UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERIA INDUSTRIAL

METODOLOGÍA PARA ESTIMAR EL IMPACTO QUE GENERAN LAS LLAMADAS REALIZADAS EN UN CALL CENTER EN LA FUGA DE LOS CLIENTES UTILIZANDO TÉCNICAS DE TEXT MINING

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

CATALINA SEPÚLVEDA JULLIAN

PROFESOR GUÍA

ANDRÉS MUSALEM SAID

MIEMBROS DE LA COMISIÓN

SEBASTIÁN RÍOS PÉREZ

MARCEL GOIC FIGUEROA

SANTIAGO DE CHILE

2015

RESUMEN DE LA MEMORIA PARA OPTAR AL
TITULO DE: Ingeniera Civil Industrial
POR: Catalina Sepúlveda Jullian
FECHA: 04/06/2015
PROFESOR GUÍA: Andrés Musalem Said

La industria de las telecomunicaciones está en constante crecimiento debido al desarrollo de las tecnologías y a la necesidad creciente de las personas de estar conectadas. Por lo mismo es que presenta un alto grado de competitividad y los clientes son libres de elegir la opción que más les acomode y cumpla con sus expectativas.

De esta forma la predicción de fuga, y con ello la retención de clientes, son factores fundamentales para el éxito de una compañía. Sin embargo, dados los altos grados de competitividad entre las distintas empresas, se hace necesario innovar en cuanto a modelos de fuga utilizando nuevas fuentes de información, como lo son las llamadas al Call Center. Es así como el objetivo general de este trabajo es medir el impacto que generan las llamadas realizadas en el Call Center en la predicción de fuga de los clientes.

Para lograr lo anterior se cuenta con información de las interacciones que tienen los clientes con el Call Center, específicamente el texto de cada llamada. Para extraer información sobre el contenido de las llamadas se aplicó un modelo de detección de tópicos sobre el texto para así conocer los temas tratados y utilizar esta información en los modelos de fuga.

Los resultados obtenidos luego de realizar diversos modelos logit de predicción de fuga, muestran que al utilizar tanto la información de las llamadas como la del cliente (demográfica y transaccional), el modelo es superior en accuracy en un 8.7% a uno que no utiliza esta nueva fuente de información. Además el modelo con ambos tipos de variables presenta un error tipo I un 25% menor a un modelo que no incluye el contenido de las llamadas.

Tras los análisis realizados es posible concluir que las llamadas al Call Center sí son relevantes y de ayuda al momento de predecir la fuga de un cliente, ya que logran aumentar la capacidad predictiva y ajuste del modelo. Además de que entregan nueva información sobre el comportamiento del cliente y es posible detectar aquellos tópicos que puedan estar asociados con la fuga, lo que permite tomar acciones correctivas.

AGRADECIMIENTOS

Primero quiero agradecer a mi familia por estar siempre apoyándome durante toda mi época universitaria, por creer y confiar en mí y por hacer que yo también confíe en mis habilidades.

A mi Nico que ha sido un apoyo tremendo todo este tiempo y ha estado presente en mis momentos de máximo estrés. Gracias por estar siempre a mi lado y apoyarme en todo lo que hago.

A mis amigos de la U que sin ellos la universidad no hubiera sido tan entretenida, por todos los momentos de risas, carretes, estudio y estrés. Especialmente a la Pame, mi partner de todos los días.

A mis lindas amigas de la vida, que siempre han confiado en mí y esperado lo mejor, sobre todo a la Gina que ha sido un apoyo desde siempre.

A todo el team CEINE por la ayuda y la buena onda, a los profesores Sebastián y Andrés por la buena disposición, el compromiso y todo lo que aprendí gracias a ellos.

TABLA DE CONTENIDO

1. INTRODUCCIÓN

Hoy en día las industrias de telecomunicaciones juegan un rol muy importante en la vida diaria de las personas, ya que, en una primera instancia, otorgan servicios tanto de telefonía como de internet, lo que permite estar conectado con el mundo de una forma rápida y sencilla. Luego, la conectividad ha tenido un gran impacto en la población, sobre todo con la aparición de las redes sociales y el desarrollo tecnológico asociado. Además, según un estudio publicado el año 2014 por la Comisión de Banda Ancha de la International Telecommunication Union (ITU)[1], que es parte de la ONU, alrededor del 40% de la población del mundo tiene acceso a internet, y específicamente para Chile, este porcentaje alcanza el 67,3%.

De esta forma, las empresas de telecomunicaciones nacionales tienen un gran potencial para seguir creciendo y aumentando sus clientes con el paso del tiempo. Cada vez se espera una mayor penetración de estos servicios, ya que según un estudio de la Subsecretaría de Telecomunicaciones (SUBTEL), las conexiones a internet acumulan un crecimiento del 22.4% de Enero a Septiembre del año 2014. Con respecto a la telefonía, los abonados móviles alcanzan 128.6 abonados por cada 100 habitantes en Septiembre 2014, y las líneas de telefonía fija registran un aumento de 5.6% de Diciembre 2013 a Septiembre 2014. La televisión alcanza una penetración del 48.6% en hogares y los suscriptores de TV pagada aumentaron un 6.6% en el último año.

En el Gráfico 1, Gráfico 2 y Gráfico 3, se puede ver la penetración y evolución de los servicios de internet y telefonía tanto fija como móvil en Chile en los últimos años.

[1] http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx

Internet Fijo y Móvil

Evolución de Accesos y Penetración cada 100 hab.

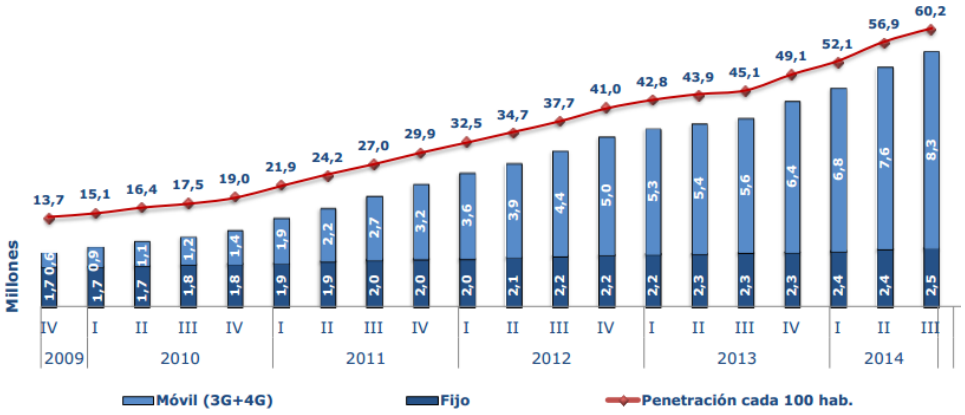


Gráfico 1. Internet Fijo y Móvil, evolución de accesos y penetración cada 100 habitantes.
Fuente: SUBTEL.

Telefonía Móvil

Evolución de abonados y penetración cada 100 hab.

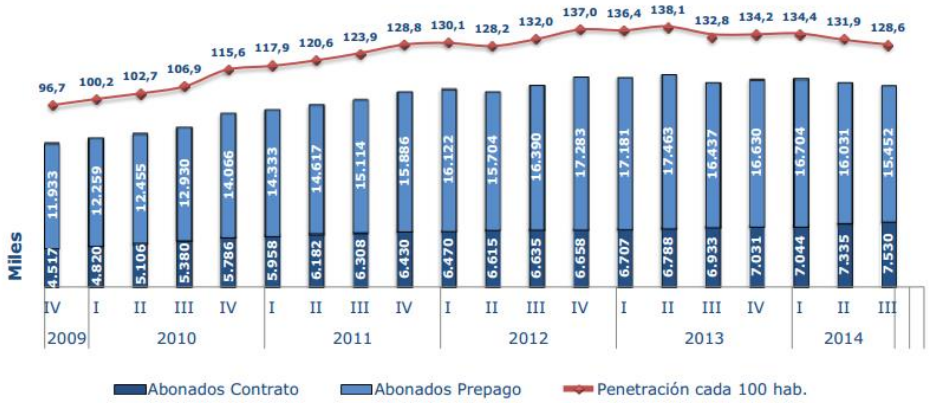


Gráfico 2. Telefonía Móvil, evolución de abonados y penetración cada 100 habitantes.
Fuente: SUBTEL.

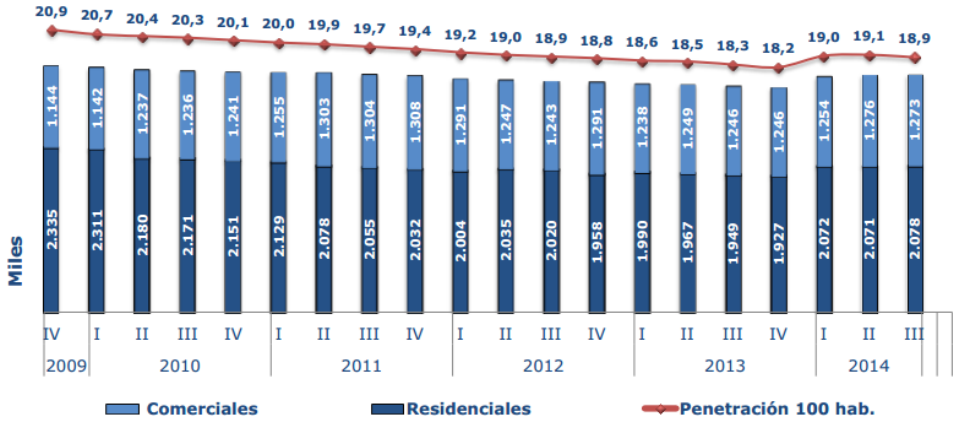


Gráfico 3. Telefonía Fija, evolución líneas fijas y penetración cada 100 habitantes.
Fuente: SUBTEL

De lo anterior se puede extraer que con respecto a las conexiones a internet (tanto fijas como móvil), crecieron un 35.5% entre Septiembre 2013 y Septiembre 2014. Para el caso de telefonía móvil y fija, se observa que las líneas móviles van en crecimiento y las líneas fijas presentan una tendencia a la baja, lo que se puede explicar por el gran desarrollo de las líneas móviles y la conectividad que esto conlleva.

Este crecimiento en las industrias de telecomunicaciones trae consigo un aumento de potenciales usuarios y con ello la necesidad de las empresas de seguir entregando un buen servicio a su cartera de clientes. Para mantener los niveles de servicio y evitar perder a un cliente frente a la competencia, la satisfacción y experiencia del usuario se transforma en un factor fundamental.

Lo anterior se hace más fuerte con la instauración de la Portabilidad Numérica en el año 2011, con la nueva Ley N°20.471[2]. Con esto los clientes tienen más alternativas y son libres de elegir la opción que más les acomode según sean sus necesidades y expectativas, lo que genera una mayor rotación de usuarios entre las distintas empresas de telecomunicaciones.

[2] Ley N°20.471 Portabilidad Numérica. http://www.fne.gob.cl/marco-normativo/otras-leyes/ley-20-471-portabilidad-numerica/

Es por esto que las empresas han tenido que cambiar su enfoque estratégico y comenzar a conocer a sus clientes, desarrollando métodos para fidelizarlos y al mismo tiempo atraer nuevos, ya que ofrecer calidad ya no es suficiente para tener éxito.

Tomando en cuenta lo anterior, y la importancia que tiene para las empresas el nivel de satisfacción de sus usuarios, la SUBTEL realizó un ranking de satisfacción general del usuario con la empresa de telecomunicaciones a la cual éste está afiliado. El Gráfico 4 muestra que la compañía con un mayor porcentaje de satisfacción de sus usuarios es Entel seguido por Nextel, mientras que Claro y Movistar presentan los niveles más bajos.

Gráfico 4. Ranking Satisfacción usuarios telecomunicaciones.
Fuente: SUBTEL

Una de las herramientas para medir la satisfacción de los clientes dentro de una empresa es tomar en cuenta los reclamos que hacen los usuarios. Es por esto que la SUBTEL también incluyó en su estudio la cantidad de reclamos recibidos por empresa. En el Gráfico 5, Gráfico 6, Gráfico 7, Gráfico 8 y Gráfico 9 se muestra la cantidad de reclamos por cada 10.000 usuarios en los distintos servicios entregados por las empresas de telecomunicaciones.

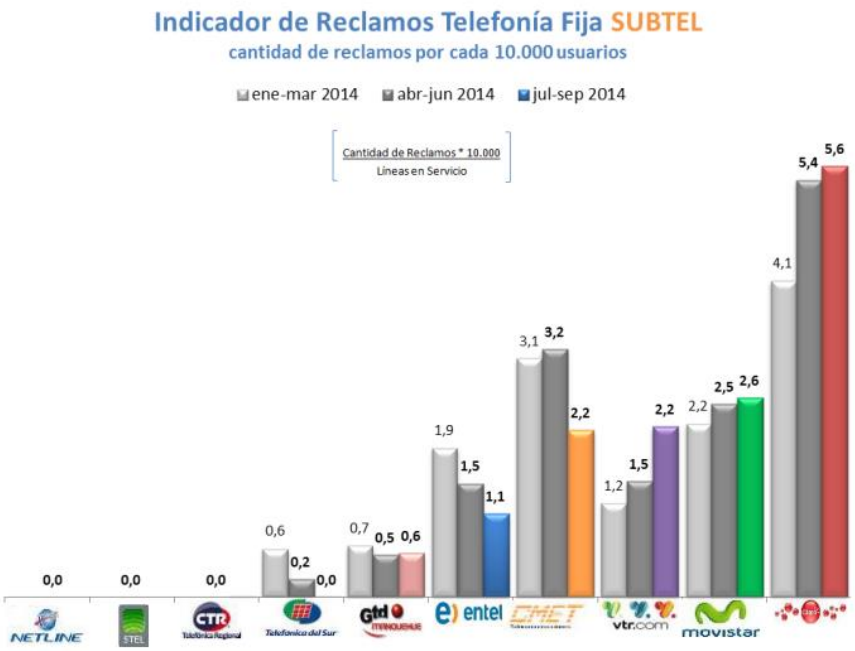


Gráfico 5. Indicador reclamos telefonía fija.
Fuente: SUBTEL

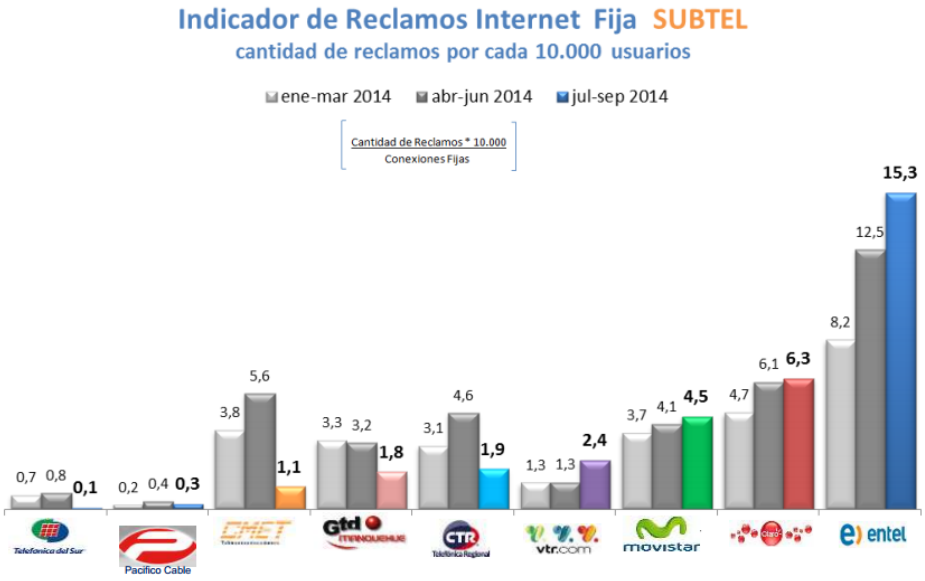


Gráfico 6. Indicador reclamos Internet Fija.
Fuente: SUBTEL

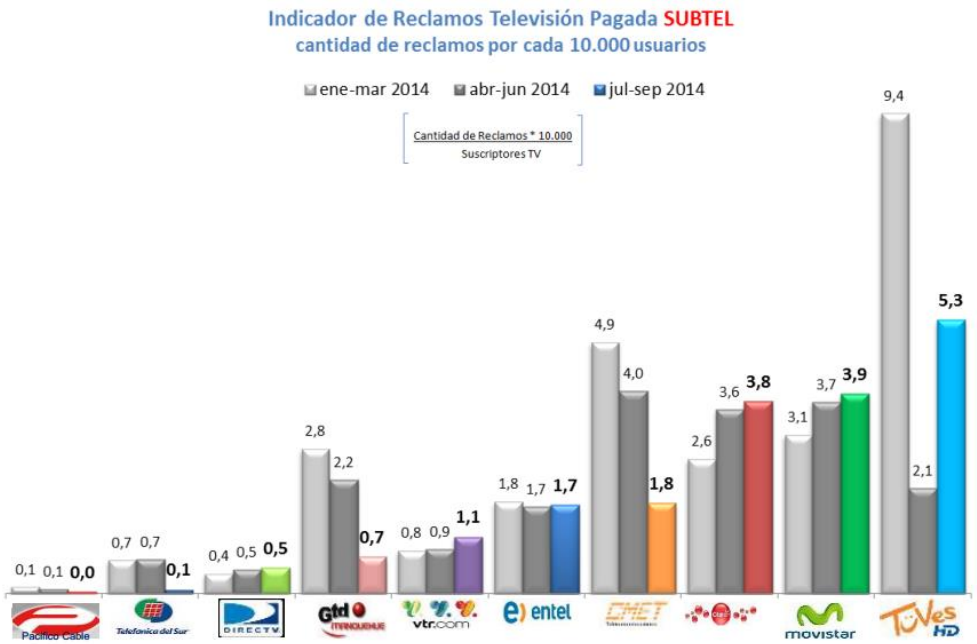


Gráfico 7. Indicador reclamos TV pagada.
Fuente: SUBTEL

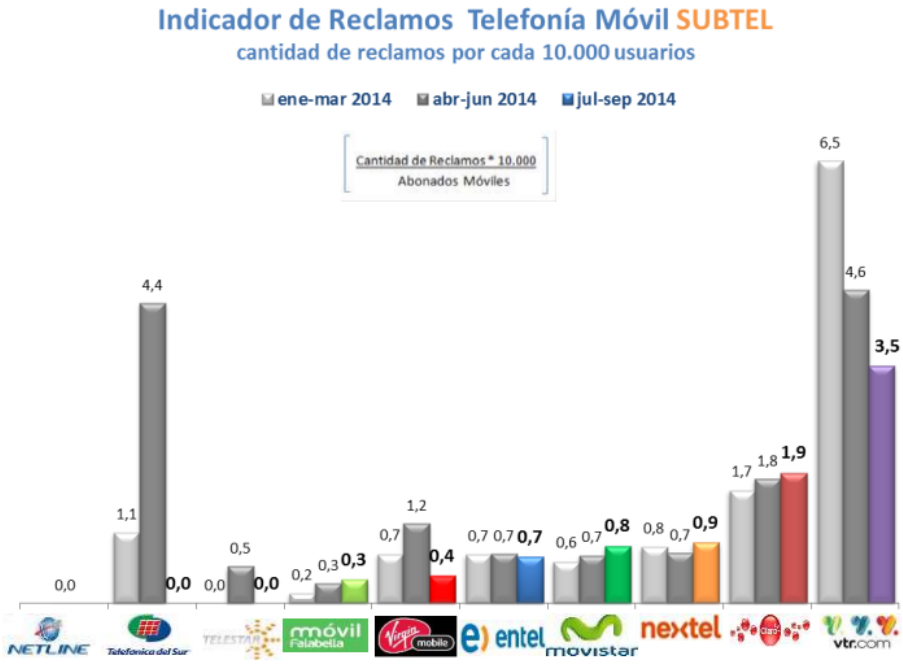


Gráfico 8. Indicador reclamos telefonía móvil.
Fuente: SUBTEL

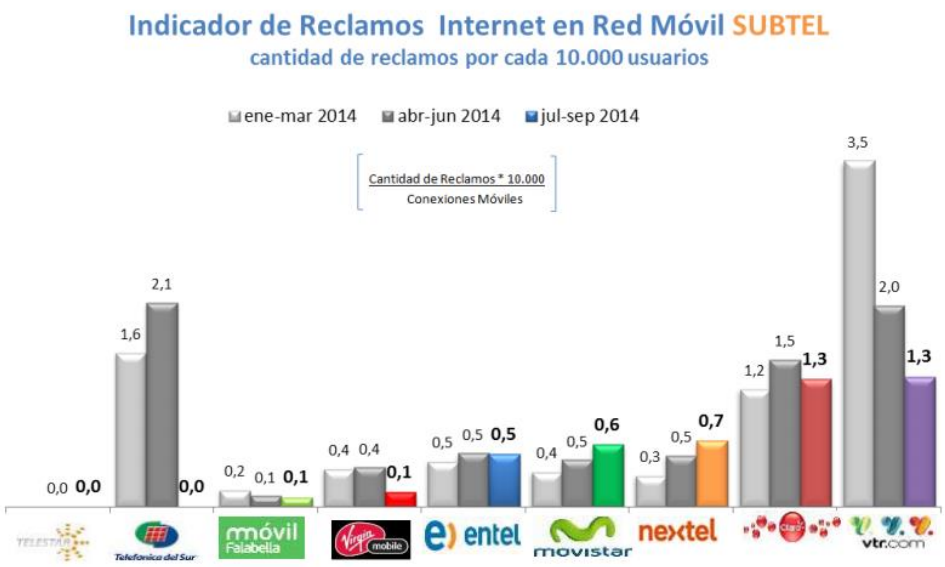


Gráfico 9. Indicador reclamos internet móvil.
Fuente: SUBTEL

Analizando los gráficos se puede ver que en general, se recibe una mayor cantidad de reclamos para los servicios de hogar, es decir, telefonía fija, banda ancha y televisión.

