Optimización de Contactos Telefónicos Efectivos en Gestión de Cobranzas Mediante un Modelo de Mejor Horario de Llamada (Best Time to Call)

Resumen: Los Centros de llamadas (en inglés Call Centers) representan una industria consolidada a nivel mundial. Una de sus posibles actividades es la gestión de cobranzas. Una vez que un Call Center ya posee lo último en tecnología ¿Qué queda por hacer para diferenciarse de los pares con igual nivel tecnológico? La respuesta está en aplicar Inteligencia de Negocios.

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 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 grande escala situada en Ecuador. Los individuos, objeto de este análisis, son prestatarios que se encontraban en mora en productos de crédito de consumo y de microcrédito. El estudio incluye el análisis de aproximadamente 6000 individuos y el tratamiento de 139 variables explicativas recogidas en un período histórico entre enero y septiembre de 2016.

El modelo consigue relacionar los efectos de diferentes variables con la probabilidad de contactabilidad. 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 de saldo en atraso al corto 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 a consolidated industry worldwide. One of its possible activities is the management of collections. Once a Call Center already has the latest in technology What remains to be done to differentiate itself from peers with the same technological level? The answer lies in applying Business Intelligence. The present work proposes a statistical predictive model to increase the probability of telephone contactability in the management of collections through the best call time. This leads directly to consider more than two possibilities, that is to say, we are faced with a multi-categorical response problem by which a multinomial model is specified. The cross-sectional data used in the empirical analysis come 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 in a historical period between January and September 2016. The model manages to relate the effects of different variables to the likelihood of contactability. The results suggest that historical contact information, day of the week, characteristics of the delinquent contract and the propensity to pay (given the short and long-term overdue balance) are determinants of an effective telephone contact.

Keywords: Multinomial Regression, Business Intelligence, Business Analysis, Big Data, Call Center, Schedule.

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 disponibilizado grandes conjuntos de datos y por lo tanto, requieren de análisis estadísticos de punta para describir comportamientos en aplicaciones que son tan grandes (de terabytes a exabytes) y complejas (de sensores a datos de redes sociales) que requieren almacenamiento avanzado de datos, gestión centralizada e inclusive tecnologías de visualización.

Las estrategias usuales en gestión de cobranza se basan en segmentación de carteras por días mora y tipo de producto 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.

A medida que la edad de mora avanza, la probabilidad de recuperación disminuye, porque entre otras cosas, los clientes no son contactados fácilmente y en muchos casos son inubicables.

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 y para ello, se definen los denominados árboles de gestión que son parametrizables por tipo de empresa. El árbol de gestión telefónica, por ejemplo, incorpora acciones desde el proceso de marcado hasta el manejo de las objeciones por parte del gestor, con esto se logra una estandarización tanto en los procesos de llamadas como en los resultados de la gestión y se obtienen indicadores para cada acción.

En la Figura 1 se presenta un ejemplo de árbol de gestión para telefonía.

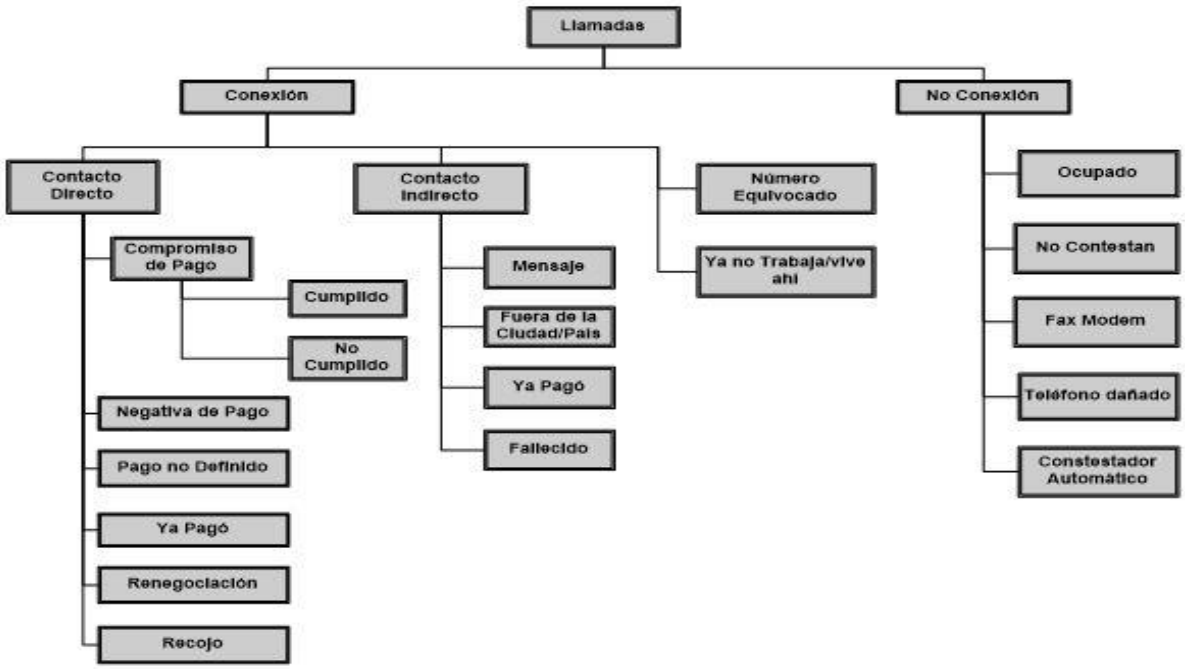

Figura 1. *Árbol de Gestión de Telefonía*

Los indicadores de gestión se relacionan con el proceso realizado en el canal correspondiente, por lo que su definición se basa en las respuestas del árbol construido.

Para el caso de gestión telefónica, los indicadores nacen del resultado de la llamada. La contactabilidad se mide directamente como contactos efectivos sobre las llamadas realizadas y depende de varios factores. Entre ellos: la calidad de los datos, la estrategia de discado o marcación, la tecnología de marcado disponible, la adhesión a los procedimientos de los agentes, la tecnología de apoyo a la gestión disponible, la capacidad de aplicar inteligencia de negocios a la distribución de llamadas.

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 lo he contactado en otras ocasiones, sin embargo ¿cómo aplicarlo en una lista de 100.000 clientes?, ¿Qué pasa con aquellos que nunca he contactado 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.

