, H. (2011). Editorial: Design science, grand challenges, and societal impacts. *ACM Transactions on Management Information Systems (TMIS)*, *2*(1), 1-10. https://doi.org/10.1145/1929916.1929917

Cunningham, P., Martin, D., Brick, J. M. (2003). *An experiment in call scheduling*. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 59-66. Nashville, TN, USA: American Statistical Association. Retrieved from http://ww2.amstat.org/sections/srms/proceedings/y2003/Files/JSM2003-000306.pdf

Durrant, G. B., D'Arrigo, J., Steele, F. (2011). Using field process data to predict best times of contact conditioning on household and interviewer influences. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *174*(4), 1029-1049.

Hand, D. J., Till, R. J. (2001). A simple generalization of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45*(2), 171-186. https://doi.org/10.1023/A:1010920819831

Kreuter, F., Müller, G. (2015). A note on improving process efficiency in panel surveys with para data. *Field Methods*, *27*(1), 55-65. https://doi.org/10.1177/1525822X14538205

Landgrebe, T., Duin, R. P. W. (2006). *A simplified extension of the area under the ROC to the multiclass domain*. Delft, The Netherlands: Delft University of Technology. Disponible en https://pdfs.semanticscholar.org/dc70/1e7fca147e2bf37f14481e35e1b975396809.pdf?_ga=2.3211 1294.1451026522.1529313596-700634931.1501938280

Loftus, S. C., House, L. L., Hughey, M. C., Walke, J. B., H, M., Belden, L. K. (2015). *Dimension reduction for multinomial models via a Kolmogorov-Smirnov measure (KSM)* Technical No. 15-1, 19 p. Blacksburg, VA, USA: Virginia Tech. Retrieved from https://www.stat.vt.edu/about/research/research-technical-reports.html

Milone, G. (2004). *Estatistica geral e aplicada*. Editora Pioneira Thomson Learning. Retrieved from https://www.estantevirtual.com.br/livros/giuseppe-milone/estatistica-geral-e-aplicada/1704135913

Park, S. H., Huh, S. Y., Oh, W., Han, S. P. (2012). A social network-based inference model for validating customer profile data. *MIS Quarterly: Management Information Systems*, *36*(4), 1217-1238.

Wooldridge, J. (2012). *Introductory econometrics: A modern approach* (5th ed.). Michigan, US: Michigan State University, Cengage Learning.