Como se mencionó anteriormente, mantener satisfechos a los usuarios es de gran importancia en el mercado de las telecomunicaciones, ya que además de existir una alta competencia, los usuarios cada vez tienen más información y con ello una mayor facilidad para elegir qué empresa contratar según sean sus necesidades. Del mismo modo, con el desarrollo de la tecnología las mismas empresas cuentan con nuevas formas de fidelizar a sus clientes y conocerlos mejor.

La empresa con la cual se trabajará cuenta con campañas de retención, esto con el objetivo de mantener una cartera de clientes constante en el tiempo. Lo anterior debido a que retener a un cliente tiene un menor costo que buscar uno nuevo [1], por lo que los modelos que logren predecir fuga son de gran importancia.

Para tener éxito en la retención de clientes lo primero es saber quiénes son, para luego saber qué ofrecer, cuándo y cómo. Generalmente en las empresas se almacenan grandes bases de datos con todo tipo de información, pero en la mayoría de los casos no saben cómo sacarle el máximo provecho. Debido a esto no tienen un total conocimiento de quienes

son sus clientes, lo que conlleva a un mal servicio y eventualmente posteriores reclamos por parte de los consumidores.

Dentro de esta línea, el Call Center de una empresa es un importante canal de atención y de relación con el cliente, es una interesante fuente de información y una excelente palanca para la creación y gestión de la experiencia y satisfacción de los usuarios. La empresa con la cual se trabajará cuenta con servicios de Call Center, sin embargo la información que se puede extraer no está siendo utilizada en las campañas y modelos de retención de clientes.

En conclusión, esta memoria tiene como objetivo medir el impacto que generan las llamadas realizadas en un Call Center en la predicción de fuga de los clientes. Esto con el fin de cuantificar si esta información puede mejorar la capacidad predictiva de los modelos que no la utilizan, y así estimar la importancia que tiene la interacción con el Call Center en la decisión del cliente de eliminar algún contrato con la compañía.

2. DESCRIPCIÓN DEL PROYECTO Y JUSTIFICACIÓN

Hoy en día los clientes son libres de elegir la opción que más les acomode según sus necesidades, por lo que hay una mayor rotación de usuarios y competencia entre las distintas empresas de telecomunicaciones.

Dado lo anterior, es de total importancia el desarrollo de modelos de fuga que permitan predecir si un cliente se dará de baja o no, de modo de poder realizar acciones previas y evitar perder al cliente.

La empresa con la que se trabajará cuenta con un alto grado de contacto con sus usuarios, debido a que ofrece productos en el área de telecomunicaciones con sus respectivos servicios de instalación y además cuenta con servicios de atención al cliente mediante un Call Center.

Dentro de la compañía se utilizan modelos de retención basados en promociones y descuentos para así disminuir sus tasas de CHURN. Actualmente en lo que respecta a la línea de negocios Móvil se utilizan modelos de predicción de fuga basados en información transaccional y demográfica de los clientes. Sin embargo, para la línea de negocios Fijo, es decir, telefonía fija, televisión e internet, no están siendo utilizados. Lo

anterior se debe principalmente a una falta de desarrollo de éstos, ya que en algún momento sí se trabajó con modelos de fuga basados en información demográfica y transaccional.

Otro tipo de información que podría utilizarse en la construcción de modelos de fuga, es la respectiva a las llamadas realizadas al Call Center. Esta información podría entregar un conocimiento más global de los clientes y sus inquietudes, otorgando nuevas herramientas que permitan aumentar las tasas de retención de clientes propensos a fugarse.

Para estudiar y analizar el impacto que tienen las interacciones con el Call Center en la predicción de fuga de los clientes, se trabajará con la información de clientes con contrato de productos hogar, es decir, que cuenten con telefonía fija, internet y/o televisión.

Las tasas de CHURN que presenta la compañía durante Marzo y Septiembre del año 2014 se pueden ver en el Gráfico 10, el cual está separado entre los distintos productos hogar donde *BAF* es banda ancha fija, *TV* es televisión y *VOZ* es telefonía fija.

El análisis a realizar estará enfocado específicamente en las bajas del mes de Agosto, donde la tasa de CHURN para internet, telefonía y televisión son respectivamente 2.26%, 1.97% y 2.51%. El análisis se centrará específicamente en este mes debido a la disponibilidad de la información, las llamadas al Call Center que se tienen corresponden al mes de Julio. De este modo para lograr relacionar las llamadas recibidas con una posible baja es que se analizará el mes de Agosto.

Cabe destacar que este gráfico solo considera contratos de clientes *personas*, es decir personas naturales, las empresas no serán estudiadas en la presente memoria.

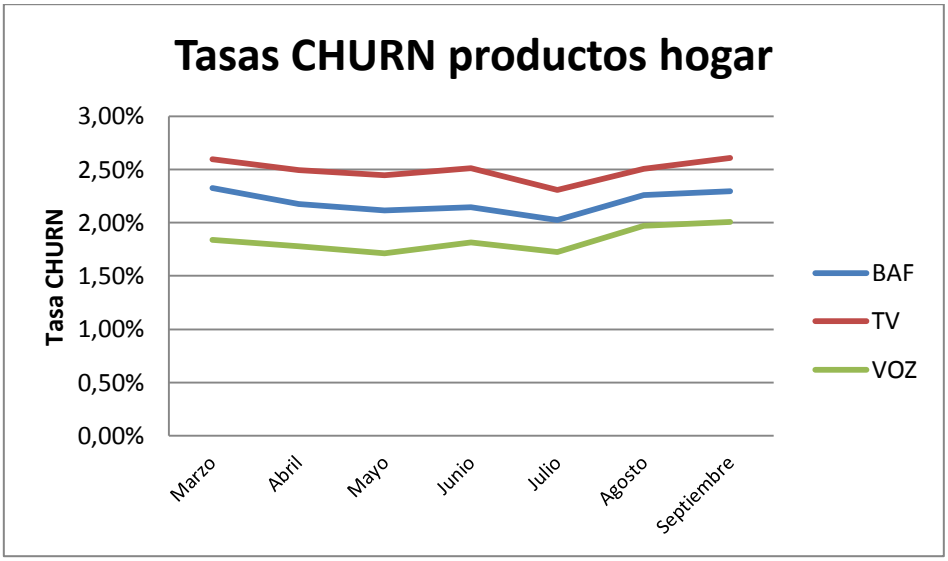


Gráfico 10: Tasa CHURN para productos hogar.
Fuente: Elaboración propia.

Las curvas para los tres productos son bastante similares, sin embargo difieren en magnitud; donde la televisión es el producto con un mayor porcentaje de bajas.

Analizando el mes de Agosto, las principales razones por las que los clientes dan de baja uno o más de sus productos se muestran en el Gráfico 11, donde se puede ver que la mayoría se está fugando por razones morosas, es decir, la misma empresa corta el contrato por deudas impagas.

Dentro del marco de esta memoria, el análisis se centrará solo en aquellos clientes que decidieron dar de baja uno o más productos por razones voluntarias o de portabilidad. La primera razón se refiere a cuando el mismo cliente termina el contrato con la empresa y se adhiere a la competencia, la segunda a que es la competencia la que notifica que un cliente se portó, y por lo tanto, ya no pertenece a la compañía.

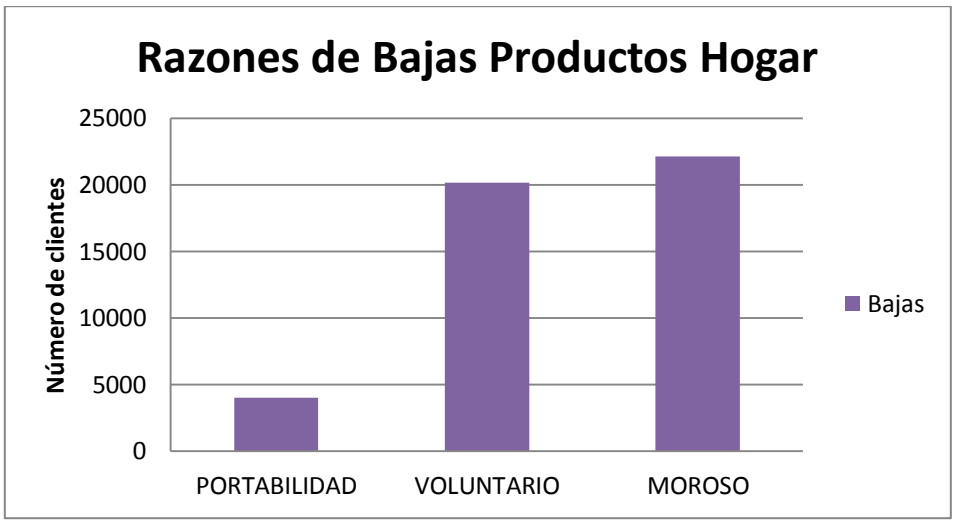


Gráfico 11: Razones de bajas de productos hogar durante Agosto 2014.
Fuente: Elaboración propia.

De esta forma, tomando solo aquellos clientes que se van por razones voluntarias o portabilidad, se tiene que en el mes de Agosto (en el cual se sitúa el análisis) el 52.20% de los fugados corresponde a una de estas dos categorías. De este porcentaje de clientes fugados, las nuevas tasas de CHURN para televisión, telefonía e internet son 1.22%, 1.08% y 1.18% respectivamente, donde la televisión sigue siendo la con mayor tasa de fuga.

Como análisis preliminar, es posible realizar una comparación entre los clientes que llaman al Call Center y los que no, de modo de determinar el rol que juega el Call Center en la fuga de los clientes.

Utilizando la información de los clientes que llaman al Call Center durante los meses de Julio y Agosto, y aquellos que cuentan con algún producto hogar contratado (televisión, internet y/o telefonía) se puede determinar qué porcentaje de ellos da de baja alguno de sus productos hogar y quiénes los conservan. En la Tabla 1, se muestra el porcentaje de clientes que llaman al Call Center en los meses de Julio y Agosto, diferenciando entre clientes catalogados como fugados y activos. De esta forma, de todos los clientes que dan de baja al menos un producto hogar durante el mes de Agosto, el 5.61% de ellos llama el mes anterior al Call Center.

Realizando una comparación, en el mes de Agosto el porcentaje de clientes que llaman al Call Center es similar en los casos de aquellos que dan de baja algún producto y los que no lo hacen. Sin embargo, igualmente es mayor para el grupo de los fugados. Para el mes de Septiembre la diferencia se hace más notoria, donde se puede ver claramente que dentro del grupo de

los clientes catalogados como fugados, un 6.13% llama al Call Center el mes anterior de dar de baja un producto, mientras que solo el 3.56% de los clientes que conservan sus productos realiza llamadas el mes previo.

Tabla 1. Porcentaje de clientes que llaman al Call Center y que al mes siguiente está catalogado como fugado o activo.

	Fugados al mes siguiente	Activos al mes siguiente
Llaman al Call Center en Julio	5.61%	5.25%
Llaman al Call Center en Agosto	6.13%	3.56%

De este modo, dentro del grupo de clientes catalogados como fugados se puede ver que hay un porcentaje mayor de éstos que realiza llamadas al Call Center el mes anterior a la baja, comparando con el grupo de clientes que conservan sus productos de un mes al otro. Así, con el simple hecho de saber si un cliente llamó o no se podría capturar bastante bien el fenómeno y se podría encontrar una relación con la posible fuga.

Sin embargo, también se tiene disponible el texto de la llamada, lo que agregaría valor al análisis a realizar ya que se podrían encontrar altas correlaciones entre un tema específico en la llamada con la fuga de un cliente. De esta forma, se podría obtener un cierto número de tópicos presentes en las llamadas, que corresponden a los temas tratados en la conversación, con distintos grados de correlación con la baja de un producto.

En conclusión, la información que se pueda obtener de las interacciones con el Call Center puede ser beneficiosa con respecto a la predicción de fuga, ergo, es interesante realizar un análisis en torno a lo anterior.

Es así como esta memoria tiene como objetivo analizar si la información disponible correspondiente a las interacciones de los clientes con el Call Center es relevante o no a la hora de predecir una baja. Es decir, si las llamadas realizadas por los clientes, y lo que se habla en ellas, puede mejorar la capacidad predictiva de los modelos de fuga que no utilicen esta información.

Para lograr lo anterior se construirán distintos modelos de predicción de fuga con y sin las variables de las interacciones con el Call Center, de modo de realizar comparaciones y poder estimar la importancia de las llamadas en cuanto a predicción de fuga.

3. OBJETIVOS

A continuación se presenta el objetivo general junto con sus objetivos específicos.

3.1. Objetivo General

"Medir el impacto que generan las llamadas realizadas en un Call Center en la predicción de fuga de los clientes"

3.2. Objetivos Específicos

- Identificar y definir variables que describan las llamadas registradas y los cambios en el comportamiento de los clientes.
- Construir un modelo de fuga en base a información transaccional y demográfica del cliente.
- Identificar tópicos presentes en las llamadas al Call Center y su relación con la posible fuga del cliente.
- Construir un nuevo modelo de predicción de fuga agregando información relativa a las llamadas realizadas en el Call Center.
- Evaluar el ajuste y capacidad predictiva de los modelos para describir la probabilidad de fuga de los clientes.
- Definir la importancia y relevancia que tienen las interacciones con el Call Center en la predicción de fuga de los clientes.

4. METODOLOGÍA

Con el objetivo de ordenar los pasos metodológicos y darle un orden lógico, se utilizará la metodología de procesamiento de datos CRISP-DM (Cross-Industry Standard Process for Data Mining). Esta consta de 6 pasos principales: Comprensión del problema, Estudio y comprensión de los datos, Preparación de los datos, Modelamiento, Evaluación y finalmente Utilización [2] [3].

4.1. Comprensión del Problema

En la empresa de telecomunicaciones con la cual se trabajará cuenta con información de las llamadas al Call Center que no está siendo explotada. Esta información está disponible mediante un software de Speech Analytics llamado IMPACT 360, el cual guarda la grabación de cada llamada realizada en el Call Center y luego la transcribe automáticamente.

Dado lo anterior, se tiene disponible información potencialmente valiosa pero no se le ha dado suficiente uso. Actualmente este software no está siendo utilizado por la empresa, por lo que para decidir si explotar esta información y ocupar más recursos en el desarrollo y uso del software, es necesario hacer una investigación y un análisis que demuestre si efectivamente agrega valor a los modelos de fuga.

Los datos que se utilizarán para lograr lo anterior corresponden a clientes que tengan algún contrato de productos hogar durante Julio 2014 (telefonía fija, internet y/o televisión) y que hayan realizado alguna llamada al Call Center en este mismo periodo. Se trabajará con contratos hogar debido a la disponibilidad de la información, ya que se cuenta con una mayor cantidad de llamadas al Call Center de usuarios con productos hogar, y al ser tan limitada la información disponible es que se decidió trabajar con esta línea de negocios.

El análisis que se llevará a cabo parte con la extracción de las llamadas registradas en el software ya mencionado, de esta forma se contará con los datos de los clientes que hayan hecho llamadas al Call Center, y también con el texto de la conversación.

Luego se construirá un modelo de predicción de fuga sin utilizar la información del Call Center, es decir, usando solo los datos demográficos y transaccionales de los clientes.

El siguiente paso consiste en construir un modelo de predicción de fuga que utilice la información de las llamadas. Al tener la transcripción de cada llamada se aplicarán técnicas de text mining con el fin de reconocer los temas tratados en cada conversación y así poder extraer información de forma sistematizada. Luego se definirán variables independientes que logren describir de mejor forma la llamada realizada y así encontrar qué tópicos son los que conducen con mayor probabilidad a la fuga.

Finalmente, se hará el mismo procedimiento anterior combinando ambos tipos de variables (de las llamadas y del cliente), con el fin de comparar los modelos obtenidos y así poder estimar el impacto de las llamadas realizadas al Call Center con el comportamiento de fuga.

4.2. Estudio y Comprensión de los Datos

4.2.1. Análisis descriptivo y limpieza de datos

Como primer paso se realizará un análisis descriptivo de los datos disponibles, con el primer objetivo de detectar posibles errores en la data como por ejemplo valores faltantes, datos duplicados, outliers, entre otros. En caso de detectar estos errores se realizará una limpieza de modo de obtener una data válida, no ambigua y consistente. El segundo objetivo del análisis descriptivo es familiarizarse con los datos y así tener una idea de la forma que tienen, es decir, distribuciones de probabilidad, parámetros de centralización como media, mediana y moda; y parámetros de dispersión como varianza y desviación típica.