En este artículo, los autores proponen un modelo estadístico de mejor horario de llamada a los clientes, para aumentar la probabilidad de contactabilidad telefónica en la gestión de cobranzas. La metodología planteada podría ampliarse fácilmente a otras situaciones de gestión e inteligencia de negocios tales como ventas, promociones, gestión de despacho, entre otros.

Además de la modelización de programación de llamadas, que por sí ya es un reto para encuestas en todo el mundo, uno de los principales desafíos en este trabajo fue 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 de datos relacionales de una empresa de cobranza de grande escala. 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 de saldo en atraso al corto y largo plazo), son determinantes de un contacto telefónico efectivo.

El resto del artículo está estructurado de la siguiente manera: La sección 2 presenta una revisión de literatura. La sección 3 describe la metodología y datos disponibles. La sección 4 presenta los resultados empíricos. Finalmente, la sección 5 concluye el trabajo.

2. REVISIÓN DE LITERATURA

En un esfuerzo por comprender mejor el estado actual de la investigación propuesta e identificar futuras fuentes de conocimiento, se analizó la literatura pertinente y las principales publicaciones académicas. La década de los 2010 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 (Chen 2011a).

La literatura presentada aborda varios aspectos del marco de investigación en Inteligencia de Negocios y Big Data. Se presentan aplicaciones a comercio electrónico e inteligencia de mercado mediante análisis de texto, web y de red; seguridad de información mediante análisis de red y análisis de datos; técnicas analíticas para sistemas de recomendación de productos; análisis de riesgo crediticio; algoritmos de llamada para aumentar la probabilidad de contactarse con la residencia; el mejor momento para usar un canal para un cliente específico; mejores tiempos de contacto para diferentes tipos de hogares, etc. En todos estos trabajos se han desarrollado 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.

Michael Chau y Jennifer Xu (2012) desarrollaron un marco para recopilar la inteligencia de negocios al recolectar y analizar automáticamente el contenido del blog y las redes de interacción de los bloggers. Un sistema desarrollado utilizando este marco se aplicó a dos estudios de caso, que reveló nuevos patrones en las interacciones blogger y las comunidades.

(Sung-Hyuk Park et al. 2012), sostienen que los sistemas de inteligencia de negocios son de valor limitado cuando tratan con datos inexactos y datos poco fiables. Los autores propusieron un marco de referencia basado en redes sociales para determinar la exactitud y fiabilidad de los perfiles de clientes auto-reportados. El marco utilizó los círculos sociales de los individuos y patrones de comunicación dentro de sus círculos. Para construir el modelo de inferencia y validación específicos se analizaron más de 20 millones de transacciones reales de llamadas móviles y se utilizó una combinación de métodos, incluyendo procesamiento de consultas, inferencia estadística, análisis de redes sociales y perfiles de usuarios.

(Raymond Lau et al. 2012) analizaron las fusiones y adquisiciones de empresas. El escaneo ambiental en línea con Web 2.0 brinda a los altos ejecutivos la oportunidad de aprovechar la inteligencia colectiva de la red para desarrollar mejores conocimientos sobre los factores socio-culturales y político-económicos que atraviesan las fronteras de fusiones y adquisiciones empresariales. Con base en el modelo de cinco fuerzas de Porter, esta investigación diseñó un modelo de puntuación para mejorar la toma de decisiones. Los autores también desarrollaron un sistema adaptativo de inteligencia empresarial que aplicaban a las actividades de fusiones y adquisiciones transfronterizas de las empresas chinas.

(Daning Hu, et al. 2012) analizaron el riesgo sistémico en los sistemas bancarios desde un enfoque de red para la gestión de riesgos. Los autores tratan a los bancos como una red vinculada con las relaciones financieras y analizan el riesgo sistémico atribuido a cada banco individual mediante simulación basada en datos del mundo real de la Corporación Federal de Seguros de Depósitos. El trabajo proporciona un nuevo medio para predecir las fallas bancarias contagiosas.

(Ahmed Abbasi et al. 2012), desarrollaron un marco de aprendizaje para la detección de fraudes financieros. Una serie de experimentos se llevó a cabo en miles de empresas legítimas y fraudulentas para demostrar la eficacia del marco de referencia sobre los métodos de referencia existentes. Los resultados de la investigación tienen implicaciones para los gestores de cartera, auditoría y reguladores.

En el artículo de (Nachiketa Sahoo et al. 2012), los autores propusieron un modelo de Markov oculto basado en el filtrado colaborativo para predecir las preferencias del usuario y hacer las recomendaciones personalizadas más apropiadas para la preferencia prevista. Los autores emplearon conjuntos de datos y simulaciones del mundo real para demostrar que, cuando las preferencias de los usuarios están cambiando, existe la ventaja de utilizar el algoritmo de Markov Oculto.

Con relación a modelos de contactabilidad, la literatura no es basta. (Cunningham P et al. 2003) en su trabajo 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 demuestra 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. Por otro lado, este algoritmo y otros que imponen mayores restricciones en lo que se refiere a cuándo se pueden realizar llamadas a un número,

tienen implicaciones en la práctica. A medida que se imponen más restricciones al momento en que se pueden hacer llamadas, el flujo de casos a los entrevistadores puede verse afectado. En el extremo, un planificador que pone demasiadas restricciones puede hacer que los entrevistadores no tengan trabajo para llamar en ciertos períodos de tiempo. Claramente, esto sería mucho menos eficiente que permitir llamadas a números en períodos menos óptimos. En tales casos, los algoritmos más sofisticados deben dar prioridad a llamar a los más propensos de beneficiarse de un intento de llamada en el período actual pero direccionar a otros períodos de llamada conforme posibilidad de tiempo de los entrevistadores.

(Frauke Kreuter y Gerrit Muller, 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.

La mayoría de los Call Centers 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 direcciones de correo electrónico, si es personal o del trabajo. Es por esta razón que (Bayrak, 2013) propone en su trabajo 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. Para los clientes con pocos o ningún dato histórico, el método proporciona una puntuación de "mejor momento para ponerse en contacto" que se basa en observaciones generales sobre la probabilidad de llegar a un cliente en un periodo de tiempo. Para los clientes con una gran cantidad de datos históricos, el método proporciona un score de "mejor momento para ponerse en contacto" que se asemeja a los patrones de accesibilidad histórica del cliente (score de riesgo).