4.2.2. Definición de variables

Se definirán 3 tipos de variables:

1. **Variables de las llamadas:** Estas variables tienen como objetivo describir las llamadas registradas en el Call Center. Las variables a utilizar son:

   ANI: Sigla *para Automatic Number Identification*. Variable que define el número desde el cual el cliente está realizando la llamada.
   Duración: Se define como el tiempo en que el ejecutivo contesta el teléfono hasta que termina la llamada. Si el cliente tiene más de una llamada registrada, corresponde al promedio de las duraciones.
   Número de Llamadas: Cantidad de veces que un cliente ha llamado al Call Center en un tiempo determinado.
   Fecha: Variable que guarda el día y hora en que se realizó la llamada al Call Center.
   Categorías: Variable categórica que etiqueta los temas presentes en cada llamada.

Tópico: Variable a crear luego de aplicar modelo de detección de tópicos. Variable numérica que definirá a partir de probabilidades los temas presentes en cada llamada.

2. **Variables Demográficas y Transaccionales:** Estas variables describen al cliente y los productos que tiene contratado, ya sea telefonía fija, internet hogar o televisión. Las variables a utilizar son:

   Mes: Corresponde al mes de estudio, particularmente Julio 2014.
   Teléfono: Número de teléfono asociado al cliente.
   RUT: Rut del cliente.
   Área: Número de área asociado al número de teléfono.
   Edad: Edad en años del cliente.
   Región: Región a la que pertenece el cliente.
   Gse: Grupo socioeconómico.
   Antigüedad Cliente: Cantidad de meses que el cliente lleva afiliado a la compañía desde sus inicios.
   Antigüedad Línea: Cantidad de meses que el cliente lleva contratando telefonía fija.
   Antigüedad Banda Ancha: Cantidad de meses que el cliente lleva contratando internet hogar.
   Antigüedad TV: Cantidad de meses que el cliente lleva contratando televisión.
   Voz: Binaria que indica si el cliente tiene contrato de telefonía fija o no.
   Banda Ancha: Binaria que indica si el cliente tiene contrato de internet fija o no.
   Televisión: Binaria que indica si el cliente tiene contrato de televisión o no.
   ARPU total: Promedio de ingresos que la compañía obtiene por cada cliente cada mes.
   Cargo total: Monto total a pagar por el cliente cada mes.
   ARPU Voz: Promedio de ingresos debidos a contratos de telefonía fija que la compañía obtiene por cada cliente.
   ARPU Banda Ancha: Promedio de ingresos debidos a contratos de internet fija que la compañía obtiene por cada cliente.
   ARPU TV: Promedio de ingresos debidos a contratos de televisión que la compañía obtiene por cada cliente.

Minutos Salientes: Cantidad de minutos salientes que utiliza el cliente al mes.
Minutos Entrantes: Cantidad de minutos entrantes que recibe el cliente al mes.
Llamadas Salientes: Cantidad de llamadas que realiza el cliente al mes.
Llamadas Entrantes: Cantidad de llamadas que recibe el cliente al mes.
Días Mora: Cantidad de días que el cliente lleva con deuda.
Titular: Binaria que indica si el cliente es el titular de la cuenta o no.
Reclamos: Cantidad de reclamos comerciales que el cliente registra por mes.
Interacciones: Cantidad de interacciones comerciales que el cliente registra con la empresa por mes.

3. Variable Dependiente: La variable dependiente a utilizar es una variable binaria llamada Fuga, que toma la siguiente forma:

$$Fuga = \begin{cases} 1, & Si\ cliente\ da\ de\ baja\ alguno\ de\ sus\ productos\ hogar\ contratados \\ 0, & Si\ no \end{cases}$$

4.3. Preparación de los Datos

Con respecto a la información demográfica y transaccional, se preparará para el análisis realizando imputaciones en los casos que sean necesarios y aplicando transformaciones a las variables que lo necesiten.

Por otro lado para la información de las interacciones con el Call Center es necesario aplicar técnicas más sofisticadas. Para poder extraer información de las llamadas registradas y lograr definirlas, se utilizará un modelo probabilístico de detección de tópicos llamado Latent Dirichlet Allocation (LDA). Este modelo encontrará los tópicos presentes en cada llamada y le asignará una probabilidad, con lo que se podrá tener conocimiento sobre lo que se habló en la llamada y así clasificarla.

El modelo LDA es un algoritmo ya definido al que se le ingresa como parámetro el número de tópicos que se desea encontrar en cada documento, y luego éste entrega los tópicos encontrados con sus respectivas probabilidades. De este modo se podrá definir la variable **Tópico**, y así clasificar las llamadas realizadas.

Este algoritmo se aplicará sobre los datos entregados por la empresa con la cual se trabajará, es decir, sobre las llamadas en formato de texto realizadas durante Julio 2014 para una muestra aleatoria de datos. Es importante realizar este paso en la metodología ya que es necesario poder definir la variable **Tópico** para tener conocimiento sobre el tema de la llamada y así poder clasificar ciertos patrones o comportamientos, además de caracterizar la llamada.

4.4. Modelamiento

Al tener las variables ya definidas y las llamadas caracterizadas, se procederá a realizar análisis bivariados y multivariados, con el propósito de definir hipótesis a corroborar y conocer el impacto que generan las llamadas realizadas en el Call Center en el comportamiento de fuga de los clientes. Para esto se realizará un modelo Logit binario, el cual tendrá como variable dependiente si el cliente se fugó o no.

Para poder definir si la información de la interacción con el Call Center agrega valor a los modelos de fuga que no la incluyan, el modelamiento se dividirá en 3 etapas:

- Primero se realizará un modelo Logit Binario cuyas variables independientes sean solo variables que expliquen al cliente, es decir, variables tanto transaccionales como demográficas. Luego se escogerá la combinación de variables que formen el mejor modelo.
- Como segundo paso, se realizará un modelo Logit Binario cuyas variables independientes sean exclusivamente de las llamadas realizadas. Con todas las variables que describen las llamadas se escogerá aquella combinación que entregue la mejor capacidad predictiva y ajuste del modelo.
- Finalmente, se hará un Logit Binario juntando ambos tipos de variables, las de las llamadas y los clientes, de modo de analizar si el modelo puede mejorar o no.

El objetivo de lo anterior es analizar si incluir variables de las llamadas agrega valor o no a los modelos de fuga y ver las diferencias existentes en los modelos dependiendo de las variables que estos contengan.

4.5. Evaluación

Para evaluar el ajuste del modelo se utilizarán métodos conocidos y estándares, como son:

- Criterio de Información de Akaike (AIC): $2k - 2\ln(L)$
- Criterio de Información Bayesiano (BIC): $-2\ln(L) + kln(n)$
- Error cuadrático medio (MSE):

$$\frac{1}{n}\sum_{i=1}^{n}(\overline{Y_i} - Y_i)^2$$

  Donde n corresponde al tamaño de la muestra, $\overline{Y_i}$ es un vector de n predicciones y Y_i es el vector de los valores verdaderos.

Para la capacidad predictiva del modelo se utilizará una muestra de testeo como grupo de control, de modo de probar el modelo y verificar que se cumplan los patrones entregados. Esto se realizará a través de matrices de confusión y curvas ROC.

4.6. Utilización

Finalmente, al tener los 3 modelos de fuga, se podrá escoger aquel modelo con un mejor ajuste y capacidad predictiva, lo que permitirá estimar el impacto que genera la información respectiva a las llamadas realizadas en el Call Center.

De esta forma, se podrá definir si la información relativa a las llamadas, que hoy no está siendo utilizada, es relevante en cuanto a la predicción de fuga y, por lo tanto, si es necesario comenzar a explotarla y utilizarla. Esto puede tener gran valor para la compañía en estudio, ya que hoy en día cuentan con las herramientas necesarias para agregar la interacción con el Call Center a sus estudios y modelos, solo es necesario saber de qué forma incorporar esta nueva fuente de información y cómo procesarla.

En conclusión, si el presente estudio tiene resultados positivos en cuanto a la inclusión de las interacciones con el Call Center, la empresa podrá comenzar a utilizar esta información, desarrollar nuevos modelos y con esto disminuir el porcentaje de fuga actual.

5. MARCO CONCEPTUAL

5.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) es un modelo probabilístico de detección de tópicos. Para describir este modelo es necesario definir ciertos conceptos básicos:

- **Palabra:** Una palabra es una unidad básica de información discreta que en este contexto se define como un elemento de un vocabulario indexado V. Para efectos del modelo, las palabras son representadas como vectores unitarios que tienen sólo una componente en 1 y todas las otras en 0. Así, la n-ésima palabra del vocabulario se define como un vector w de largo $|V|$ tal que $w^n = 1$ y $w^u = 0$ para todo $u \neq n$.

- **Documento:** Un documento es una secuencia de N palabras denotadas por $w = (w_1, w_2, w_3, \dots, w_N)$, donde w_n es la n-ésima palabra en la secuencia.

- **Corpus:** Un corpus es una colección de M documentos denotado por $D = \{w_1, w_2, \dots, w_M\}$.

- **Tópico:** Un tópico es una distribución de probabilidad sobre un vocabulario fijo. Para una colección de documentos, se asume que estas distribuciones están dadas de manera previa a la generación de cualquier documento. Por ejemplo el tópico naturaleza tiene las palabras árbol, río, mar y montañas con una alta probabilidad.

De esta forma, LDA es un modelo generativo para colecciones de datos discretos como por ejemplo un *corpus*. La noción tras este modelo es que los documentos se representan en forma de mezclas al azar sobre tópicos, donde cada tópico se caracteriza por una distribución de palabras [4].

Así, cualquier colección de documentos con la que se esté trabajando puede presentar múltiples tópicos y además, estos últimos no están relacionados entre sí. De este modo se crea una especie de cluster de tópicos o temas presentes en el documento. Es importante mencionar que los tópicos que entrega el modelo están compuestos por *bags of words*, es decir, las

palabras dentro de cada uno no tienen un orden lógico con respecto a los documentos a los que pertenecen [5].

El modelo asume el siguiente proceso para cada documento w en un cuerpo D:

1. Seleccionar $N \sim Poisson(\xi)$.
2. Seleccionar $\theta \sim Dirichlet(\alpha)$.
3. Para cada una de las N palabras en el documento w:
    a. Escoger un tópico $z_n \sim Multinomial(\theta)$.
    b. Escoger una palabra w_n de $p(w_n|z_n,\beta)$, que es una probabilidad multinomial condicionada por el tópico z_n.

Donde ξ corresponde al parámetro de la distribución de Poisson, α es el parámetro de la distribución de Dirichlet, z_n corresponde al tópico asociado a la palabra n-ésima, θ_d es la distribución de tópicos para el documento d y β es la matriz $(k \times V)$ de probabilidades de aparición de una palabra (columnas) en un tópico (filas), donde $\beta_{ij} = p(w^j = 1|z^i = 1)$.

De modo que el primer paso corresponde a definir N para representar el número de palabras en un documento. Donde ξ se puede estimar definiendo una distribución a priori y utilizando la cantidad de palabras en cada documento.

Luego, se selecciona la distribución de los tópicos según una distribución Dirichlet con parámetro α, donde α es un vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$. Usando el mismo valor para todos los elementos de α se especifica que a priori todos los tópicos son igualmente probables.

Al tener la distribución de los tópicos, para cada una de las N palabras en el documento w se escoge uno de los K tópicos obtenidos. Así el último paso refleja que cada palabra en un documento se extrae de uno de los K tópicos en proporción a la distribución de los documentos sobre tópicos.

Para el parámetro α se utiliza el valor $50/k$ [6], se define este valor ya que para $\alpha > 1$ se tiene incidencia de múltiples tópicos en cada documento. De este modo se tendrán valores de α mayores a 1 ya que generalmente no se utilizan más de 50 tópicos. Es así como la intuición detrás del valor de alfa es que a valores más altos es más probable que cada documento tenga una gran cantidad de tópicos con alta probabilidad, mientras que a valores más

bajos es más probable que cada documento presente pocos tópicos con alta probabilidad o incluso solo un tópico dominante.

El parámetro β se puede definir como $\beta \sim Dirichlet(\eta)$, donde cada fila representa una distribución Dirichlet, y η toma el valor de 0.1 [6]. Se utiliza un valor pequeño de η ya que así cada tópico tendrá un conjunto de palabras más probables más pequeña, lo que lo hace más simple de interpretar. En cambio valores grandes de η significa que es más probable que cada tópico tenga una mezcla de la mayoría de las palabras. El parámetro β se estima en el modelo a partir de los datos y el número de tópicos asignado (k).

Teniendo los parámetros α y β, la distribución conjunta de las probabilidades de aparición de los tópicos en cada documento θ, el conjunto de K tópicos z y el conjunto de N palabras w, se puede expresar por:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$$

Integrando sobre θ y sumando sobre z, se obtiene la distribución marginal de un documento:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta)\right)d\theta$$

Finalmente, tomando el producto de las probabilidades marginales de los documentos individuales, se obtiene la probabilidad del corpus:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)\right)d\theta_d$$

Donde M es la cantidad de documentos presentes en el corpus.

Es posible representar el modelo LDA gráficamente como se observa en la Figura 1. Donde los nodos representan variables aleatorias, las flechas indican posible dependencia, las variables observables están sombreadas y las cajas representan estructuras replicadas (es decir, de 1 hasta N). Por lo tanto, cada caja representa el proceso de elección de un elemento, la caja exterior representa los documentos dentro del corpus, y la caja interior el conjunto de tópicos y palabras dentro de un documento. Así es posible notar

que w es la única variable observable, ya que lo único que se tiene es un conjunto de palabras ordenadas por documentos.

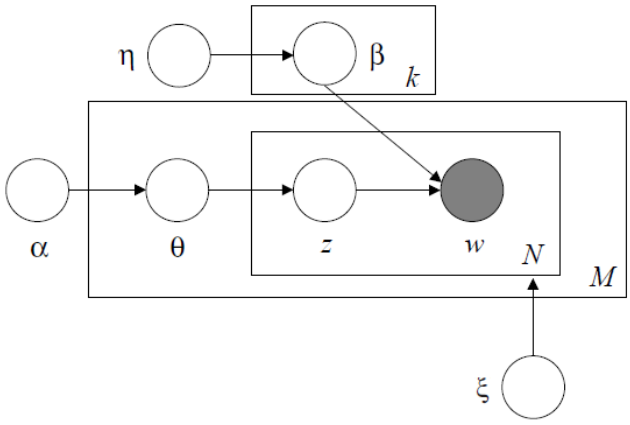


Figura 1. Representación gráfica del modelo LDA.

De esta forma se tienen 3 niveles en la representación del modelo, el corpus, los documentos y el conjunto de tópicos con las palabras. Los parámetros α y β están en el nivel del corpus, por lo que son estimados en el proceso de generar el corpus, y η corresponde al parámetro que determina la distribución de β. Las variables θ_d están en el nivel de los documentos, y se estiman para cada documento. Finalmente las variables z_{dn} y w_{dn} se encuentran en el nivel de las palabras y son estimadas para cada palabra en cada documento, donde el total de palabras N queda determinado por ξ. Es así como los tópicos son estimados repetidas veces para cada documento, por lo que los documentos pueden estar asociados con múltiples tópicos.

El problema de inferencia clave que es necesario resolver para poder usar LDA es lograr calcular la distribución posterior de las variables ocultas (aquellas que no están sombreadas en la Figura 1) para un documento, que se puede expresar como sigue:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}$$

Dada la naturaleza de la ecuación, en general esta distribución no es posible de calcular. La cantidad de estructuras de tópicos que se pueden encontrar en un corpus crece de manera exponencial y provoca que el denominador sea imposible de calcular.

Al no poder calcular la distribución posterior mediante inferencia exacta es necesario utilizar algoritmos estadísticos que la permitan estimar, uno de los

más utilizados es Gibbs Sampling, que será el aplicado en la presente memoria.

Finalmente, la principal ventaja de modelos probabilísticos generativos como el LDA es su modularidad y su extensibilidad, ya que se puede utilizar fácilmente en modelos más complejos [2].

5.2. Modelo Logit

El modelo Logit utiliza alternativas de elección discretas, las cuales tienen que cumplir con ser mutuamente excluyentes, exhaustivas y con un número finito de alternativas.

Para describir la probabilidad con que un agente n elegirá la alternativa i se asume que el individuo decide entre las alternativas maximizando su utilidad. De esta forma la utilidad será conocida por el agente, pero no por el analista, de modo que será necesario descomponer la utilidad en una componente determinística y otra estocástica. Es así como se obtiene lo siguiente:

$$P_{ni} = P\left(u_{ni} > u_{nj}, \forall j \neq i\right)$$

$$P_{ni} = P\left(v_{ni} + \ \varepsilon_{ni} > v_{nj} + \varepsilon_{nj}, \forall j \neq i\right)$$

$$P_{ni} = P(\varepsilon_{nj} - \varepsilon_{ni} < v_{ni} - v_{nj}, \forall j \neq i)$$

$$= \int \mathbf{1}_{[\varepsilon_{nj}-\varepsilon_{ni}<v_{ni}-v_{nj}, \forall j \neq i]} f(\varepsilon_n) d\varepsilon_n$$

De esta forma, el modelo Logit se obtiene a través de asumir que cada término ε_{ij} es independiente e idénticamente distribuido a través de una distribución de valor extremo o Gumbel. La función de densidad y de probabilidad acumulada de esta distribución se define a continuación:

$$f\left(\varepsilon_{ij}\right) = e^{-\varepsilon_{ij}} e^{-e^{-\varepsilon_{ij}}}$$

$$F\left(\varepsilon_{ij}\right) = e^{-e^{-\varepsilon_{ij}}}$$

Con estas funciones es posible derivar el modelo Logit y llegar a la siguiente fórmula cerrada:

$$P_{ni} = \int \left(\prod_{j \neq i} e^{-e^{-(\varepsilon_{ni} + v_{ni} - v_{nj})}} \right) e^{-\varepsilon_{ni}} e^{-e^{-\varepsilon_{ni}}} d\varepsilon_{ni}$$

$$P_{ni} = \int_{-\infty}^{0} - \exp\left(-t \sum_{j} e^{-(v_{ni} - v_{nj})} \right) dt$$

$$P_{ni} = \frac{e^{v_{ni}}}{\sum_j e^{v_{nj}}}$$

De esta forma el modelo indicará la probabilidad de que el cliente n escoja la alternativa i, o de lo contrario que elija al alternativa j, que se representa por $1 - P_{ni}$.

Lo anterior describe el caso en que se tienen n alternativas (con $n > 2$), pero como lo que se quiere modelar es si el cliente se fugará o no, se utilizará un Logit Binario.

Así es posible estimar la probabilidad de que el cliente i tome la decisión de fugarse como sigue:

$$P_i = \frac{\exp(V_i)}{\exp(V_i) + 1}$$

Donde V_i es la componente determinística de la utilidad de un cliente i. La utilidad se puede representar mediante la siguiente ecuación:

$$V_i = X_i * \delta_i$$

5.3. Métricas Matriz de Confusión

Las métricas que se utilizarán para analizar las matrices de confusión y así poder comparar los modelos construidos se describen a continuación:

- True Positive: Cantidad de clientes que el modelo clasifica correctamente como fugados.

- True Negative: Cantidad de clientes que el modelo clasifica correctamente como activos.

- False Positive: Falsos positivos, es decir, cantidad de clientes que el modelo clasifica como fugados y que en realidad son activos.

- False Negative: Falsos negativos, cantidad de clientes que el modelo clasifica como activos, cuando en realidad corresponden a fugados.

- Accuracy: Corresponde a la exactitud del modelo, utilizando la cantidad de fugados y activos clasificados correctamente sobre el total de clientes.

$$Accuracy = \frac{True\ positive + True\ Negative}{Total\ Population}$$

- Precision: Corresponde a la proporción de clientes clasificados como fugados correctamente sobre el total de clasificados como fugados.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- Sensitivity: Equivale a la *true positive rate*, es decir, la proporción de clientes fugados bien clasificados versus el total de fugados.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- Specificity: Equivale a la *true negative rate*, es decir, la proporción de clientes activos bien clasificados versus el total de clientes activos.

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

- Lift: Compara la precisión con la tasa de churn total en la base de prueba. Indica cuán mejor un clasificador predice comparado a una selección aleatoria, un valor sobre 1 indica que el modelo predice bien. P corresponde al número de fugados en el conjunto de datos (positivos) y N al número de no fugados (negativos).

$$Lift = \frac{Precision}{\frac{P}{(P+N)}}$$

6. ALCANCES

Esta memoria busca investigar y analizar si las llamadas realizadas en un Call Center mejoran los modelos de predicción de fuga que utilizan variables demográficas y transaccionales. Con respecto a los datos, el alcance de esta memoria se limitará a lo siguiente:

- Los modelos de predicción de fuga serán solo para clientes con servicios hogar contratados.
- Los clientes serán personas naturales, por lo que se deja de lado empresas y/o instituciones.
- Solo se trabajará con fugas por razones voluntarias, se deja de lado las bajas debido a deudas impagas.
- Solo se utilizarán datos proporcionados por la empresa, tanto de las llamadas como la información demográfica y transaccional de los clientes.

7. RESULTADOS ESPERADOS

Al finalizar el trabajo de memoria se espera obtener los siguientes resultados:

1. Variables que logren describir las llamadas registradas en un Call Center.
2. Encontrar qué tópicos presentes en las llamadas hacen más probable una fuga.
3. Probabilidad de fuga de clientes con y sin datos de las llamadas realizadas al Call Center.
4. Encontrar si la información respectiva a las llamadas de los clientes se puede utilizar como predictor de fuga y si mejora la capacidad predictiva.

8. ANÁLISIS DESCRIPTIVO DE LOS DATOS

8.1. Datos

Los clientes que se desean seleccionar son aquellos que han llamado al menos una vez al Call Center durante el mes de Julio del año 2014, y de los cuales se puede obtener dos grupos: clientes que hayan dado de baja al menos uno de los productos hogar durante Agosto 2014 y clientes que permanezcan con los mismos servicios contratados en este mes.

Para poder extraer esta muestra, se tiene la siguiente información:

- Base de datos de clientes con contrato fijo: Información de cada cliente y sus servicios hogar contratados por mes (donde los servicios pueden ser telefonía fija, internet y/o televisión). Esta información se extrae de las bases de datos proporcionadas por la empresa.

- Base de datos de llamadas: Texto de las llamadas y números de teléfono desde los cuales se llamó al Call Center por periodo. Esta información se extrae del software de Speech Analytics con que cuenta la empresa, el cual transcribe automáticamente las llamadas a texto. Sin embargo, no es posible extraer la información de forma automatizada, por lo que es necesario hacerlo manualmente. El gran problema tras lo anterior, es que la base de datos del software se guarda solamente por 3 meses móviles, lo que dificulta la extracción de datos.

- Base de datos de clientes fugados: Lista de clientes que eliminan uno o más de sus servicios hogar contratados. Esta información se extrae de las bases de datos proporcionadas por la empresa.

El proceso para extraer la muestra se ejecutó en los siguientes pasos:

1. Primero se realiza la extracción de clientes fugados:
    a) Se hizo un cruce entre la base de datos de clientes fugados del mes de Agosto 2014 con la base de datos de llamadas del mes de Julio 2014 sobre la variable que indica el número de teléfono del cliente (ANI).

b) Del cruce anterior, se obtiene un grupo de clientes que llamó al Call Center en Julio desde su teléfono fijo de la compañía y que están catalogados como fugados en Agosto del mismo año. De este grupo se saca una muestra aleatoria simple de 74 clientes.

2. Luego se realiza la extracción de clientes activos:
   a) Se hizo un cruce entre la base de datos de clientes fugados del mes de Agosto 2014 con la base de datos de llamadas de Julio sobre la variable que indica el número de teléfono del cliente (ANI), pero esta vez quedándose solo con aquellos clientes que estén en la base de datos de las llamadas y no en la base de datos de fugados. De modo de obtener a los clientes que llamaron al Call Center y que no están catalogados como fugados en Agosto.

   b) Con el resultado anterior se hizo un cruce con la base de datos de clientes con contratos fijo de Julio sobre la variable que indica el número de teléfono del cliente, para así poder identificar cuántas de estas llamadas fueron realizadas desde un teléfono fijo de la compañía. De este grupo se saca una muestra aleatoria simple de 65 clientes, quienes permanecen con los mismos servicios contratados que el mes anterior.

De esta forma, la muestra cuenta con un total de 139 personas, de las cuales 74 se fugan durante el mes de Agosto del 2014, y 65 permanecen con los mismos servicios hogar contratados durante este mes. La muestra tiene este tamaño debido a la dificultad para extraer la información, al tener que hacerlo manualmente, y contra el tiempo, es que solamente se alcanza a obtener las llamadas de 139 personas.

Al hacer un primer análisis de la muestra y limpieza de datos, se encontró 24 casos anómalos, que corresponden a clientes que no tienen contratado telefonía fija durante el mes de Julio, pero que según el cruce que se realizó, sí llamaron al Call Center durante este mes desde su teléfono fijo asociado a la compañía. Al estudiar estos 24 casos se encontró lo siguiente:

- 15 casos corresponden a clientes que efectivamente realizaron llamadas al Call Center durante Julio desde su teléfono fijo asociado a la compañía, pero que dieron de baja este producto durante el mismo mes de Julio. Por lo que al crear la base de datos de clientes de Julio

(que se hace a fin de mes) estos clientes ya aparecían sin tener contrato de telefonía fija. De este modo, la llamada que realizaron fue por motivos de una baja que ya se realizó, y no la que se quiere estudiar durante el mes de Agosto. Igualmente estos casos podrían permanecer en la muestra, pero para mantener la estructura del análisis se decide eliminarlos, ya que las llamadas realizadas por estos clientes serán por distintos motivos a la baja que se está estudiando.

- 4 casos corresponden a clientes que no tienen contratado telefonía fija, por lo que la compañía le asocia un número virtual, el cual coincide con un número real de teléfono de la competencia. Por lo que el cliente que llamó al Call Center no es el mismo que aparece en la muestra, sino que es una coincidencia de números.

- Los 5 casos restantes corresponden a errores en la base de datos, por lo que no se puede verificar si estos clientes tenían o no contratado telefonía fija durante Julio para poder realizar llamadas al Call Center desde su número.

Dado lo anterior, se eliminaron estos 24 casos ya que no se ajustan al análisis que se quiere realizar. Con esto finalmente se obtiene una muestra de 115 clientes que realizaron llamadas al Call Center en Julio desde su teléfono fijo.

Con respecto a la información demográfica y transaccional de los clientes de la muestra, la cual se extrae de la base de datos de clientes con contrato hogar, se cuenta con un total de 131 variables. Al tener tantas variables fue necesario realizar una selección previa donde se eliminó todas aquellas variables con más del 50% de datos faltantes y aquellas con cero varianza.

De este modo se eliminaron 77 variables, de las 54 restantes se analizó cuáles de estas podrían servir para el análisis y si es posible realizar algún tipo o técnica de imputación. Luego de este segundo filtro, se eliminaron 26 variables por contar con valores faltantes que no se pueden rellenar en la muestra.

Finalmente, se tienen 28 variables que logran describir al cliente, donde se aplicó técnicas de imputación en las variables *edad*, *grupo socioeconómico* y *región*. Para la edad se tenían 34 casos faltantes (29.56%), los cuales se imputaron utilizando una regresión lineal con un R^2 de 0.9167 basada en los

ruts de los clientes. Con respecto al grupo socioeconómico se tienen 13 casos faltantes, a los cuales se les creó una nueva categoría llamada *desconocido.* Finalmente para la región se cuenta con 29 casos faltantes (25.25%), para rellenarlos se utilizó los datos de la variable *área*, la cual tiene la información del número de área desde donde se realizó la llamada.

Para poder utilizar estas variables se realizaron algunas transformaciones logarítmicas a aquellas variables con valores muy altos en magnitud, como ser la cantidad de minutos hablados o el cargo total a pagar por cada cliente[3]. De esta forma se crearon nuevas variables aplicando las transformaciones que se muestran a continuación:

- Log(ARPU Voz) = $\log(ARPU\ Voz + 1)$
- Log(ARPU Banda Ancha) = $\log(ARPU\ Banda\ Ancha + 1)$
- Log(ARPU TV) = $\log(ARPU\ TV + 1)$
- Log(ARPU total) = $\log(ARPU\ total + 1)$
- Log(Cargo total) = $\log(Cargo\ total + 1)$
- Log(Minutos Salientes) = $\log(Minutos\ Salientes + 1)$
- Log(Minutos Entrantes)= $\log(Minutos\ Entrantes + 1)$
- Log(Llamadas Salientes) = $\log(Llamadas\ Salientes + 1)$
- Log(Llamadas Entrantes) = $\log(Llamadas\ Entrantes + 1)$

Para las variables categóricas, tales como región y grupo socioeconómico, se crearon nuevas variables binarias por categoría. Por lo tanto si una variable tiene *n* categorías, se crearon (*n-1)* variables binarias. Luego de esto se analizaron las variables dummy creadas con el fin de evitar aquellas con muy poca información, por lo que se agruparon en los casos en que esto ocurría. De esta forma se crearon las siguientes variables:

- MEDIAALTA = Variable binaria que agrupa las variables ABC1 y C2 definidas por la empresa.
- MEDIABAJA = Variable binaria que agrupa las variables C3 y D definidas por la empresa.
- SUR = Variable binaria que agrupa las regiones al sur del país, sin tomar en cuenta la VIII región.
- NORTE = Variable binaria que agrupa las regiones al norte del país.
- METROP = Variable binaria que indica si el cliente pertenece a la región metropolitana.

[3] El detalle y descripción de cada variable se encuentra en la sección 4.2.2 Definición de variables.

- VIII = Variable binaria que indica si el cliente pertenece a la VIII región.

Las regiones Metropolitana y VIII no se agruparon debido a que son las que cuentan con un mayor número de clientes, por lo que se consideran regiones importantes. Las demás regiones fueron agrupadas geográficamente según si están al norte o al sur de la región Metropolitana ya que por separado contaban con una cantidad muy baja de registros.

Además de estas variables, se crearon variables nuevas utilizando la información de meses anteriores. Con el fin de analizar el comportamiento con respecto a los minutos y cantidad de llamadas realizadas, se crearon nuevas variables que muestran si el cliente ha aumentado, disminuido o mantenido constante el consumo de minutos (o cantidad de llamadas) en los últimos 6 meses. Para poder realizar esto se calculó el promedio de minutos (llamadas) hablados en los últimos 6 meses (desde enero a junio) y luego se calculó la diferencia de este promedio con los minutos (llamadas) del mes de julio. De esta forma se pueden tener valores negativos y positivos, donde un valor negativo indica que el cliente disminuyó su consumo con respecto a su promedio en los últimos 6 meses. Para los casos en que los clientes no lleven 6 meses con contrato de telefonía fija el promedio se calcula con los meses que lleven activos.

Lo anterior se realizó para las variables *Minutos Salientes, Minutos Entrantes, Llamadas Salientes* y *Llamadas Entrantes*, donde también se aplicó la transformación logarítmica y se obtuvo lo siguiente:

- Log6(Minutos Salientes) =
$$\log(MinutosSalientes\_6m + |\min(Minutos\ Salientes\_6m)| + 1\ )$$
- Log6(Minutos Entrantes) =
$$\log(MinutosEntrantes\_6m + |\min(MinutosEntrantes\_6m)| + 1)$$
- Log6(Llamadas Salientes) =
$$\log(LlamadasSalientes\_6m + |\min(LlamadasSalientes\_6m)| + 1)$$
- Log6(Llamadas Entrantes) =
$$\log(LlamadasEntrantes\_6m + |\min(LlamadasEntrantes\_6m)| + 1)$$

En este caso como se tienen valores negativos, la transformación logarítmica se utiliza sumando el valor de la variable con el valor absoluto del mínimo y se suma uno para el caso en que el valor de la variable corresponda al

mínimo. De esta forma las variables estarán normalizadas y el mínimo en vez de ser negativo tendrá valor cero.

Finalmente, en la Tabla 2 se puede ver un resumen de la muestra a utilizar para el análisis. El Universo corresponde a la cantidad total de clientes que llamaron al Call Center desde su teléfono fijo de la compañía durante el mes de Julio del año 2014.

Tabla 2. Detalles muestra.

	Muestra	Universo
Fugados	63	766
Activos	52	44787
Total	115	45553

Como se puede ver, la muestra no representa las proporciones que se dan en el universo, ya que el porcentaje de fuga en el universo es de 1,68%, mientras que en la muestra es de un 54,78%. Esto ocurre porque al momento de extraer la muestra, los datos se separaron en dos grupos; fugados y activos, por lo que se extrajeron proporciones similares de ambos conjuntos. Lo anterior se hizo con el objetivo de tener una cantidad suficiente de clientes fugados que permita construir un modelo para predecir la fuga.

Para casos en que se tiene una muestra basada en el valor de la variable dependiente, según [7], [8] y [9], es necesario hacer una corrección al valor del intercepto dentro del modelo, el cual se describe a continuación:

$$\dot{\alpha} = \alpha + \ln\left(\frac{Q(1)/H(1)}{Q(0)/H(0)}\right)$$

Donde α corresponde al valor original del intercepto, $Q(1)$ es la proporción de clientes fugados en el universo, $Q(0)$ es la proporción de clientes activos en el universo, $H(1)$ es la proporción de clientes fugados en la muestra y $H(0)$ es la proporción de clientes activos en la muestra. De esta forma, los valores serían los siguientes:

$$Q(1) = \frac{766}{45553} = 0.016$$

$$Q(0) = \frac{44787}{45553} = 0.9831$$

$$H(1) = \frac{63}{115} = 0.5478$$

$$H(0) = \frac{52}{115} = 0.4521$$

Y la fórmula quedaría, para esta muestra, como se ve a continuación:

$$\dot{\alpha} = \alpha - 4.3101$$

Cabe destacar que esta corrección solo es necesario realizarla en caso de implementación del modelo, por lo que los valores cambiarían ya que se estaría implementando en una muestra distinta. De esta forma, en caso de implementación tiene que ser conocida la distribución de probabilidad del universo y de la muestra que se está utilizando. Es por esto que esta corrección no se utilizará en los análisis de los distintos modelos, ya que no es necesario ajustar el intercepto en esta etapa.

Una de las limitaciones de este estudio, es que la base de datos de las llamadas al Call Center es reducida y además solo presenta el número de teléfono desde el cual se realizó la llamada, por lo que este número es la única forma de identificar al cliente. Lo anterior genera los siguientes sesgos y trae consigo distintas consecuencias:

- El software desde el cual se extrae la información solo guarda el 30% del total de llamadas recibidas en el Call Center.
- Clientes que llamen al Call Center desde teléfonos no afiliados a la empresa no podrán ser detectados por lo tanto no serán considerados en el análisis.
- Clientes que llamen al Call Center desde su celular móvil asociado a la empresa podrán ser detectados, pero no serán incluidos en el estudio, ya que la muestra contempla solo a clientes que llamen desde su teléfono fijo.
- El motivo de las llamadas de los clientes es independiente del número del cuál llamen, es decir, puede ocurrir que un cliente llame desde su teléfono fijo y quiera hacer un reclamo sobre su contrato de internet móvil, y viceversa.

Con respecto al primer punto no hay nada que se pueda hacer, por lo que la muestra a utilizar solo consideró el 30% de las llamadas al momento de extraer los datos.

El segundo punto es un sesgo que no es posible medir, ya que son clientes que no podrán ser detectados. Por otro lado, el tercer punto sí es cuantificable ya que se tiene acceso a las bases de datos de los clientes que tienen contratos de telefonía móvil (equipo y plan). Esto genera un sesgo de selección al momento de extraer la muestra de datos, ya que los clientes que realizaron llamadas desde sus teléfonos móviles no están considerados. De todas formas una gran cantidad de las llamadas debieran ser realizadas desde teléfonos fijos, debido a que la mayoría se realizan en horarios en que la gente está en la casa y lo natural sería ocupar el teléfono fijo. En Anexo 1 y Anexo 2 se muestra el detalle de lo mencionado anteriormente, donde se puede ver que existe un sesgo al momento de sacar la muestra.

Con respecto a los clientes catalogados como fugados, de aquellos que son identificables se tiene que el 74.08% realiza las llamadas desde su teléfono fijo asociado a la empresa, por lo que el 25.91% de los casos no está siendo considerado en la extracción de la muestra (que son aquellos clientes que llaman al Call Center desde su teléfono móvil asociado a la empresa y dan de baja alguno de los servicios de hogar).

Y con respecto a los clientes catalogados como activos, es decir, que permanecen con los mismos servicios hogar contratados, el sesgo es similar al anterior. De aquellos clientes que son identificables, el 78.67% realiza las llamadas desde su teléfono fijo afiliado a la compañía, por lo que el 21.32% de los casos no está siendo considerado en la extracción de la muestra.

En síntesis, por la forma en que se obtuvo la muestra y la información disponible sobre las llamadas al Call Center es que se tiene una muestra con un cierto grado de sesgo de selección, ya que solo se cuenta con clientes que hayan realizado al menos una llamada al Call Center en Julio del año 2014 y que la hayan hecho desde su teléfono fijo asociado a la empresa, de lo contrario no podría ser detectado.

8.2. Análisis Descriptivo

8.2.1. Variables Demográficas y Transaccionales

Algunos estadísticos descriptivos que se pueden obtener con respecto a la información de los clientes presente en la base de datos de clientes con contratos hogar se muestran en la Tabla 3.

Tabla 3. Estadísticos descriptivos sobre la información de los clientes en servicios hogar.