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.

Además, el "mejor momento" también se puede describir en términos relativos en el caso de mensajes instantáneos, por ejemplo, la escala de tiempo pertinente puede ser "minutos después de que un cliente esté disponible para chatear". De manera similar, la definición de un "intento de contacto exitoso" puede ser diferente para diferentes canales.

Durrant (2011) propone un trabajo donde investiga los mejores tiempos de contacto para diferentes tipos de hogares y la influencia de un encuestador en establecer contacto. 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 o paradatas, 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.

3. MARCO TEÓRICO/METODOLOGÍA

3.1 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 como

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

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, es de la forma

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

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 siguiente manera

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

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

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

3.2 Estimación por máxima verosimilitud.

De igual forma que en la regresión logística clásica, cuando se dispone de datos a nivel individual o micro, para estimar los coeficientes del modelo de regresión multinomial, se usa el método de máxima verosimilitud que consiste en encontrar los valores de los coeficientes que maximicen la probabilidad de obtener los valores de la variable dependiente en función de los datos proporcionados por la muestra.

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.

Se tiene una muestra aleatoria de tamaño N con Q combinaciones diferentes de valores de las variables explicativas $X_1, X_2, \dots, X_n$. Se denota por $x_q = \left(x_{q0}, x_{q1}, \dots, x_{qn}\right)'$ con $x_{q0} = 1 \; \forall q = 1, \dots, Q$. En cada una de estas combinaciones se tiene una muestra aleatoria de d_q observaciones independientes de la variable de respuesta Y, de entre las cuales se denota por $y_{j/q}$ al número de observaciones que cae en la categoría de respuesta $Y_j \; \forall j = 1, \dots, k$. Aquí se verifica que $\sum_{j=1}^{k} y_{j/q} = d_q$ y $\sum_{q=1}^{Q} d_q = N$.

Los vectores $\left(y_{1/q}, \dots, y_{k/q}\right)' \; \forall q = 1, \dots, Q$ siguen distribuciones de probabilidad multinomiales independientes, siendo $p_{j/q} = P\left[Y = Y_j | X = x_q\right]$ y verificando que $\sum_{j=1}^{k} p_{j/q} = 1$. Por tanto, la función de verosimilitud de los datos viene dada por

$$V = \prod_{q=1}^{Q} \left(\frac{d_q!}{\prod_{j=1}^{k} (y_{j/q})!} \prod_{j=1}^{k} p_{j/q}^{y_{j/q}} \right)$$

Normalmente, en lugar de utilizar la función de verosimilitud se utiliza la función auxiliar

$$\Lambda = -2\ln\left(V\right)$$

Por lo que el problema de maximizar la verosimilitud equivale al de minimizar esta función auxiliar.

Para obtener los estimadores de máxima verosimilitud se debe resolver k-1 sistemas de n+1 ecuaciones no lineales. Con este método se obtiene el estimador de los parámetros $\hat{b}$, que es una matriz de dimensión (n+1)×(k-1) formada por las siguientes columnas:

$\hat{b} = (\hat{b}'_1, \hat{b}'_2, \dots, \hat{b}'_{k-1})'$ siendo $\hat{b}_j'$ el estimador de máxima verosimilitud del vector de parámetros asociado a la categoría Y_j de la variable respuesta Y.

3.2 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, como la regresión logística clásica, es común usar la prueba de Kolmogorov-Smirnov para seleccionar las mejores variables explicativas de acuerdo al valor que tenga este estadístico, es decir, a mayor valor del estadístico mayor poder de predicción de la variable.

En este caso, la variable respuesta tiene más de dos categorías, por tanto surge la necesidad de extender este concepto para el caso multinomial. La prueba de Kolmogorov-Smirnov (también conocida como prueba K-S) es una prueba no paramétrica que determina la bondad de ajuste de dos distribuciones continuas independientes, contrastando la hipótesis de que si estas son idénticas o no. Utilizando el trabajo realizado por Loftus (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

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

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. En el presente trabajo, se 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**. Tabla de Contingencia*

$\mathbf{Y \backslash X}$	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_{2p}	$n_{2.}$
.							.
.							.
.							.
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	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.p}$	n

De la Tabla 1 se tiene

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

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

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).

H0: X e Y son independientes

H1: X e Y no son independientes

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

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

El valor del estadístico asociado a la prueba, puede ser calculado por

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

La medida contraria a la independencia es la asociación y se dice que dos variables están asociadas si aparecen juntas en mayor número de veces que el esperado si fuesen independientes, por tanto, si se rechaza la hipótesis nula entonces existe asociación y según sea la tendencia a coincidir o no se tendrán distintos grados de asociación. En general, en tablas de k x p se utiliza el coeficiente de contingencia de Pearson definido por

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

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.

3.3 Validación del modelo.

Dado que el modelo multinomial se puede ver como varios modelos logit clásicos, para validarlo, aparte de usar las estadísticas comunes, como pruebas sobre los coeficientes (estadístico de Wald), bondad de ajuste del modelo (estadístico chi-cuadrado), análisis de multicolinealidad y residuos, se utilizan el estadístico KS y el área bajo la curva ROC (AUROC) para evaluar la calidad de discriminación.

3.4 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

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

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).

3.5 Medidas de poder de discriminación

Al estadístico KS se lo define como:

$$KS = \sup_x |F_X - G_X|$$

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 (escore estimado por el modelo).

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).

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. Utilizando el trabajo de Landgrebe y Duin (2006) se extiende 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 $Y_1, Y_2, \ldots, Y_k$ 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:

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

Luego para cada individuo se tomará

$$argmáx_{j=1}^{k} p(Y_j|x)$$

En la práctica, se desconocen las distribuciones condicionales de las categorías, éstas 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 por las ecuaciones (15) y (16) a cada individuo le corresponderá $\mathrm{máx}_{\mathrm{j=1}}^{\mathrm{k}} \mathrm{p_j(x)}$.

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 y verdaderas negativas respectivamente.