	Fugados (N=63)		Activos (N=52)	
	Promedio	Desviación Estándar	Promedio	Desviación Estándar
Edad	47.96	14.22	51.86	14.17
Antigüedad Cliente	89.46	104.27	121.71	112.12
Antigüedad Línea	55.60	72.60	103.78	105.25
Antigüedad Banda Ancha	33.33	33.34	27.25	33.93
Antigüedad TV	16.39	23.55	13.88	25.62
ARPU Voz	191.95	1523.57	588.96	2068.69
ARPU Banda Ancha	10540.53	5667.38	11054.26	5152.34
ARPU TV	12953.60	12884.86	10373.07	11171.13
ARPU Total	32329.23	27588.65	29956.48	12873.17
Cargo Total	35046.44	30775.35	33879.71	15507.95
Minutos Salientes	1100.53	7389.53	304.30	400.54
Minutos Entrantes	98.03	118.65	105.76	88.32
Llamadas Salientes	114	392	76	87
Llamadas Entrantes	45	47	50	35
Días Mora	0.619	2.23	3.07	16.19
Reclamos	0.12	0.38	0.01	0.13
Interacciones	6.84	6.90	7.21	8.12

Referente a la antigüedad que lleva el cliente en la compañía, los clientes activos llevan en promedio un 36.04% más meses que los fugados. Por lo tanto se puede concluir que la cantidad de meses que el cliente lleva afiliado a la compañía es un factor a considerar al momento de decidir dar de baja un producto. De este modo, los clientes que deciden dar de baja algún producto llevan menos tiempo asociados a la empresa.

Analizando lo anterior pero diferenciando en tipo de producto contratado, se tiene que con respecto a los contratos de telefonía fija los clientes activos presentan una antigüedad superior a los clientes fugados, siendo esta un 86.65% mayor. Para contratos de banda ancha hogar y televisión ocurre lo

contrario, los clientes catalogados como fugados presentan una antigüedad superior a los clientes activos, siendo estas un 18.24% y 15.31% mayor, respectivamente.

Con respecto al promedio de ingresos totales que la compañía obtiene por cada usuario, el cual se mide en ARPU, se tiene que en promedio el valor comercial que otorgan los clientes fugados es un 7.33% mayor al de los clientes activos. Por lo tanto los clientes que se están fugando son valiosos para la compañía, en términos de ARPU. Esto también puede deberse a que los clientes fugados están utilizando un 72.34% más de minutos al mes que los clientes activos, y realizando un 33.33% más de llamadas.

Analizando la cantidad de reclamos comerciales registrados, los clientes catalogados como fugados a pesar de registrar más del doble que los activos, en ambos casos el promedio es cercano a cero. Los fugados están más cercanos a tener 1 reclamo registrado, mientras que los activos se acercan al cero.

Para poder analizar si existe alguna diferencia significativa entre las medias de estas variables para ambos grupos (fugados y activos) se realizaron diversos test de media y de proporciones. El test *t* de media permite decidir si dos variables aleatorias tienen la misma media y funciona analizando si una diferencia en la media muestral entre dos muestras es estadísticamente significativa. La hipótesis nula del test es que ambas medias son iguales ($\mu_A = \mu_B$), por lo tanto si se rechaza esta hipótesis es posible afirmar que las dos muestras corresponden a distribuciones de probabilidad de media poblacional distinta. El test de proporciones tiene el mismo objetivo y se utiliza para comparar variables binarias, donde la hipótesis nula corresponde a que las proporciones para ambos grupos son iguales. En la Tabla 4 y Tabla 5 se muestra el detalle.

Tabla 4. Test t de media.

Variables	p-valor
Edad	0.1495
Antigüedad Cliente	0.1161
Antigüedad Línea	0.0063
Antigüedad Banda Ancha	0.3379
Antigüedad TV	0.5885
ARPU Voz	0.2531
ARPU Banda Ancha	0.612
ARPU TV	0.2526
ARPU Total	0.5452
Cargo Total	0.793
Minutos Salientes	0.3965
Minutos Entrantes	0.6897
Llamadas Salientes	0.4539
Llamadas Entrantes	0.4809
Días Mora	0.2824
Reclamos	0.0401
Interacciones	0.7954

Tabla 5. Test de proporciones.

Variables	p-valor
Televisión	0.0058
Banda Ancha	0.0202
MEDIAALTA	1
MEDIABAJA	0.5841
SUR	0.3978
NORTE	0.7838
VIII	0.2541
METROP	0.5089

De lo anterior es posible concluir que la cantidad de meses que el cliente lleve con contrato de telefonía fija, la cantidad de reclamos comerciales realizados, y si tiene televisión y/o banda ancha contratada, son variables que presentan una diferencia estadísticamente significativa entre las medias de ambos grupos, por lo que serían interesantes de analizar y podrían ser incluidas en el modelo.

Por otro lado, también es posible analizar cuáles son los servicios de hogar más contratados dentro de la muestra. Donde de los clientes catalogados como fugados el 73.01% tiene contratado televisión y el 92.06% cuenta con banda ancha hogar. De los clientes activos, el 59% de ellos cuenta con

televisión y el 69.23% tiene contratos de banda ancha. En ambos grupos de clientes los contratos de internet son superiores a la televisión, alcanzando el 81.73% en el total de la muestra, mientras que los contratos de televisión en el total de la muestra alcanza el 62.60%.

El producto telefonía fija tiene una tasa de contrato del 100% dentro de la muestra, ya que como ésta cuenta solo con clientes que llamaron desde su teléfono fijo al Call Center, todos los clientes de la muestra tienen contrato de telefonía.

8.2.2. Variables de las Llamadas

Con respecto a la información de las llamadas (lo que se pudo extraer del software de Speech Analytics), las variables que se tienen disponibles corresponden a la duración de la llamada y a la cantidad de veces que un cliente llamó al Call Center en un periodo determinado. La media y desviación estándar de estas dos variables diferenciando entre clientes fugados y activos se muestran en la Tabla 6.

Tabla 6. Estadísticos descriptivos sobre la información de las llamadas.

	Fugados		Activos	
	Promedio	Desviación Estándar	Promedio	Desviación Estándar
Duración [min]	8.90	7.63	11.22	9.82
Número de Llamadas	1.19	0.53	1.42	0.95

Analizando la tabla anterior, se tiene que las llamadas realizadas por los clientes catalogados como activos son, en promedio, un 26.06% más largas que las de los clientes fugados. Del mismo modo, los clientes activos llaman al Call Center un 19.32% más que los fugados.

Por lo tanto se tiene que en promedio los clientes activos realizan más llamadas al Call Center que los fugados y además estas son más largas. Esto puede deberse a que si un cliente quiere dar de baja un producto, llamará al Call Center específicamente para esto, por lo que la llamada podría ser más corta.

Para analizar si la diferencia entre las medias de ambas variables para cada grupo son estadísticamente significativas se realizó un test t de media,

donde se encontraron los siguientes p-valores: 0.2179 y 0.118 para las variables Duración y Número de llamadas respectivamente. Esto quiere decir que no existe una diferencia significativa entre las medias de ambas variables para activos y fugados, por lo que es información que no aportará mucho en cuanto a la predicción de fuga de los clientes.

Otro tipo de información disponible en cuanto a las llamadas registradas en el Call Center es la categoría a la cual se asocia cada llamada. Estas categorías son creadas por distintas personas que utilizan el software, y se generan a partir de palabras presentes en las llamadas. De este modo, se definen distintos conjuntos de palabras donde cada uno está asociado a una categoría en particular, y así cada llamada tendrá asociado un set de categorías según sean las palabras presentes en la conversación.

Las categorías más frecuentes (que aparecen en más del 10% de las llamadas) diferenciando entre llamadas de clientes fugados y activos se muestran en la Tabla 7 y Tabla 8.

Tabla 7. Categorías definidas para clientes activos.

ACTIVOS		
Categorías	Frecuencia	%
A-Insatisfacción	38	0.458
Problemas Técnicos	33	0.398
P-Banda Ancha	29	0.349
P-Televisión	21	0.253
Internet	20	0.241
P-Línea Fija	20	0.241
Z 0. Reiteradas	20	0.241
Z 0. No Resueltas OK	10	0.229
Z 0. Insatisfacción	10	0.12
Z 0. Transferidas	10	0.12

Tabla 8. Categorías definidas para clientes fugados.

FUGADOS		
Categorías	Frecuencia	%
Z 0. Reiteradas	29	0.377
Z 0. No Resueltas OK	26	0.338
Z 0. Transferidas	25	0.325
Ind P - No tengo servicio	23	0.299
Ind F - Decodificador	18	0.234
Z 0. Insatisfacción	16	0.208
Ind P - Pago no reflejado	12	0.156
Ind F - Canales	10	0.13
Z 0. No Resuelta - Nuevo contacto telefónico	9	0.117
Z 0. No Res. Remite a otro numero	8	0.104
Ind F - Otros servicios adicionales	8	0.104
Z 1. Pagos	8	0.104

Como es posible apreciar de las tablas anteriores, hay categorías que no son totalmente claras y presentan prefijos no definidos, esto ya que como se mencionó anteriormente son categorías creadas por distintas personas. Para el mes de Julio, el total de categorías encontradas es de 46, por lo que solo se muestran las más relevantes, es decir, las que aparecen en más del 10% de las llamadas.

Además es importante mencionar que cada llamada puede tener un número arbitrario de etiquetas asociadas, para la muestra en estudio se pueden tener desde cero a 46 categorías.

Debido a que estas categorías son creadas por personas externas con las cuales no se tiene contacto, y que además lo hicieron con métodos desconocidos en función de palabras claves, es que se aplicará un modelo más sofisticado de detección de tópicos para así generar nuevas categorías de las cuales se tenga total conocimiento. De todos modos se construirá un modelo que utilice estas categorías para así compararlo con el modelo basado en las nuevas categorías encontradas.

9. RESULTADOS

9.1. Selección de Modelo de Fuga a partir de Variables Demográficas y Transaccionales

Las variables independientes a utilizar en este modelo corresponden a las presentes en la base de datos de clientes con contratos hogar, las cuales son tanto demográficas como transaccionales, además de las que fueron creadas y a las cuales se le aplicó una transformación logarítmica.

De este modo, se partirá construyendo un modelo base, el cual no utiliza variables de las interacciones con el Call Center.

Con este conjunto de variables, los pasos a seguir para la elección de aquellas que se incluirán en el modelo logit son realizar un Stepwise en sus tres formas: forward, backward y bidirectional. *Forward* parte probando el modelo sin ninguna variable y en cada etapa va agregando variables hasta encontrar el modelo con menor AIC. *Backward* es lo contrario, empieza con todas las variables y en cada iteración testea la eliminación de cada una, hasta encontrar la combinación con menor AIC. Finalmente *Bidirectional* es la combinación de los dos anteriores, donde en cada paso prueba agregar y eliminar variables hasta encontrar el modelo con menor AIC. Lo anterior se aplicará a los siguientes casos:

1. Utilizando las variables sin aplicar transformaciones.
2. Utilizando las variables luego de aplicar transformaciones logarítmicas.
3. Utilizando las nuevas variables creadas y sus transformaciones correspondientes.

De esta forma, se escogerá aquel modelo que cuente con un menor AIC/BIC.

El punto 1 y 2 cumple el objetivo de verificar si efectivamente la transformación logarítmica aplicada a algunas variables mejora su desempeño, y por lo tanto, el del modelo. Tras realizar varias pruebas se encontró que el usar las variables transformadas mejora el modelo, por lo tanto serán utilizadas con la transformación de ahora en adelante.

El punto 3 tiene como objetivo identificar si las variables creadas (log6(Minutos Salientes), log6(Minutos Entrantes), log6(Llamadas Salientes) y log6(Llamadas Entrantes)) ayudan a construir un modelo superior, por lo

que se buscará el mejor modelo sin usar estas variables y luego utilizándolas, para así poder comparar los modelos encontrados.

Es así como los modelos con menor AIC y BIC encontrados se muestran en la Tabla 9.

Tabla 9. Modelos seleccionados tras realizar stepwise en modelo logit.

	Modelo	Número de Variables	AIC	BIC
Sin utilizar variables creadas	1	13	137.1	172.74
	2	12	137.5	170.47
Utilizando variables creadas	3	14	132.7	171.13
	4	13	133	168.64

Para poder elegir uno de los modelos anteriores se realizaron matrices de confusión con distintas proporciones de entrenamiento, además de utilizar distintos puntos de corte, 0.5 y 0.7 para definir como fugado a un cliente.

En Anexo 3 se muestran las matrices de confusión correspondientes dentro de la muestra, donde las que tienen mejor desempeño corresponden a los modelos 2 y 4, en la proporción 80/20 (entrenamiento/testeo) y punto de corte 0.5 con un accuracy de 0.8260 para ambos.

También se realizó un análisis de los resultados de las matrices de confusión con la curva ROC, que es una representación gráfica de la medida sensitivity frente a (1- specificity). La métrica de interés es el área bajo la curva (AUC) que se puede interpretar como la probabilidad de que un clasificador puntuará una instancia positiva elegida al azar más alta que una negativa, y toma valores entre 0.5 y 1.

De esta forma, analizando los valores del AUC para distintas proporciones dentro de la muestra (Anexo 4) se tiene que el mayor AUC se alcanza en la proporción 80/20, el cual alcanza un valor de 0.9149 por el modelo 3. Los modelos 1, 2 y 4 cuentan con un AUC de 0.8832, 0.8774 y 0.9125 respectivamente para esta misma proporción.

Con lo anterior se tiene que los modelos 1, 2, 3 y 4 se desempeñan bien y de forma similar en la muestra de entrenamiento, por lo que será necesario analizarlos en el conjunto de testeo para poder elegir uno.

Utilizando la partición 80/20, las métricas de las matrices de confusión para los modelos 1, 2, 3 y 4 se muestran en la Tabla 10.

Tabla 10. Métricas matrices de confusión en muestra de testeo.

Medidas	Valores			
	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Accuracy	**0.6521**	0.6086	0.5217	0.5217
Precision	**0.6363**	0.5833	0.5	0.5
Sensitivity	**0.6363**	0.6363	0.5454	0.5454
Specificity	**0.6666**	0.5833	0.5	0.5
False Positive	**0.3333**	0.4166	0.5	0.5
False Negative	**0.3636**	**0.3636**	0.4545	0.4545

Tomando en cuenta que la muestra que se está utilizando es pequeña (N=115), dividirla en test y train puede afectar el poder estadístico para estimar los parámetros del modelo. Es por esto que además se realizó un análisis utilizando la técnica Leave one out Cross Validation [10], la cual funciona utilizando solo un registro como muestra de testeo, pero iterando con cada uno. Es decir, el método selecciona 1 registro de la muestra para utilizarla como grupo de testeo y entrena el modelo con el resto, luego hace lo mismo pero con un registro diferente hasta probar con todos los existentes.

De este modo, el modelo será testeado con todos los casos y el grupo de entrenamiento irá cambiando dependiendo del caso de prueba, por lo que el número de iteraciones corresponde al tamaño de la muestra con la que se esté trabajando. Este método entrega el error cuadrático medio del modelo, el cual se obtiene calculando la media aritmética de cada error obtenido por iteración. Por lo tanto, se obtienen los errores cuadráticos medios 0.2077, 0.2035, 0.2179 y 0.2153 para los modelos 1, 2, 3 y 4 respectivamente.

Los errores son bastante similares, pero con los resultados anteriores es posible seleccionar el modelo 1, ya que presenta un mayor accuracy y sensitivity. El área bajo la curva (AUC) para el modelo 1 en la muestra de testeo es de 0.6818. Esto significa que existe un 68.18% de probabilidad de que el modelo entregue una predicción más correcta para una persona que se fuga que para una que se queda en la compañía escogida al azar.

Este modelo no cuenta con las variables creadas, lo que demuestra que incluirlas no mejora la capacidad predictiva, por lo que no serán utilizadas. Las variables de este modelo y sus betas asociados tras realizar un logit se muestran en la Tabla 11.

Tabla 11. Variables y betas asociados al modelo 1.

Variables	Betas	Error Estándar	P-valor
Intercepto	3.2744	1.1785	0.0054
Antigüedad Línea	-0.0159	0.0058	0.0061
Antigüedad Cliente	0.0077	0.0049	0.1143
SUR	-2.3701	0.9130	0.0094
Televisión	2.6010	0.7811	0.0008
Antigüedad Banda Ancha	0.0198	0.0092	0.0323
Log(Minutos Salientes)	-2.2355	0.8026	0.0053
Log(Llamadas Salientes)	2.3691	1.0131	0.0193
Log(ARPU Voz)	-0.6093	0.4109	0.1381
Log(ARPU Banda Ancha)	-0.5046	0.2485	0.0423
Log(ARPU TV)	-0.4447	0.2012	0.0270
Días Mora	-0.0771	0.0556	0.1659
METROP	-0.9582	0.6300	0.1282

Analizando lo anterior, lo primero es notar que el beta con mayor peso corresponde al intercepto. Haciendo un análisis básico y considerando que todas las variables toman su valor promedio, la probabilidad de que un cliente dé de baja un producto es de un 77.63%.

Las variables *Antigüedad Cliente*, *Televisión*, *Antigüedad Banda Ancha* y *log(Llamadas Salientes)* tienen valores de beta positivos, lo que aumenta la probabilidad del modelo de predecir fuga. De este modo, analizando la variable *Televisión* que es significativa y con un beta de alto peso ($\beta = 2.60$) es posible concluir que para el modelo tener contratado televisión con la empresa aumenta la probabilidad de que el cliente dé de baja algún producto. Esto puede explicarse con el hecho de que el producto hogar con una mayor tasa de CHURN es la televisión.

Las variables *Antigüedad Banda Ancha* y *Antigüedad Cliente* presentan coeficientes pequeños, por lo que su efecto no es tan grande dentro del modelo. La antigüedad del cliente en la empresa no es significativa, y es la que presenta un coeficiente más pequeño por lo que su efecto es casi marginal en comparación con las otras variables. Con respecto a la cantidad de meses que el cliente tiene contratado banda ancha hogar, la relación que se establece es que mientras más meses lleve aumentará levemente la probabilidad de fuga. Para que esta variable sea distinta de cero es necesario

tener contratado banda ancha, por lo tanto el aumento en la probabilidad de fuga puede explicarse con que al tener un producto más, (cuya tasa de CHURN es de 1.18%, sobre telefonía y bajo televisión) existe una mayor probabilidad de estar desconforme con el servicio y por lo tanto dar de baja algún producto.

De forma contraria, las variables significativas con valores de beta negativos son *Antigüedad Línea*, *SUR*, *log(Minutos Salientes)*, *log(ARPU Banda Ancha)* y *log(ARPU TV)*. La variable con un valor de beta de mayor peso es *SUR* ($\beta = -2.37$), lo que estaría diciendo que clientes que sean de regiones al sur de la región Metropolitana (sin considerar la VIII región) son más propensos a mantener sus productos contratados, por lo que pertenecer a una de estas regiones reduce la probabilidad de fuga en el modelo.

Con respecto a las variables *log(ARPU Banda Ancha)* y *log(ARPU TV)* se puede concluir que mientras el valor del ARPU (tanto en banda anda como en televisión) sea mayor, los clientes presentaran una probabilidad menor de dar de baja un producto. Por lo tanto un cliente con un valor de ARPU mayor, tendrá una menor inclinación a la fuga, lo que puede explicarse con que clientes con un mayor ARPU (mayor valor para la empresa) son clientes importantes, ergo, son tratados de mejor forma por la empresa.