En este caso se obtiene un gráfico de sensibilidad vs especificidad (Figura 2), donde:

- **sensibilidad:** Probabilidad de ser un verdadero positivo.(r_{11})
- **especificidad:** Probabilidad de ser un verdadero negativo. (r_{22})

***Tabla 2**. Matriz de Confusión*

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

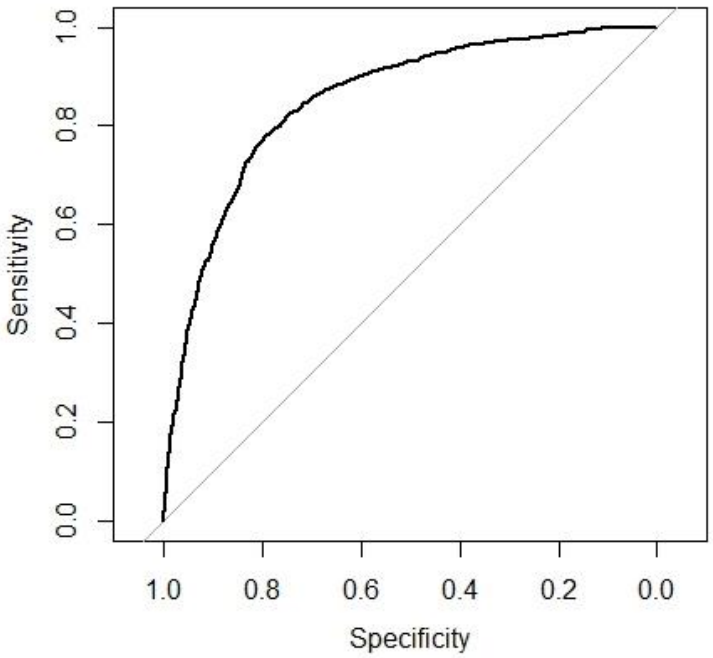

Figura 2. **Curva ROC**

La clasificación perfecta resulta cuando el área bajo la curva es igual a 1, una clasificación pobre cuando el área es cercana a 0 y una clasificación aleatoria cuando el valor es 0,5 (puesto que es una porción del cuadrado unitario). Esta área se conoce como el área bajo el ROC (AUROC) y puede ser escrita como

$$AUC = \int r_{22} dr_{11}$$

La extensión del AUC al caso multinomial lleva al cálculo de el 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 (Volumen Under the Surface) simplificado se puede escribir como:

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

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.

Si sólo se consideran estos resultados, el VUS es similar al AUC en que si la clasificación es buena, resultará en un VUS alto y clasificaciones más pobres en un puntaje más bajo. Sin embargo, antes de que el VUS se aplique ciegamente, es importante caracterizar y comprender los límites de rendimiento entre clasificadores aleatorios y perfectos.

La medida del rendimiento del AUC está estrechamente relacionada con el coeficiente de Gini, que a veces se utiliza como medida alternativa.

Esto se define más comúnmente como el doble del área entre la curva ROC y la diagonal (siendo esta área tomada como negativa en el raro caso de que la curva esté por debajo de la diagonal). La geometría elemental muestra que Gini+1 = 2 × AUC.

En este artículo se tabaja en términos de AUC, pero los resultados se aplican igualmente al coeficiente de Gini.

Finalmente, a partir del modelo ajustado, se clasifica cada observación en la categoría más probable, construyendo así una matriz de clasificación observados - predichos y se utiliza el porcentaje de clasificaciones correctas como una medida de la calidad de predicción. Se define como la proporción de individuos clasificados correctamente por el modelo y se calcula como el cociente entre el número de observaciones clasificadas correctamente y el tamaño muestral N. Un individuo es clasificado correctamente por el modelo cuando su valor observado de la variable respuesta Y coincide con su valor estimado por el modelo.

4. RESULTADOS Y DISCUSIÓN

Los datos fueron estructurados y recopilados a través de diversos sistemas de gestión de bases de datos relacionales de una empresa de cobranza de grande escala. 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 julio y agosto del 2016, a estos meses se los llama puntos de observación.

4.1 Fuente 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 a 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 a (Frauke Kreuter y Gerrit Muller, 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 (Figura 3). 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.

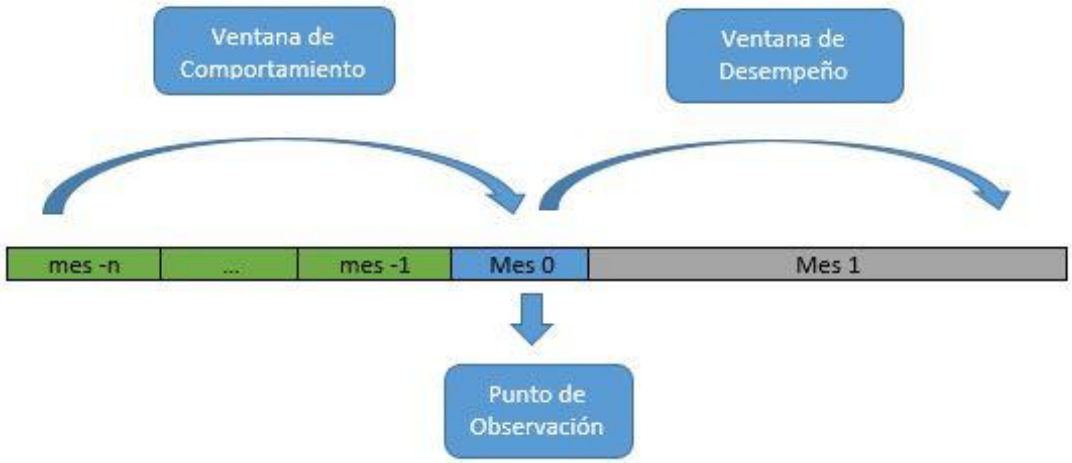

Figura 3. Ventanas de Tiempo

Una vez que la metodología y la base de datos fueron definidas, se procede a definir a la variable dependiente categórica Y, donde cada categoría será un horario del día.

4.2 Variable Dependiente. La variable dependiente Y es una variable cualitativa con más de dos categorías, 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 las informaciones fueron establecidas las siguientes cuatro categorías de horario de llamadas: 1: 7am a 9am; 2: 9am a13pm; 3: 13pm a 16pm; 4: 16pm a 21pm. En la Figura 4 se muestra el patrón de conexiones efectivas, donde en los horarios 7-9am y 13-16pm los porcentajes de conexión efectiva son más altos que en los horarios de 9-13pm y 16-21pm.

Se define como pce_j al porcentaje de conexión efectiva en el horario j, j = {1, 2, 3, 4}:

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

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 se define a Y como sigue:

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

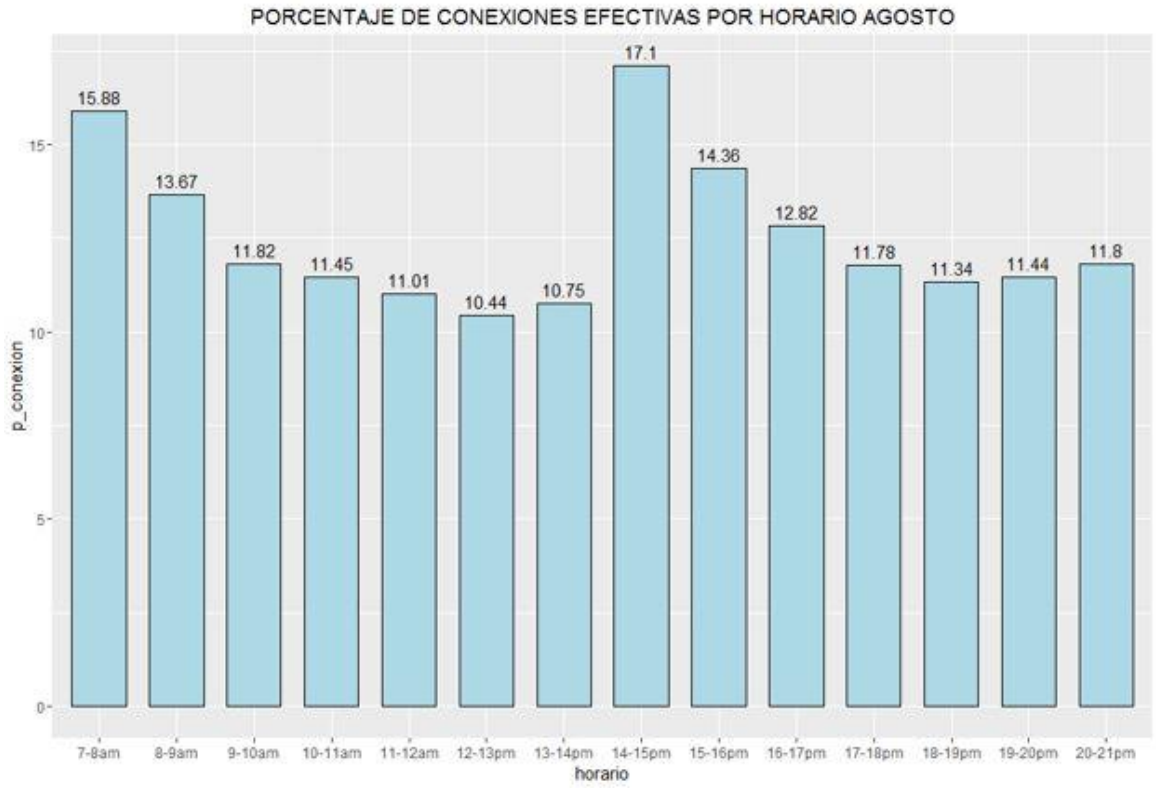

Figura 4. Porcentaje de Conexiones Efectivas Agosto 2016

4.3 Selección de Variables Numéricas Continuas. Para el filtrado de las variables numéricas continuas, se utiliza la medida KSM definida a detalle en la sección 2.

En la Figura 5 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.

4.3.1 Selección de Variables Categóricas. Para el filtrado de variables categóricas se utiliza el coeficiente de contingencia de Pearson (CCP), definido a detalle en la sección 2.

En la Figura 6 se presenta el gráfico del CCP por variable, en este caso se dispone de una cantidad baja de variables y de estas se pueden descartar con seguridad dos.

Una vez obtenidas las mejores variables entre numéricas continuas y categóricas (un total de 48 variables), se procedió a seleccionar el mejor modelo mediante la técnica stepwise (método pasa a paso). (Robert B. Bendel y A.A. Afifi, 2012) dan detalles sobre este método.