Las variables *log(Llamadas Salientes)* y *log(Minutos Salientes)* están altamente correlacionadas, presentan un coeficiente de correlación de 0.9046. La diferencia es que una representa cantidad y la otra duración. La variable *log(Llamadas Salientes)* es la cantidad de llamadas (salientes) que el cliente realiza desde su teléfono fijo durante el mes, mientras que *log(Minutos Salientes)* son los minutos que el cliente utiliza durante el mes, ambas con transformación logarítmica.

En el modelo los coeficientes son bastante similares salvo que tienen signos opuestos, por lo que la cantidad de llamadas estaría indicando una mayor probabilidad de fuga, mientras que los minutos utilizados lo contrario.

Como presentan una alta correlación, los efectos dentro del modelo se estarían anulando. El coeficiente de *log(Llamadas Salientes)* es un poco mayor (en términos absolutos), mientras que los valores que toma la variable son, en la mayoría, un poco más pequeños. Por lo tanto, la cantidad de llamadas y duración de estas no tendría un mayor efecto en lo que es la probabilidad de fuga.

Según lo anterior, la diferencia entre estas dos variables multiplicadas por un factor, es lo que realmente se ve en el modelo. Si la cantidad de minutos es mayor a la cantidad de llamadas, la diferencia entre ambas variables tendrá signo negativo, por lo que disminuirá la probabilidad de fuga. Lo anterior ocurre si la cantidad de minutos es mayor a las llamadas por una diferencia de α, la cual cambiará según los valores de las variables que se tengan. En la muestra que se está utilizando, α toma los valores (0,0.2) para el mínimo y máximo respectivamente.

De forma contraria, si la cantidad de llamadas supera los minutos, existe una inclinación hacia la fuga. Esto quiere decir que se están realizando llamadas cortas, por lo que el cliente estaría realizando llamadas sin utilizar muchos minutos, lo que se puede asociar con que los esté ahorrando para usar después o no quiera pasarse del plan con el que cuenta. En cambio realizar llamadas largas y utilizar una gran cantidad de minutos podría indicar que el cliente no tiene problemas con el plan, y por lo tanto, se asocia con una menor probabilidad de fuga.

9.2. Aplicación Modelo de Tópicos

Es importante destacar que los tópicos que se encontrarán serán los presentes en las llamadas en su totalidad, es decir, tanto lo que habla el cliente como lo que responde el ejecutivo. Esto ya que la información que se tiene disponible de las llamadas no diferencia entre el usuario y el ejecutivo, por lo que aparecen ambos diálogos.

Este modelo se aplicó en el software R y para poder utilizarlo es necesario primero hacer una limpieza de los datos, ya que el texto de las llamadas al ser transcrito por una máquina posee caracteres extraños, espacios extra entre las palabras, entre otras cosas.

Al tener el texto limpio, los inputs necesarios para correr el modelo son crear un diccionario de palabras y definir el número de tópicos que se quiere encontrar.

Para crear el diccionario es necesario analizar cuáles son las palabras más frecuentes dentro de la muestra de datos. Como las llamadas están en español, es necesario eliminar todas los tildes, puntuaciones y caracteres que no sean letras. También es necesario eliminar las stopwords, que

corresponden a palabras que no entregan información, como por ejemplo artículos y preposiciones.

Luego de esto es posible armar un diccionario de palabras, dejando aquellas que sean relevantes para el análisis. Esto se hizo eliminando las palabras que aparecen en menos de 5 documentos, las que aparecen solo una vez en el corpus y aquellas con las mayores frecuencias, ya que si una palabra aparece en todos los documentos no sirve para diferenciar los tópicos [11] [6] [12]. Así se obtiene el diccionario de palabras a utilizar y es posible definir N, que corresponde a 1016 palabras.

Luego es necesario definir el número de tópicos a utilizar. La elección de estos es un problema persistente en *topic modeling*, ya que en algunos casos dependerá de la formulación del problema, del investigador o de los resultados obtenidos al hacer varias pruebas con distinto número de tópicos [11]. De todos modos se probaron distintos métodos propuestos en la literatura para encontrar el número correcto.

Como primer método se utilizó el planteado en [6], el cual propone obtener el número óptimo de tópicos maximizando la verosimilitud marginal, y luego estimarla con la media harmónica. El inconveniente con este método es que el LDA al ser probabilístico las verosimilitudes varían, lo que provoca que al repetir el método descrito los resultados cambian debido a la alta varianza de la verosimilitud. De esta forma no se puede llegar a un resultado estable, por lo que no es posible obtener el número de tópicos a utilizar.

El segundo método que se utilizó fue el planteado por [13], donde se propone utilizar un modelo bayesiano que elija el número de tópicos, pero este modelo entrega como resultado el número mínimo de tópicos (2), lo que no sirve para el caso en estudio, por lo que se descarta la opción. De todas maneras se realizarán pruebas utilizando solo 2 tópicos para asegurar el descarte de la opción.

Al no poder determinar un número correcto de tópicos utilizando un método estándar, el número que se considerará correcto se escogerá dependiendo del modelo que se utilizará, se probarán modelos con distinto número de tópicos y finalmente se escogerá aquel con mejor ajuste y capacidad predictiva de fuga.

El valor de α a ingresar en el algoritmo es estándar y corresponde a $50/k$ [6] siendo k el número de tópicos que se escoja, mientras que β es estimado en el modelo y corresponde a una matriz de k tópicos por 1016 palabras.

El modelo LDA entrega como output las palabras con mayor probabilidad de pertenecer a un tópico y la distribución de probabilidad de cada uno en cada documento. De esta forma es posible etiquetar cada tópico según las palabras con mayor probabilidad de pertenecer a éste, y obtener la probabilidad de que cada tópico esté presente en un determinado documento (que en este caso serían las llamadas al Call Center).

Cada documento se forma por la combinación de cada tópico, es decir, un documento tendrá presente todos los tópicos, pero con distintas probabilidades (las cuales suman 1 en cada documento).

De este modo si el modelo es de X tópicos, se tendrán X variables "Tópico", con valores entre 0 y 1.

Tomando 10 tópicos de ejemplo, se tendrá que una llamada escogida aleatoriamente presenta las proporciones de tópicos mostrados en el Gráfico 12. Donde se puede ver que en esta llamada el tópico dominante corresponde al número 10.

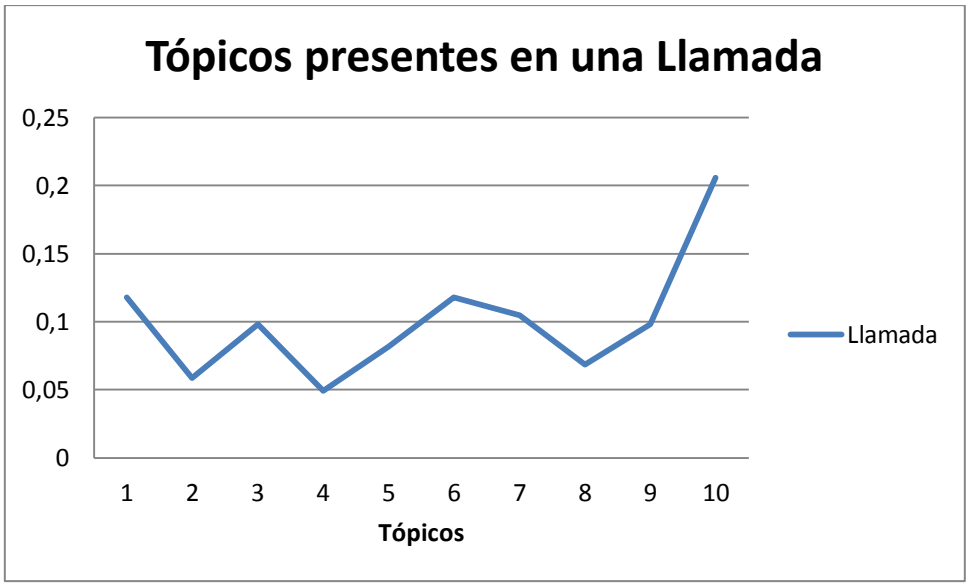


Gráfico 12. Tópicos presentes en una llamada.

Luego de tener esta información, es necesario realizar el match con los clientes para así saber qué cliente fue el que habló de X tópico de manera dominante y poder encontrar patrones que diferencien los tópicos según si un cliente da de baja o no un producto hogar de la compañía.

Para poder realizar esto, fue necesario hacer una transformación a las variables, ya que el output del LDA es por documentos (llamadas), y la información de fuga de los clientes es por persona. Por lo tanto se puede dar el caso en que una misma persona realizó múltiples llamadas, y es por esto que los datos de las llamadas tiene más registros que la de la fuga.

Para solucionar lo anterior, la transformación que se realizó consistió en sumar los valores de cada variable tópico de aquellos clientes que tengan más de una llamada, y luego normalizarlos (multiplicándolos por un factor) para que sigan sumando 1. De esta forma es posible unir ambos conjuntos de datos y así diferenciar aquellas llamadas de clientes que se hayan fugado y de los que siguen activos.

9.3. Selección de Modelo de Fuga a partir de Variables de las Llamadas

Para seleccionar el modelo que utilice variables de las llamadas primero se realizará un modelo que no incluya el contenido de estas y luego uno que sí lo incluya. Lo anterior con el objetivo de estimar y analizar la importancia de utilizar Speech Analytics y text mining en cuanto a la información de la interacción con el Call Center.

9.3.1. Variables que describen las llamadas al Call Center

Como primer análisis se utilizarán las variables de las interacciones con el Call Center sin considerar el contenido de las llamadas. Las variables disponibles que cumplen con lo anterior corresponden a la duración promedio de las llamadas de los clientes, y la cantidad de llamadas que el cliente realizó al Call Center en el mes de estudio.

Lo anterior tiene como fin comparar de manera más "limpia" el impacto que generan las variables correspondientes al contenido de las llamadas, es decir, los tópicos.

Si se utilizan solo las variables Duración y Número de llamadas, el modelo con menor AIC/BIC incluye ambas variables y presenta los valores de 159.91 y 168.14 para AIC y BIC respectivamente. Las matrices de confusión en la muestra de entrenamiento para las distintas particiones con puntos de corte 0.5 y 0.7 se muestran en Anexo 5.

Las métricas asociadas a las matrices de confusión dentro y fuera de la muestra para la partición 80/20 y punto de corte 0.5 se muestran en la Tabla 12.

Tabla 12. Métricas asociadas a Matriz de Confusión dentro y fuera de la muestra.

Medida	Valor	
	Muestra de Entrenamiento	Muestra de Testeo
Accuracy	0.6195	0.5217
Precision	0.6103	0.5
Sensitivity	0.9038	0.9090
Specificity	0.25	0.1666
False Positive	0.75	0.8333
False Negative	0.0961	0.0909

El área bajo la curva ROC (AUC) para la muestra de entrenamiento y de testeo es de 0.5584 y 0.7045 respectivamente. Mientras que el error cuadrático medio del modelo corresponde a 0.2506.

Los betas asociados a este modelo, así como también su error estándar y significancia se muestran en la Tabla 13.

Tabla 13. Indicadores modelo Logit con variables que describen las llamadas.

Variables	Betas	Error Estándar	P-valor
Intercepto	1.0563	0.4642	0.0229
Duración	-0.0287	0.0221	0.1939
Número de Llamadas	-0.4373	0.2717	0.1076

De lo anterior es posible notar que la duración y número de llamadas no son significativas en el modelo. Ambas cuentan con betas con signo negativo por lo que disminuyen la probabilidad de fuga en el modelo. Esto quiere decir que una mayor cantidad de llamadas al Call Center y una mayor duración promedio de éstas no está asociado con la baja de algún producto, es más, disminuye la probabilidad de fuga en el modelo.

De todos modos, el efecto que generan ambas variables en el modelo es pequeño en comparación con el intercepto, el cual al tener un coeficiente positivo aumenta la probabilidad de fuga del modelo. Dado lo anterior es posible concluir que las variables Duración y Número de llamadas no son buenos predictores de fuga al utilizarlas por sí solas.

9.3.2. Variables que describen el contenido de las llamadas al Call Center

Las variables que describen el contenido de las llamadas al Call Center son las encontradas tras utilizar el modelo LDA, que corresponden a las variables *Tópico*. Sin embargo, como se mencionó en la sección 8.2.2, el software desde el cual se obtuvo la información dispone de categorías pre definidas asociadas a las llamadas. De este modo, con el único objetivo de comparar las categorías existentes con los tópicos creados, es que se construirá un modelo logit que utilice exclusivamente estas categorías.

Es importante mencionar que las categorías existentes solamente serán utilizadas con fines informativos y de comparación, ya que como se mencionó, no se tiene información suficiente sobre cómo se crearon y cuál es su significado. Por esta razón se hace necesario crear nuevas categorías utilizando un modelo de detección de tópicos.

9.3.2.1. Categorías Existentes

Para utilizar las categorías se crearon variables binarias que indican si la llamada contiene o no cierta categoría. Cada llamada puede contener más de una etiqueta, sin embargo también hay casos en que no tiene asociada ninguna.

Como la muestra que se está utilizando es pequeña y contiene solamente 115 registros, hay categorías con una frecuencia de aparición muy baja. Es por esto que para la construcción del modelo solo se tomarán en cuenta las más frecuentes, es decir, que aparezcan en más del 15% de las llamadas de los clientes. Las categorías que cumplen lo anterior son las siguientes:

- A-Insatisfacción
- Ind F – Decodificador
- Ind P – No tengo servicio

- Internet
- P – Banda Ancha
- Problemas Técnicos
- P – Televisión
- Z 0. Insatisfacción
- Z 0. No Resueltas OK
- Z 0. Reiteradas
- Z 0. Transferidas

El modelo seleccionado cuenta con un AIC de 89.69 y BIC de 106.16. Las métricas de la matriz de confusión para la partición 80/20 con punto de corte 0.5 y las variables seleccionadas con sus respectivos coeficientes se muestran en la Tabla 14 y Tabla 15.

Tabla 14. Métricas matriz de confusión modelo categorías existentes.

Medida	Valor	
	Muestra de Entrenamiento	Muestra de Testeo
Accuracy	0.8695	0.7391
Precision	0.8225	0.6470
Sensitivity	0.9807	1
Specificity	0.725	0.5
False Positive	0.275	0.5
False Negative	0.0192	0
AUC	0.9373	0.7917

Tabla 15. Variables y coeficientes modelo categorías existentes.

Variables	Betas	Error Estándar	P-valor
Intercepto	0.3976	0.3341	0.2341
Ind F – Decodificador	1.6921	1.0999	0.1239
P – Banda Ancha	-3.5822	1.1782	0.0023
Problemas Técnicos	-4.3749	1.3901	0.0016
Z 0. Reiteradas	1.7266	0.8779	0.0492
Z 0. Transferidas	2.2409	0.9806	0.0223

El problema existente al utilizar categorías pre definidas como variables es que no es posible otorgar una interpretación acertada, ya que solo se cuenta con la etiqueta, no con su significado. De lo anterior se podría extraer que llamadas que presenten las categorías *Ind F – Decodificador, Z 0. Reiteradas y Z 0. Transferidas* aumentaría la probabilidad de fuga de un cliente, donde solo las últimas dos son significativas.

9.3.2.2. Tópicos Encontrados

El modelo a construir tendrá como variables independientes los tópicos encontrados de las llamadas y como dependiente la Fuga.

Para poder encontrar la combinación de variables que forme un mejor modelo (en términos de ajuste y capacidad predictiva) se utilizaron 3 métodos:

- Se probaron todas las combinaciones posibles entre las variables Tópicos para el modelo logit, y luego se escogió aquella combinación que entregara un menor AIC y BIC.

- Se probaron todas las combinaciones posibles entre las variables Tópicos para el modelo logit, pero esta vez incluyendo interacciones entre variables, es decir, del tipo *Tópico1*Tópico2*. De la misma forma anterior, se elige el modelo con menor AIC y BIC.

- Se utiliza el método Stepwise en sus tres formas; *forward*, *backward* y *bidirectional*, el cual busca automáticamente el modelo con mejor ajuste.

Lo anterior se hizo para distintos casos de LDA, utilizando desde 2 a 12 tópicos, ya que como se mencionó anteriormente, para elegir la cantidad correcta de tópicos se probará con distintos números y se elegirá aquel que entregue el mejor modelo.

Haciendo la diferencia entre modelos con y sin interacciones entre variables, los modelos escogidos se pueden ver en la Tabla 16 y Tabla 17, respectivamente.

Tabla 16: Modelos elegidos sin interacciones entre variables.

Modelo	Número de Tópicos	Número de Variables	AIC	BIC
1	4	2	151.6	157.09
2	8	2	151.5	156.99

Tabla 17: Modelos elegidos con interacciones entre variables.

Modelo	Número de Tópicos	Número de Variables	AIC	BIC
3	8	18	146.3	195.72
4	6	4	150.1	161.03
5	8	15	145.1	186.29

Donde el Número de Tópicos representa el input del modelo LDA, es decir, el modelo de 4 tópicos de la Tabla 16 indica que solo se tienen las variables Topico1, Topico2, Topico3 y Topico4, porque el modelo de tópicos recibe como input encontrar solo 4 tópicos en la muestra.

El criterio para elegir los modelos fue escoger aquel con menor AIC y menor BIC, para el caso sin interacciones el modelo 1 con 4 tópicos cumplía ambas condiciones. Para el caso con interacciones, el modelo con menor AIC fue el modelo 3, con 8 tópicos y el con menor BIC fue el modelo 4, de 6 tópicos.

Los modelos 2 y 5 fueron escogidos de forma separada ya que ninguno cuenta con la constante en la ecuación. De este modo, son los modelos elegidos tras realizar los tres pasos mencionados anteriormente pero para modelos sin el intercepto.

Para poder elegir uno de los modelos anteriores se realizaron matrices de confusión con distintas proporciones de entrenamiento, además de utilizar distintos puntos de corte, 0.5 y 0.7 para definir como fugado a un cliente.

En Anexo 6 se muestran las matrices de confusión dentro de la muestra con sus distintas proporciones y con ambos puntos de corte. Las matrices con mejor desempeño en todas las proporciones corresponden a los modelos 3 y 5, ambos con 8 tópicos.

En la proporción 80/20 con punto de corte 0.5 se alcanza el mayor accuracy, el cual corresponde al modelo 3. Las métricas correspondientes se pueden ver en la Tabla 18 para los modelos 3 y 5.

Tabla 18. Métricas de los modelos 3 y 5 en el conjunto de entrenamiento.

Medida	Valor	
	Modelo 3	Modelo 5
Accuracy	**0.7826**	0.7065
Precision	**0.7857**	0.7192
Sensitivity	**0.8461**	0.7884
Specificity	**0.7**	0.6
False Positive	**0.3**	0.4
False Negative	**0.1538**	0.2125

Los valores del área bajo la curva (AUC) para cada proporción se pueden ver en Anexo 7.

En este caso se observa que ambos modelos con 8 tópicos son los que entregan mejores resultados, el modelo con mayor AUC es el modelo 3 en la proporción 90/10, que da un área bajo la curva de 0.86. Esto significa que existe un 86% de probabilidad de que el modelo entregue una predicción más correcta para una persona que se fuga que para una que se queda en la compañía escogida al azar.