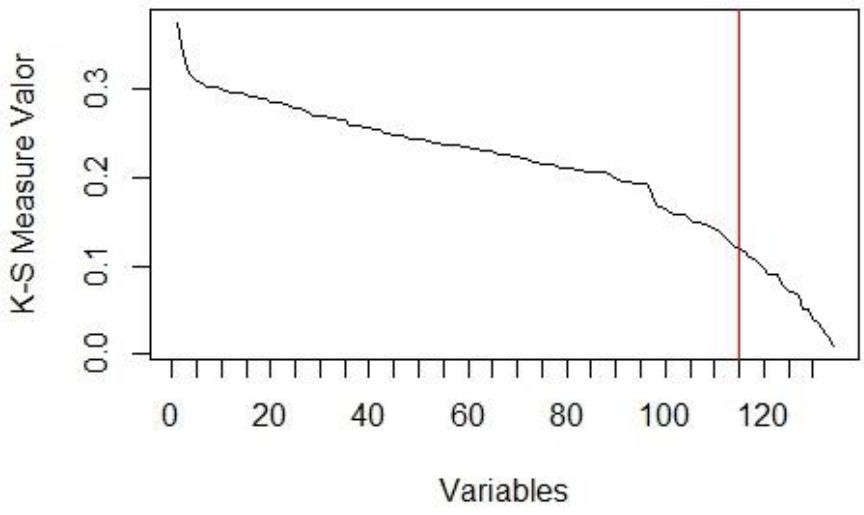

Figura 5. KSM por Variable

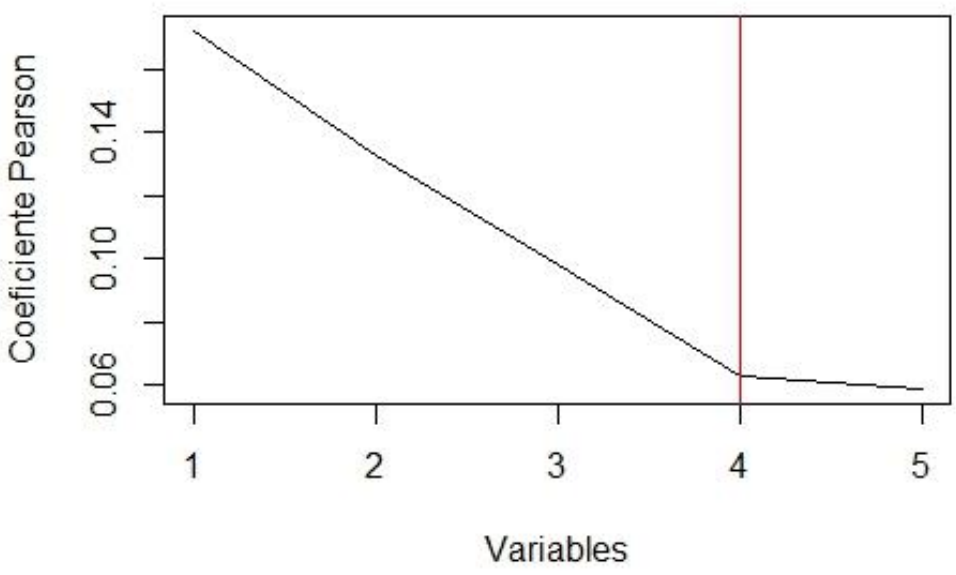

Figura 6. CCP por Variable

Para elegir el mejor modelo se analizó la significancia de las variables en cada uno de los modelos estimados para cada horario de la variable respuesta, de tal forma de obtener un modelo con mayor diversidad de variables significativas en todos los modelos para cada horario y, además, respetando el principio de parsimonia.

4.4 Modelo Propuesto

El modelo propuesto está constituido por las siguientes variables explicativas:

- num_conex: Variable continua de gestión. Número de conexiones efectivas al punto de observación.

- p_conex: Variable continua de gestión. Porcentaje de conexión efectiva al punto de observación.

- DIAGESTION: Variable categórica referente a los días de semana: IS (Inicio Semana): Lunes y Martes, MS (Mitad de Semana): Miércoles y Jueves y FD (Fin de Semana): Viernes y Sábado. Esta variable fue re categorizada 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: Variable continua de gestión. Máximo de los porcentajes de conexión efectiva en el largo plazo.

- num_conex_2M: Variable continua de gestión. Número de conexiones efectivas en el corto plazo.

- min_num_conex_4M: Variable continua de gestión. Mínimo del número de conexiones efectivas en el mediano plazo.

- rsaldo_inicial_2M_26: Variable continua de comportamiento crediticio. Saldo en atraso en el corto y largo plazo.

En las Tablas 3, 4, 5 y 6 se presentan los modelos estimados para las categoría de horario 7-9am, 9-13pm, 13-16pm y 16-21pm respectivamente. En las tablas se muestra el coeficiente estimado, error estándar, significancia y los extremos del intervalo de confianza al 95 %,

Tabla 3. Variables Explicativas 7-9am

7-9am	Coeficiente	Std. Error	Significancia	Sup	Inf
Intercepto	-6,100	0,400	0,000	-6,884	-5,315
num_conex	0,483	0,045	0,000	0,394	0,571
p_conex	0,027	0,004	0,000	0,019	0,035
DIAGESTION	0,886	0,269	0,001	0,358	1,414
PRODUCTO	0,541	0,210	0,010	0,130	0,953
max_porc_conex_6M	0,008	0,003	0,002	0,003	0,014
num_conex_2M	0,096	0,025	0,000	0,047	0,146
min_num_conex_4M	0,419	0,121	0,001	0,182	0,657
rsaldo_inicial_2M_6M	0,800	0,270	0,003	0,272	1,328

Tabla 4. Variables Explicativas 9-13pm

9-13pm	Coeficiente	Std. Error	Significancia	Sup	Inf
Intercepto	-3,976	0,261	0,000	-4,488	-3,465
num_conex	0,356	0,041	0,000	0,276	0,436
p_conex	0,026	0,003	0,000	0,019	0,032
DIAGESTION	0,185	0,141	0,189	-0,091	0,461
PRODUCTO	0,548	0,148	0,000	0,257	0,839
max_porc_conex_6M	0,007	0,002	0,001	0,003	0,010
num_conex_2M	0,079	0,020	0,000	0,039	0,120
min_num_conex_4M	0,372	0,103	0,000	0,169	0,574
rsaldo_inicial_2M_6M	0,673	0,228	0,003	0,225	1,120

Tabla 5. Variables Explicativas 13-16pm

13-16pm	Coeficiente	Std. Error	Significancia	Sup	Inf
Intercepto	-5,126	0,343	0,000	-5,798	-4,453
num_conex	0,430	0,042	0,000	0,347	0,513
p_conex	0,028	0,004	0,000	0,021	0,035
DIAGESTION	0,732	0,194	0,000	0,352	1,113
PRODUCTO	0,820	0,182	0,000	0,462	1,177
max_porc_conex_6M	0,011	0,002	0,000	0,007	0,015
num_conex_2M	0,097	0,022	0,000	0,053	0,140
min_num_conex_4M	0,145	0,114	0,205	-0,079	0,369
rsaldo_inicial_2M_6M	0,337	0,289	0,243	-0,229	0,904

Tabla 6. Variables Explicativas 16-21pm

16-21pm	Coeficiente	Std. Error	Significancia	Sup	Inf
Intercepto	-4,131	0,266	0,000	-4,653	-3,608
num_conex	0,370	0,040	0,000	0,292	0,449
p_conex	0,026	0,003	0,000	0,020	0,033
DIAGESTION	0,265	0,139	0,056	-0,007	0,538
PRODUCTO	0,874	0,156	0,000	0,568	1,180
max_porc_conex_6M	0,004	0,002	0,036	0,000	0,008
num_conex_2M	0,091	0,020	0,000	0,053	0,130
min_num_conex_4M	0,400	0,101	0,000	0,203	0,597
rsaldo_inicial_2M_6M	0,584	0,230	0,011	0,134	1,035

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 $\exp(b_{sj})$.

Cuando se interpreta las odds ratios de cada variable, se asume que el resto de variables independientes se mantienen fijas. 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 7 se presentan los coeficientes estimados junto con los odds ratio para el horario de 7-9am. La Tabla 7 sugiere que:

La ventaja de contactar en el horario de 7-9am frente a no contactar en ningún horario 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 7-9am frente a no contactar en ningún horario es de 1,027 veces a medida que el porcentaje de conexión efectiva aumenta en una unidad ceteris paribus.