Finalmente se utilizó la técnica de Cross Validation usando el método Leave one out (el cual se explicó en la sección 9.1), ya que como se está utilizando una muestra pequeña, dividirla en *test* y *training* genera que se tenga muy poca información para entrenar el modelo. En la Tabla 19 se observan los errores obtenidos para cada modelo.

Tabla 19. Error cuadrático medio de cada modelo tras realizar validación cruzada.

Modelo	MSE (Mean Square Error)
1	0.2339
2	0.2335
3	0.2378
4	0.2332
5	0.2377

Los errores para cada modelo son bastante similares, por lo que no es posible seleccionar un modelo por sobre otro basándose solo en el error cuadrático medio.

Según los análisis realizados, que consideran tanto el ajuste como la capacidad predictiva del modelo, es posible establecer que un modelo utilizando 8 tópicos entrega los mejores resultados. La diferencia existente

entre los modelos 3 y 5 es bastante pequeña: el modelo 5 tiene tres variables menos que el 3 y el resto permanece igual, por lo que ambos podrían servir para la predicción de fuga con buenos resultados.

Para poder seleccionar uno de los modelos anteriores, se analizaron las matrices de confusión y curvas ROC en la muestra de testeo utilizando la partición 80/20. En la Tabla 20 se muestran los resultados obtenidos.

Tabla 20. Métricas matriz de confusión para modelos 3 y 5 en muestrea de testeo.

Medida	Valor	
	Modelo 3	Modelo 5
Accuracy	0.5652	**0.6956**
Precision	0.5384	**0.6428**
Sensitivity	0.6363	**0.8181**
Specificity	0.5	**0.5833**
False Positive	0.5	**0.4166**
False Negative	0.3636	**0.1818**

El análisis de la curva ROC para ambos modelos en la data de testeo (80/20) muestra un área bajo la curva similar, siendo el valor de AUC para el modelo 5 un poco mayor, alcanzando un 0.7576.

Con lo anterior es posible seleccionar el modelo 5, ya que presenta un mayor accuracy y una mayor área bajo la curva (AUC) en la muestrea de testeo. Además de que es una versión simplificada del modelo 3, por lo que por el principio de parsimonia se escoge aquel modelo que tenga una menor cantidad de variables, ya que ambos se ajustan de manera similar.

Los parámetros utilizados en el algoritmo del LDA para $k = 8$ tópicos corresponden a un valor de $\alpha = \frac{50}{8} = 6.25$, un diccionario de $N = 1016$ palabras, y la matriz $\beta = 1016 \times 8$ con valores estimados en el modelo.

Las variables del modelo 5 con 8 tópicos y los betas asociados se muestran en la Tabla 21.

Tabla 21. Variables modelos con 8 tópicos.

Variables	Betas	Error Estándar	P-valor
Tópico 1	-28.10	19.35	0.1463
Tópico 2	30.97	19.61	0.1143
Tópico 3	-36.40	15.06	0.0156
Tópico 4	-30.77	22.05	0.1629
Tópico 5	-45.57	22.96	0.0471
Tópico 7	44.49	18.07	0.0138
Tópico 8	33.71	23.73	0.1554
Tópico1*Tópico2	-186.06	107.61	0.0838
Tópico1*Tópico5	399.95	136.32	0.0033
Tópico2*Tópico4	200.32	94.71	0.0344
Tópico2*Tópico8	-239.30	137.24	0.0812
Tópico3*Tópico4	414.81	150.26	0.0005
Tópico4*Tópico5	-441.40	168.52	0.0088
Tópico5*Tópico8	456.97	177.15	0.0098
Tópico7*Tópico8	-343.22	144.11	0.0172

Luego de elegir el modelo final y la cantidad de tópicos a utilizar, es posible darle sentido al modelo tomando en cuenta el significado de los tópicos dados los resultados del modelo LDA.

Este modelo entrega las palabras más recurrentes por cada tópico, y a cada tópico le asigna una probabilidad de aparición en un documento. De este modo, cada documento tiene asignada distintas probabilidades de tener presente un tópico, sin embargo nunca tendrá probabilidad cero en uno de ellos.

Utilizando esta información, es posible etiquetar cada tópico con un nombre según sean las palabras que aparecen con mayor frecuencia (en Anexo 8 se muestran las primeras 10 palabras más frecuentes por cada tópico), de la misma forma las probabilidades de cada tópico se pueden utilizar para encontrar el tema dominante en cada llamada, que corresponde al tópico que presente una mayor probabilidad de aparición.

Para poder ver gráficamente las distribuciones de probabilidad de los tópicos diferenciando entre clientes fugados y activos, se calculó el promedio de las probabilidades de aparición utilizando todas las llamadas realizadas por clientes fugados y clientes activos. De esta forma, para cada tópico se tendrá una probabilidad de aparición correspondiente al promedio de todas las llamadas del grupo respectivo.

En el Gráfico 13 se puede ver en promedio cuáles son los tópicos más recurrentes o con mayor probabilidad de aparición para cada grupo de clientes, ordenado según la diferencia de probabilidad entre grupos. Así es posible observar que el Tópico 2 es el que presenta la menor brecha entre grupos, mientras que el Tópico 8 cuenta con la mayor, siendo el tópico más predominante en los clientes fugados.

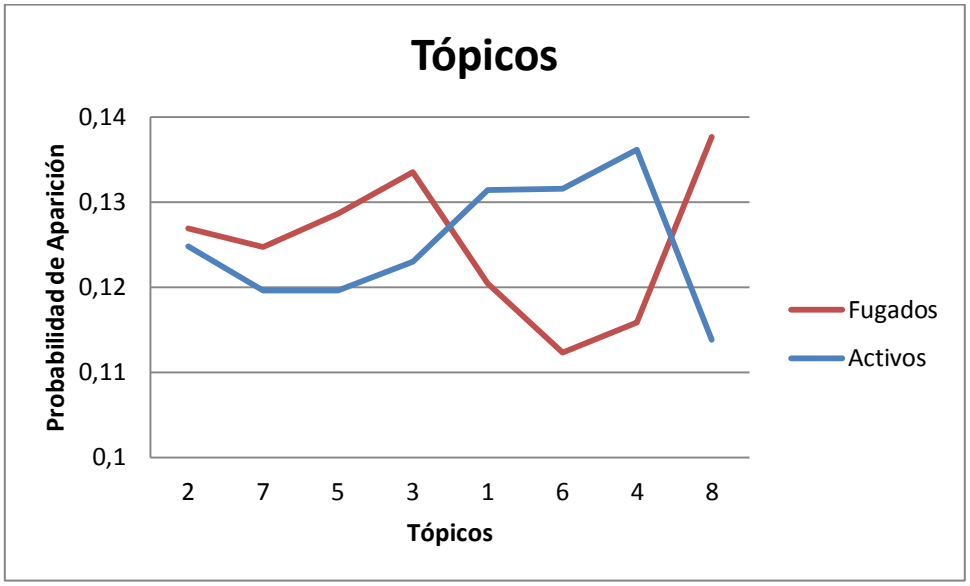


Gráfico 13. Distribución de probabilidad para cada tópico para clientes fugados y activos.
Fuente: Elaboración propia.

De esta forma, los tópicos 2, 3, 5, 7 y 8 son los que están más presentes en las llamadas realizadas por los clientes fugados, mientras que los tópicos 1, 4 y 6 tienen una presencia mayor en las llamadas realizadas por los clientes que permanecen activos. Siendo los tópicos 8 y 4 los más predominantes para cada grupo, fugados y activos respectivamente.

Analizando solamente los tópicos dominantes dentro de cada llamada, que como ya se explicó, corresponden a los tópicos que tengan una mayor probabilidad de aparición dentro de un documento, es posible crear la Tabla 22 y Tabla 23 donde se muestran los distintos tópicos para clientes fugados y activos, ordenados por frecuencia de aparición en las llamadas.

Además es posible asignar una etiqueta a cada tópico para saber de qué se trata. Para etiquetar los tópicos se utilizó las primeras 30 palabras más

frecuentes, se leyeron algunas llamadas con el fin de corroborar la etiqueta y se entrevistó a expertos para su validación.

Tabla 22. Resultados LDA con 8 Tópicos para clientes fugados.
Fuente: Elaboración propia.

FUGADOS			
Tópicos	Nombre	Frecuencia	%
8	Transferir	18	21.69%
5	Contactabilidad	16	19.28%
2	Problema Internet	14	16.87%
7	Cambio/Intento de baja	13	15.66%
6	Problema TV	11	13.25%
3	Reclamo	10	12.05%
1	Bloqueo/Desbloqueo	8	9.64%
4	Página Web	4	4.82%

Tabla 23. Resultados LDA con 8 Tópicos para clientes activos.
Fuente: Elaboración propia.

ACTIVOS			
Tópicos	Nombre	Frecuencia	%
4	Página Web	13	16.88%
6	Problema TV	13	16.88%
1	Bloqueo/Desbloqueo	11	14.29%
8	Transferir	11	14.29%
3	Reclamo	10	12.99%
5	Contactabilidad	9	11.69%
7	Cambio/Intento de baja	9	11.69%
2	Problema Internet	8	10.39%

Con las etiquetas ya definidas se logran entender los distintos tópicos y clasificar las llamadas. La explicación de cada etiqueta se describe a continuación:

- **Tópico 1 "Bloqueo/Desbloqueo":** Este tema se relaciona con el bloqueo o desbloqueo de algún producto que tenga el cliente. Entre sus palabras más probables es posible encontrar *Bloqueo, Pin, Puk y Sim.* Los códigos PIN y PUK se utilizan para casos de emergencia en situaciones de bloqueo. Al tener las palabras SIM, PIN y PUK pareciera que es un tópico perteneciente a productos Móvil, ya que los celulares son los que presentan estos códigos y tarjeta SIM.

- **Tópico 2 "Problema Internet":** Hace referencia directa a que el cliente tiene algún tipo de problema con su conexión a internet. Entre las palabras más probables se encuentran *Internet, Módem, Conexión, Red y Computador.*
- **Tópico 3 "Reclamo":** Este tema es bastante general, ya que se asocia con cualquier tipo de reclamo que el cliente quiera realizar. Por lo tanto una llamada que presente este tópico con alta probabilidad estaría indicando que el cliente está haciendo algún tipo de reclamo.
- **Tópico 4 "Página Web":** Este tema está relacionado con la página web de la empresa, ya sean dudas, problemas al ingresar o canjeo de puntos. Entre las palabras más probables se encuentran *Página, Configuración, Ingresar y Contraseña.*
- **Tópico 5 "Contactabilidad":** Este tema está relacionado con el interés del cliente de contactarse con alguien, ya sea directamente con el Call Center o para realizar llamadas particulares.
- **Tópico 6 "Problema TV":** Hace referencia directa a que el cliente tiene algún tipo de problema con la configuración de su televisión. Entre las palabras más probables se encuentran *Señal, Pantalla, Televisión y Decodificador.*
- **Tópico 7 "Cambios/Intento de baja":** Está relacionado con que el cliente desea cambiar algún producto o tratar de dar de baja uno de ellos. Como la mayoría de las empresas de telecomunicaciones, los productos hogar se venden generalmente en packs o promociones, por lo que si un cliente quiere cambiar su plan o contrato puede estar asociado a la baja de alguno de los productos.
- **Tópico 8 "Transferencia":** Este tema está relacionado con las transferencias de líneas que ocurren durante una llamada. Cuando el cliente llama por algún tema en específico que el ejecutivo no es capaz de atender, éste debe ser transferido a un área determinada donde se puedan resolver sus inquietudes. Es el tópico con mayor presencia en los clientes fugados, lo que se puede deber a que cuando un cliente desea dar de baja un producto, inmediatamente es transferido a un ejecutivo especialista en bajas para evitar que éste termine el contrato.

De esta forma es posible darle interpretación a las variables presentes en el modelo. Lo primero que se puede observar es que las variables Tópico 2, Tópico 7 y Tópico 8 son positivas, lo que significa que si una llamada presenta alguno de estos tópicos de forma dominante es más probable que la predicción del modelo se incline hacia la fuga. De forma contraria ocurre

con las variables Tópico 1, Tópico 3, Tópico 4 y Tópico 5, que son negativas en el modelo.

Los signos de los tópicos en el modelo concuerdan con los análisis de los resultados del LDA, exceptuando por el Tópico 3 y Tópico 5, que según los análisis corresponden a tópicos más presentes en los fugados pero la regresión está indicando que disminuye la probabilidad de fuga.

Lo anterior se puede explicar recordando que cada llamada presenta todos los tópicos en distintas probabilidades, por lo tanto el Tópico 5 o el Tópico 3 podrían presentar una alta probabilidad de aparición en las llamadas de los fugados, pero dependerá de la combinación con los demás tópicos. Es decir, una llamada puede tener 2 o 3 tópicos con mayor probabilidad de aparición, por lo que dependerá de esta combinación de lo que se trate efectivamente la llamada.

Viendo la etiqueta del Tópico 5, *Contactabilidad*, es un tema que puede variar dependiendo con qué otro tópico ocurra, ya que es bastante general. Lo mismo ocurre con el Tópico 3, que está etiquetado como *Reclamo.*

Analizando las variables positivas, la que tiene más peso es Tópico 7 (β = 44.49), que aparece como tópico dominante en un 15.66% de las llamadas que realizan los clientes catalogados como fugados, mientras que aparece en una menor proporción en las llamadas de los clientes que permanecen con los mismos productos.

La variable Tópico 2 también es interesante de analizar, ya que es el tercer tópico más frecuente dentro de los fugados, y el último en los clientes activos, de modo que logra diferenciar notoriamente ambos grupos. Por lo que llamadas por problemas con los servicios de internet en el hogar podría ser propio de clientes fugados.

Finalmente, la variable Tópico 8 (β = 33.71) es la más frecuente dentro de las llamadas de los fugados, por lo que es lógico que tome un valor positivo. Este tópico está etiquetado como Transferir, por lo que son llamadas con un objetivo particular que el cliente realiza porque tiene un problema en específico.

Con respecto a las variables con betas negativos, específicamente el Tópico 4 es el más frecuente entre los activos y el menos frecuente entre los clientes fugados. Por lo que es una variable que logra diferenciar ambos grupos, y

está asociada a que los clientes mantendrán sus productos afiliados a la compañía.

Las demás variables con betas negativos corresponden a tópicos que tienen que ver con el bloqueo/desbloqueo de productos (Tópico 1), contactabilidad (Tópico 5) y reclamos en general (Tópico 3).

Con todo el análisis anterior se puede concluir que llamar al Call Center para solicitar cambios de productos y/o problemas con el Internet (lo que puede ocurrir a través de una transferencia de línea) implica una alta probabilidad de dejar la compañía, mientras que llamar por dudas o consultas sobre la página web de la empresa no presenta mayores indicios de que el cliente quiera terminar el contrato. Esto es lógico ya que actualmente los servicios de internet se están usando constantemente y son una necesidad en el día a día, de modo de que si la empresa no está entregando un buen servicio, el cliente se puede sentir incentivado a cambiar a una compañía que le entregue lo que necesita. Y con respecto a los cambios, si un cliente llama porque quiere cambiar algún producto/servicio, implica que no está satisfecho con él. En cambio problemas o dudas al ingresar a la página web de la empresa son temas que, al parecer, no son tan relevantes para el cliente promedio.

Los valores de los betas para cada variable son de gran magnitud (especialmente para las interacciones entre variables), esto se debe a que los valores de las distintas variables de tópicos se encuentran entre 0 y 1, ya que la suma de los 8 tópicos es 1. Por lo tanto al multiplicar las variables, los valores se hacen aún más chicos, lo que se ve reflejado en los valores de los betas.

9.3.2.3. Comparación

Tras construir ambos modelos es posible realizar una comparación con fines informativos en cuanto al ajuste y capacidad predictiva. En la Tabla 24, se muestran las métricas correspondientes a las matrices de confusión de ambos modelos en la muestra de testeo.

Tabla 24. Métricas para modelos que utilizan categorías y tópicos.

	Valor	
Medidas	Categorías	Tópicos
Accuracy	**0.7391**	0.6956
Precision	**0.6470**	0.6428
Sensitivity	**1**	0.8181
Specificity	0.5	**0.5833**
False Positive	0.5	**0.4166**
False Negative	**0**	0.1818

Lo primero es notar que la exactitud de ambos modelos es bastante similar, lo que es esperable ya que ambos modelos utilizan variables que describen el contenido de las llamadas. Sin embargo, es necesario analizar más que la medida de accuracy para definir qué modelo presenta una mejor capacidad predictiva.

Ambos modelos presentan una precisión del 64%, lo que quiere decir que del total de clientes que el modelo clasifica como fugados, lo hace correctamente el 64% de las veces. Analizando la sensibilidad, el modelo que utiliza categorías es superior, del total de clientes fugados el modelo logra clasificar bien al 100%. Mientras que el modelo de tópicos logra clasificar bien al 81%.

Sin embargo, analizando el error al clasificar como fugado a un cliente que no lo es (false positive), se tiene que el modelo que utiliza categorías presenta una mayor error, el cual alcanza el 50%.

Lo anterior quiere decir que a pesar de que el modelo de categorías clasifique correctamente como fugado al 100% del total de fugados, también se equivoca el 50% de las veces. Esto significa que es un modelo que clasifica a la mayoría de los clientes como fugados, específicamente al 50.58%, lo que trae consigo costos innecesarios en campañas de retención.

En síntesis, ambos modelos son similares, siendo el modelo de tópicos levemente superior, ya que a pesar de tener una menor sensibilidad, también cuenta con un menor error.

9.3.3. Variables que describen las llamadas versus variables que describen su contenido

Finalmente es posible concluir que utilizar variables que describan y clasifiquen el contenido de las llamadas a través de tópicos o categorías, aumenta la capacidad predictiva de los modelos de predicción de fuga. Por lo tanto es valioso tener información sobre lo que tratan las llamadas que los clientes realizan y no solo saber la cantidad de veces que llamó o la duración promedio de estas. Es así como el modelo seleccionado que utiliza variables de las llamadas corresponde al de 8 tópicos con una exactitud del 69%.

En conclusión, un modelo que utilice variables del contenido de las llamadas aumenta su exactitud en un 17.38% con respecto a un modelo que no utilice esta información. De esta forma, el aumento en la exactitud del modelo representa el valor que otorga realizar *Speech Analytics* y aplicar métodos que permiten categorizar las llamadas.

9.4. Selección de Modelo de Fuga utilizando Ambos Tipos de Variables

Las variables a utilizar en este modelo corresponden a variables tanto demográficas y transaccionales como variables respectivas a las llamadas al Call Center. Se cuenta con dos tipos de variables de esta última categoría:

- Variables asociadas al contenido de las llamadas registradas en el Call Center: Estas corresponden a las variables de tópicos que se obtienen luego de aplicar técnicas de text mining sobre las llamadas.

- Variables asociadas a las interacciones con el Call Center: Estas variables tienen que ver con la cantidad de llamadas que realizan los clientes y la duración promedio de éstas.

De esta forma, utilizando todo este conjunto de variables se construirá un modelo logit para explicar la probabilidad de fuga de los clientes. Para realizar lo anterior el análisis se separará en las siguientes etapas:

1. Se aplicará técnicas de stepwise (backward, forward y bidirectional) utilizando solo las variables demográficas, transaccionales y de tópicos. Donde se elegirá aquel modelo con mínimo AIC y/o BIC.