La ventaja de contactar en el horario de 7-9am entre semana (Miércoles o Jueves) o en fin de semana (Viernes o Sábado) frente a no contactar en ningún horario en inicio de semana (Lunes o Martes) es de 2,425 veces ceteris paribus.

La ventaja de contactar en el horario de 7-9am 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 7-9am 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 7-9am 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 7-9am 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 7-9am 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 7. Odds Ratio 7-9am

7-9am	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	1,718
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.

4.5 Poder de Discriminación del Modelo

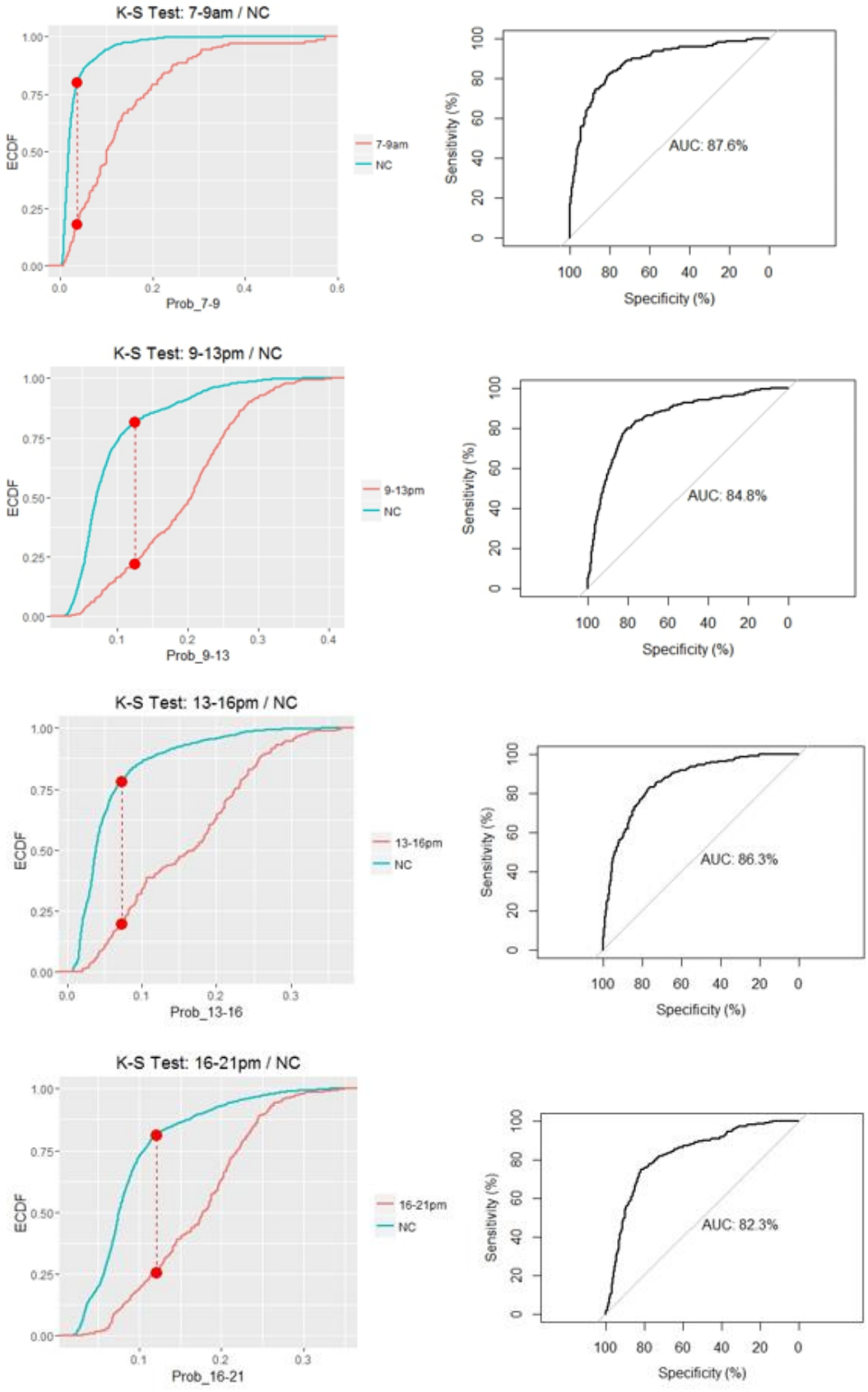

Figura 7. KS y Curva ROC Validación

Tabla 8. Medidas de Calidad de Discriminación

	MODELAMIENTO			VALIDACION		
	KS	AUROC	GINI	KS	AUROC	GINI
7-9am	**0,636**	**0,882**	**0,752**	**0,6268**	**0,876**	**0,765**
9-13pm	0,559	0,836	0,672	0,599	0,848	0,697
13-16pm	**0,596**	**0,863**	**0,727**	**0,593**	**0,863**	**0,726**
16-21pm	0,548	0,831	0,664	0,564	0,823	0,647

En la Figura 7 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 8 se muestran los valores de las medidas de calidad de discriminación para cada horario de la variable respuesta. En la industria se esperan valores superiores a 50% como punto de referencia para los modelos de comportamiento, por lo tanto los resultados muestran una buena capacidad de discriminación entre un horario específico de llamada y el no llamar al cliente.

Por otro lado, en la Tabla 10 se presentan los indicadores globales (KSM y VUS) del modelo. Aquí 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 8 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, celular u oficina, indicador si el cliente realizó una llamada al Call Center y tiempos de llamadas (Bayrak, 2013).

Tabla 9. Matriz de Confusión Validación

$Y \setminus \hat{Y}$	7-9am	9-13pm	13-16pm	16-21pm	NC
7-9am	**3,94**	4,93	10,84	30,05	50,25
9-13pm	0,22	**3,68**	5,63	25,97	64,5
13-16pm	1,2	4,19	**6,89**	35,03	52,69
16-21pm	0,78	3,11	4,66	**27,38**	64,08
NC	0,1	0,62	0,67	3,99	**94,62**

Tabla 10. KSM y VUS

MODELAMIENTO		VALIDACION	
KSM	VUS	KSM	VUS
0,3988	0,669	0,406	0,660

5. CONCLUSIONES

La importancia de trabajarse 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 grande 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, reducciones de tiempo, desarrollo de nuevos productos, estrategias, ofertas optimizadas y toma de decisiones inteligentes.

En este artículo, los autores proponen un modelo estadístico de mejor horario de llamada a los clientes, para aumentar la probabilidad de contactabilidad telefónica en la gestión de cobranzas. La metodología planteada podría ampliarse fácilmente a otras situaciones de gestión e inteligencia de negocios tales como ventas, promociones, gestión de despacho y por tanto se espera contribuir en diferentes áreas de investigación académica aportando con una solución relevante y duradera.

Los resultados sugieren que información histórica de contactabilidad, el día de la semana, características del contrato moroso y la propensión de pago (dada por la razón de saldo en atraso al corto y largo plazo), son determinantes de un contacto telefónico efectivo.

Tomando en cuenta las medidas de KS, AUROC y GINI el mejor modelo, en términos de modelamiento es el horario de 7-9am seguido por el horario de 13-16pm. Siendo así se podrían implementar mayores esfuerzos en estos horarios para aumentar la contactibilidad. Por otro lado, al tomar la máxima probabilidad estimada de contactabilidad para cada individuo, las medidas KSM y VUS disminuyen (lo que es esperado). Incluir variables relevantes (que no fueron incluidas en el presente estudio por ausencia de información) tales como estado civil, profesión, sector de residencia, números de teléfonos activos, si el teléfono es de casa, celular u oficina, indicador si el cliente realizó una llamada al Call Center y tiempos de llamadas, debería mejorar el poder de discriminación entre horarios.

Una forma de obtener mayor flexibilidad con las variables de los modelos (y una oportunidad para futuras investigaciones) 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 apoyándose en árboles de decisión donde se pueden encontrar las interacciones entre las variables que determinen el contacto efectivo en un horario determinado y así obtener variables diferentes en cada modelo logístico binomial. Esta metodología llevaría a construir otro tipo de modelo multinomial que podría dar mejores resultados de discriminación.

6. REFERENCIAS

Adomavicius, G., and Tuzhilin, A. 2005. "Toward the Next Generation of Recommender Systems: A Survey of the State-ofthe-Art and Possible Extensions," IEEE Transactions on Knowledge and Data Engineering (17:6), pp. 734-749.

Anderson, C. 2004. "The Long Tail," WIRED Magazine (12:10) (http://www.wired.com/wired/archive/12.10/tail.html).

Agresti, A. (2007). An Introduction to Categorical Analysis. Florida, United States: Wiley-Interscience.

Ahmed Abbasi, Conan Albrecht, Anthony Vance, and James Hansen (2012). Metafraud: a meta-learning framework for detecting financial fraud1. MIS Quarterly Vol. 36 No. 4, pp. 1293-1327

Arriaza, M. (2006). Guía Práctica de Análisis de Datos. Andalucía, España: IFAPA.

Bayrak, H., Bulbul A., Conser E., Dorai Ch. And Veen A. (2013). Determining best time to reach customers in a multi-channel world ensuring right party contact and increasing interaction likelihood. INTERNATIONAL BUSINESS MACHINES CORPORATION. vol 13, 224-757.

Chen, H. 2011a."Design Science, Grand Challenges, and Societal Impacts," ACM Transactions on Management Information Systems (2:1), pp. 1:1-1:10

Conover, W. J. (1965). Several k-sample Kolmogorov-Smirnov test. JSTOR, vol 36, 1019-1026.

Croissant, Y. (2015). Estimation of multinomial logit models in R: The mlogit packages, Université de la Réunion.

Cunningham P., Martin D., Brick J. M.. 2003. An experiment in call scheduling. Proceedings of the Section on Survey Research Methods, American Statistical Association, 59–66.

Daning Hu, J. Leon Zhao and Zhimin Hua, Michael C. S. Wong (2012). Network-based modeling and analysis of systemic risk in banking systems1. MIS Quarterly Vol. 36 No. 4, pp. 1269-1291

Davenport, T. H. 2006. "Competing on Analytics," Harvard Business Review (84:1), p. 98-107.

Durrant, G., D'Arrigo, J. and 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). pp. 1029-1049

Frauke Kreuter and Gerrit Muller (2014). A Note on Improving Process Efficiency in Panel Surveys with Paradata. Field Methods 2015, Vol. 27(1) 55-65

Hand, D. J. and Till, R. J. (2001) A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Kluwer Academic Publishers, vol 45, 171-186.

Hosmer, D. W. and Lemeshow, S.(2000). Applied Logistic Regression, Ohio, United States: Wiley-Interscience.

Landgrebe, T. and Duin, R. P.(2006). A Simplified Extension of the Area Under the ROC to the Multiclass Domain. Landgebre, vol 1, 241-245.

Loftus, S. C., House, L. L., Hughey, M. C., Walke, M. H, Walke, J. B., and Belden, L. K. (2015). Dimension Reduction for Multinomial Models via Kolmogorov-Smirnov measure (KSM), stat.vt, vol 1, 1-19.

Milone, Giuseppe (2009). Estatística geral e aplicada. Pioneira Thomson Learning.

Michael Chau, Jennifer Xu (2012). Business intelligence in blogs: understanding consumer interactions and communities1. MIS Quarterly. vol. 36 no. 4, pp. 1189-1216

Nachiketa Sahoo, Param Vir Singh, Tridas Mukhopadhyay (2012). A hidden Markov model for collaborative filtering. MIS Quarterly. Volume 36 Issue 4. Pages 1329-1356.

Pang, B., and Lee, L. 2008. "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval (2:1-2), pp. 1-135.

Raymond Y. K. Lau, Stephen S. Y. Liao, K. F. Wong, Dickson K. W. Chiu (2012). Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions. MIS Quarterly Vol. 36 No. 4—Appendices

Robert B. Bendel, A.A. Afifi (2012).Comparison of Stopping Rules in Forward "Stepwise" Regression. Journal of the American Statistical Association. Volume 72. 1977. Issue 357.

Sung-Hyuk Park, Soon-Young Huh, Wonseok Oh, and Sang Pil Han (2012). A Social Network-Based Inference Model for Validating Customer Profile Data. MIS Quarterly. Volume 36. Issue 4. Pages 1217-1237.