2. Se aplicará técnicas de stepwise (backward, forward y bidirectional) utilizando las variables demográficas, transaccionales y las variables asociadas a la interacción con el Call Center (Número de llamadas y Duración). Donde se elegirá aquel modelo con mínimo AIC y/o BIC.

3. Finalmente se aplicarán las mismas técnicas de stepwise pero a todo el conjunto de variables, encontrando también los mejores modelos que cuenten con el mínimo AIC y BIC.

El objetivo de lo anterior es lograr ver el efecto que tienen los tópicos en el modelo, y analizar si efectivamente agrega un valor adicional con respecto a las variables que ya se tienen disponibles. Los modelos seleccionados en cada etapa anterior se muestran en la Tabla 25.

Tabla 25. Modelos elegidos por etapa luego de realizar stepwise.

Etapa	Modelo	Número de Variables	AIC	BIC	Criterio de selección
1	1	16	126	172.64	Min AIC/BIC
2	2	13	131.9	170.36	Min AIC/BIC
3	3	16	126.6	173.27	Min AIC/BIC

Los modelos 1 y 3 son bastante parecidos, cuentan con el mismo número de variables y los valores de AIC y BIC son cercanos. Sin embargo, el modelo 1 tiene valores de AIC y BIC menores, por lo que es posible seleccionar los modelos 1 y 2 para un análisis más profundo. Las matrices de confusión de ambos modelos en el conjunto de entrenamiento para distintas particiones y puntos de cortes se encuentran en Anexo 9.

La matriz que alcanza el mejor desempeño es con la partición 80/20 y punto de corte 0.5, donde el modelo 1 presenta un accuracy de 0.8804 y el modelo 2 de 0.8152.

Las curvas ROC y sus respectivos valores del área bajo la curva (AUC) se muestran en Anexo 10 para las distintas particiones. El mayor AUC se alcanza en la partición 90/10 por el modelo 1, el cual corresponde a 0.9451. El modelo 2 presenta un AUC de 0.897.

Según lo anterior el modelo 1 sería el que cuenta con un mejor desempeño en lo que respecta al conjunto de entrenamiento, pero también es importante analizar la muestra de testeo, ya que también es necesario que el modelo elegido tenga un buen desempeño en este conjunto, de lo contrario podría tratarse de un modelo sobreajustado.

Las métricas de las matrices de confusión en la proporción 80/20 para ambos modelos tanto en el conjunto de entrenamiento como de testeo se muestran en la Tabla 26 y Tabla 27.

Tabla 26. Métricas matriz de confusión modelo 1.

Modelo 1	Valor	
Medidas	Muestra de entrenamiento	Muestra de testeo
Accuracy	0.8804	0.7391
Precision	0.9019	0.8571
Sensitivity	0.8846	0.5454
Specificity	0.875	0.9166
False Positive	0.125	0.0833
False Negative	0.1153	0.4545

Tabla 27. Métricas matriz de confusión modelo 2.

Modelo 2	Valor	
Medidas	Muestra de entrenamiento	Muestra de testeo
Accuracy	0.8152	0.5652
Precision	0.8181	0.5555
Sensitivity	0.8653	0.4545
Specificity	0.75	0.6666
False Positive	0.25	0.3333
False Negative	0.1346	0.5454

Los valores del área bajo la curva en el conjunto de testeo para el modelo 1 y 2 son 0.7121 y 0.6364 respectivamente. Dado lo anterior, el modelo 1 presenta un desempeño superior en la muestra de testeo, al igual que en el conjunto de entrenamiento.

El último análisis a realizar corresponde a Leave one out Cross Validation, que entrega un error cuadrático medio de 0.2034 y 0.2162 para los modelos 1 y 2 respectivamente. Los errores son bastante similares entre ellos, siendo un poco menor el correspondiente el modelo 1.

De esta forma, es posible seleccionar al modelo 1 como aquel modelo con un mejor ajuste y capacidad predictiva utilizando variables demográficas,

transaccionales y las obtenidas mediante técnicas de text mining aplicadas a las llamadas registradas en el Call Center.

El análisis realizado anteriormente no cuenta con interacciones entre las variables de tópicos (como se realizó en la sección 9.3.2), por lo que también sería interesante incluir interacciones y ver si el modelo es capaz de mejorar.

Para poder agregar las interacciones entre los tópicos es necesario realizar una reducción de la dimensión de la data, ya que las interacciones equivalen a 28 variables extras que incluir en el modelo.

Para esto se aplicará Análisis de Componentes Principales para las variables de tópicos y sus interacciones, que corresponden a 36 variables.

Tras realizar lo anterior se obtienen 6 componentes principales que logran explicar el 92,99% de la varianza. Aplicando técnicas de rotación como Varimax, se obtienen los coeficientes de los componentes, los que se utilizarán en el modelo.

En Anexo 11 se puede ver gráficamente la naturaleza de cada componente principal, donde es posible notar que en los 6 componentes principales las interacciones entre los tópicos no tienen mucho protagonismo. Los coeficientes de las interacciones son cercanos a cero en todos los componentes, por lo tanto es muy similar a ocupar solo las variables de tópicos, sin sus interacciones. Dado esto es muy probable que el modelo con interacciones no sea superior al anterior, de todas formas se aplicará la misma metodología (stepwise), con el fin de encontrar el modelo con menor AIC y BIC (el que se muestra en la Tabla 28).

Tabla 28. Modelo elegido tras realizar stepwise en regresión logística con interacciones entre tópicos.

Modelo	Número de Variables	AIC	BIC	Criterio de selección
4	16	125.6	169.48	Min AIC/BIC

Las matrices de confusión en el conjunto de entrenamiento para cada partición se pueden ver en Anexo 12, el mayor accuracy se alcanza en la partición 80/20 con punto de corte 0.5, que corresponde a 0.8695.

Analizando el valor de AUC de la curva ROC se tiene 0.899, 0.9127 y 0.8654, para las particiones 80/20, 90/10 y 95/5 respectivamente.

Para poder hacer la comparación con el modelo anterior, en la Tabla 29 se muestran las métricas de la matriz de confusión en la partición 80/20 en el conjunto de entrenamiento y de testeo.

Tabla 29. Métricas matriz de confusión en partición 80/20 modelo 4.

Medidas	Valor	
	Muestra de entrenamiento	Muestra de testeo
Accuracy	0.8695	0.6521
Precision	0.8846	0.7142
Sensitivity	0.8846	0.4545
Specificity	0.85	0.8333
False Positive	0.15	0.1666
False Negative	0.1153	0.5454

El área bajo la curva ROC de este modelo en el conjunto de testeo corresponde a 0.697, mientras que el error cuadrático medio calculado por Cross Validation es de 0.2176.

De esta forma, el modelo 1 sigue siendo superior al modelo 4, por lo que las interacciones entre tópicos no serán incluidas en la regresión. Los betas asociados a esta regresión se muestran en la Tabla 30.

Tabla 30. Betas asociados a logit modelo 1.

Variables	Betas	Error Estándar	P-valor
Intercepto	1.1768	2.4967	0.6373
Antigüedad Línea	-0.0119	0.0039	0.0027
SUR	-1.6743	1.0018	0.0946
VIII	1.4694	0.9043	0.1042
Edad	0.0413	0.0241	0.0876
Televisión	3.2729	0.8890	0.0002
Antigüedad Banda Ancha	0.0306	0.0105	0.0037
Log(Minutos Salientes)	-2.6176	0.9154	0.0042
Log(Llamadas Salientes)	2.3634	1.0644	0.0264
Log(ARPU Voz)	-1.2908	0.8169	0.1140
Log(ARPU Banda Ancha)	-0.7725	0.3367	0.0217
Log(ARPU TV)	-0.3873	0.2119	0.0675
Días Mora	-0.0886	0.0611	0.1470
Tópico 2	9.6378	5.2373	0.0657
Tópico 8	21.9199	8.0077	0.0061
Tópico 1	-17.7057	7.2362	0.0144
Tópico 4	-8.1103	5.0441	0.1078

Las variables demográficas y transaccionales seleccionadas en el modelo son similares a las seleccionadas en la sección 9.1, la diferencia es que en este modelo se agregan las variables *VIII* y *Edad*, mientras que se quitan *Antigüedad Cliente* y *METROP*. Los signos de los coeficientes se comportan de la misma forma, por lo que la interpretación es la misma.

De las variables *VIII* y *Edad*, solo *Edad* es significativa, y al tener un coeficiente positivo indica que a mayor edad aumenta la probabilidad de dar de baja un producto. De todas formas el valor de beta es bajo ($\beta = 0.04$), por lo que el efecto que produce la edad en el modelo no es alto.

Con respecto a las variables de tópicos, se puede notar que los signos de éstos se comportan igual que en el modelo de la sección 9.3, donde las variables Tópico 2 y Tópico 8 inclinan el modelo hacia la fuga, y las variables Tópico 1 y Tópico 4 disminuyen esta probabilidad.

Lo anterior se puede explicar viendo la Tabla 22, donde es posible notar que las variables Tópico 2 y Tópico 8 son las que aparecen con mayor frecuencia entre las llamadas de los clientes fugados, por lo que son tópicos importantes para poder estimar la probabilidad de fuga de un cliente. Mientras que las variables Tópico 1 y Tópico 4 son las que tienen una menor frecuencia entre los fugados.

De esta forma llamar por problemas con la conexión a internet (Tópico 2) y llamar por temas específicos que necesitan ser atendidos por un especialista (Tópico 8) aumentan la probabilidad de fuga en el modelo.

La interpretación del Tópico 8 (transferencia) se asocia a que cuando un cliente está insatisfecho con un producto o servicio y llama al Call Center, lo hace con un objetivo específico que en ciertos casos puede ser la baja del producto. Este tipo de llamadas tiene que ser atendida por un ejecutivo en especial, ya sea de soporte técnico o el encargado de las bajas, por lo que es necesario transferir al cliente a una línea especial. Es por esto que este tópico se encuentra en primer lugar en la tabla de frecuencias de los clientes fugados, ya que es un tema que se repite en la mayoría de las llamadas.

9.5. Comparación entre Modelos

En la Tabla 31 se muestran los 3 modelos elegidos en las secciones anteriores, con sus valores de AIC y BIC y el número de variables de cada uno.

Tabla 31. Modelos elegidos.

Modelo	Número de variables	AIC	BIC
Modelo A	13	137.1	172.74
Modelo B	15	145.1	186.29
Modelo C	16	126	172.64

El Modelo A corresponde al seleccionado utilizando variables demográficas y transaccionales, el Modelo B es el seleccionado utilizando las variables de las llamadas, y finalmente el Modelo C utiliza todo el conjunto de variables.

Para iniciar la comparación se estudiarán las métricas correspondientes a la matriz de confusión y las curvas ROC para cada modelo, tanto en el conjunto de entrenamiento como en el de testeo en la proporción 80/20 con punto de corte 0.5. Además se incluye el valor del error cuadrático medio (MSE) calculado utilizando la técnica Leave one Out Cross Validation y la medida *Lift.*

Para iniciar la comparación se partirá analizando cada modelo por separado según las métricas mostradas en la Tabla 32.

Tabla 32. Métricas matriz de confusión para Modelo A, Modelo B y Modelo C.

Entrenamiento	Modelo A (Demográfico y Transaccional)	Modelo B (Tópicos)	Modelo C (Demográfico y Transaccional + Tópicos)
Accuracy	0.8152	0.7065	**0.8804**
Precision	0.8070	0.7192	**0.9019**
Sensitivity	**0.8846**	0.7884	**0.8846**
Specificity	0.725	0.6	**0.875**
False Positive	0.275	0.4	**0.125**
AUC	0.8832	0.8077	**0.9361**
Lift	1.42	1.27	**1.59**
Testeo			
Accuracy	0.6521	0.6956	**0.7391**
Precision	0.6363	0.6428	**0.8571**
Sensitivity	0.6363	**0.8181**	0.5454
Specificity	0.6666	0.5833	**0.9166**
False Positive	0.3333	0.4166	**0.0833**
AUC	0.6818	**0.7576**	0.7121
Lift	1.33	1.34	**1.79**
Leave One Out CV			
MSE	0.2077	0.2377	**0.2034**

Ninguno de los tres modelos es superior al otro en todas las métricas estudiadas, la elección dependerá de qué métrica es considerada más relevante para el análisis.

Las métricas más importantes para este análisis son accuracy, sensitivity, false positive y MSE. Esto ya que accuracy mide la exactitud general del modelo, por lo tanto es una métrica importante al momento de realizar comparaciones. Como la clase de interés son los clientes fugados, las métricas sensitivity y false positive son relevantes para el análisis, debido a que indican la proporción de fugados bien clasificados por sobre el total de fugados y la cantidad de fugados mal clasificados, respectivamente. Finalmente el error cuadrático medio (MSE) es relevante al realizar comparaciones ya que representa el error de predicción de los modelos. De esta forma es posible hacer una selección del modelo considerando cada uno de estos indicadores.

Analizando brevemente el conjunto de entrenamiento, es posible observar que el modelo C es superior en accuracy y false positive a los modelos A y B, y en lo que respecta a la medida sensitivity presenta el mayor valor junto al modelo A. Por lo tanto el modelo C es superior dentro de la muestra.

Ahora analizando el conjunto de testo, tomando en cuenta la medida accuracy, el modelo seleccionado corresponde al modelo C, con un accuracy de 73.91%. Siendo mayor al modelo A en un 8.7% y al modelo B en un 4.35%.

Considerando la medida sensitivity, el mejor modelo sería el modelo B, con un 81.81%. Siendo mayor al modelo A en un 18.18%, y al modelo C en un 27.27%.

Con respecto al false positive, el mejor modelo es el modelo C ya que presenta el menor error al clasificar como fugado a un cliente, que corresponde a un 8.3%. El modelo A se equivoca un 25.03% más, y el modelo B un 33.36% más.

Finalmente considerando el error cuadrático medio (MSE) el mejor modelo corresponde al modelo C, con un 20.34%. Siendo menor al modelo A en un 0.43%, y al modelo C en un 3.43%.

Dado el análisis anterior, es posible seleccionar el modelo C, ya que es superior en la mayoría de las métricas. El único punto en contra es la sensibilidad, pero esto se equilibra al tener un bajo false positive (en comparación a los demás modelos).

De esta forma es posible concluir que un modelo que solo utiliza variables demográficas y transaccionales, efectivamente puede ser mejorado incluyendo información sobre las interacciones que tienen los clientes con el Call Center.

Para poder estimar el impacto que tienen las variables de tópicos en el modelo, es necesario comparar el modelo A con el modelo B y C, para así ver en cuánto mejoran los modelos si se le agregan variables de las llamadas. En la Tabla 33 se muestra esta comparación, donde el signo positivo indica que el modelo es superior en ese porcentaje al modelo A.

Tabla 33. Comparación modelo A con modelo B y C.

	Modelo B	Modelo C
Accuracy	+4.35%	+8.7%
Sensitivity	+18.18%	-9.09%
False Positive	-8.33%	+25.03%
MSE	-3.00%	+0.43%

De esta forma, agregar las interacciones con el Call Center a los modelos de predicción de fuga que solo utilizan variables demográficas y transaccionales mejora la exactitud en un 8.7%.

Por otra parte, a pesar de que el modelo B sea un 4.35% más exacto que un modelo que solo utiliza variables demográficas y transaccionales, presenta errores de clasificación mayores. De este modo un modelo solo con variables de tópicos o solo con variables demográficas y transaccionales no será superior a uno que utilice ambas.

Además de concluir que los tópicos sí son relevantes y de importancia para los modelos de fuga, es posible definir cuáles son los más importantes y cuáles están más asociados con la baja de un producto.

Según los modelos construidos, es posible separar los tópicos entre los que aumentan la probabilidad de fugarse y aquellos que no. Los que se asocian con la fuga son los siguientes:

Tópico 2 = Problemas con la conexión a internet hogar.

Tópico 7 = Cambios de algún producto/servicio o intentos de baja.

Tópico 8 = Llamadas específicas que necesitan ser transferidas a un ejecutivo especialista.

Por lo que si un cliente realiza una llamada que presente estos tópicos en mayor proporción podría tratarse de una posible baja.

Los tópicos que no están asociados con la fuga son los que se muestran a continuación.

Tópico 1 = Bloqueos o desbloqueos de algún producto.

Tópico 3 = Reclamos en general, llamadas por algún tipo de problema.

Tópico 4 = Llamadas asociadas a problemas o consultas sobre la página web de la empresa.

Tópico 5 = Contactabilidad.

Tópico 6 = Problemas técnicos con la televisión o el decodificador.

De los tópicos anteriores, los tópicos 3 y 5 a pesar de no estar asociados con la fuga, son tópicos con una alta frecuencia de aparición en las llamadas de

los clientes fugados. Para el caso del Tópico 3 que corresponde a reclamos en general, esto se debe a que la mayoría de las llamadas al Call Center (de clientes que se van a fugar o no) son por reclamos o problemas por parte del cliente, sino no estaría llamando. Por lo que a pesar de ser un tópico más presente en el grupo de los clientes fugados al ser general no aumenta la probabilidad de fuga. Lo mismo con el Tópico 5, está más presente en los clientes fugados pero también es un tópico que se repite en muchas llamadas y no es tan específico, por lo que no es asociado directamente con la fuga.

Con esta diferenciación entre los tópicos, se puede concluir que los clientes que se fugan realizan llamadas más específicas, es decir, cuentan directamente su problema y quieren ser atendidos. Por lo mismo es que presentan en promedio un menor número de llamadas al Call Center que los clientes que permanecen con los mismos productos.

En cambio, los clientes que permanecen activos (en la mayoría de los casos) realizan llamadas más generales, para hacer consultas sobre sus planes contratados, o sobre la página web y sus puntos, etc.

9.6. Evaluación Económica

Llevando esta comparación entre los modelos a lo que implica para la empresa el uso de uno u otro, es posible realizar un análisis económico basándose en las métricas mencionadas anteriormente, en el valor promedio de los clientes fugados para la empresa (medido en ARPU) y en los costos asociados a efectuar acciones de retención.

El análisis se realizará para distintos escenarios, definidos como positivo, intermedio y negativo, donde en cada escenario varía el porcentaje de clientes efectivamente retenidos luego de acciones correctivas por parte de la empresa.

Utilizando la medida sensitivity (calculada con un punto de corte de 0.5), la cual entrega la proporción de clientes fugados bien clasificados versus el total de fugados, es posible calcular la Fuga de Ingresos Prevenida (FIP) para cada modelo:

$$\left[\left(Sensitivity * (N°Clientes * TasaChurn)\right) * ARPU * Escenario\right]$$

Donde $(N° Clientes * TasaChurn)$ es la cantidad promedio de clientes que se dieron de baja durante Agosto, y al multiplicarlo por la medida $Sensitivity$ se obtiene la cantidad de clientes fugados que fueron bien clasificados por el modelo, y por lo tanto, detectados. Luego con el valor de $ARPU$ se tiene la valoración monetaria de este grupo de clientes para la empresa.

El valor promedio del ARPU para los clientes que dieron de baja algún producto en el mes de Agosto corresponde a $32835.22, y la tasa de CHURN promedio para este mes es de 1.16%.

El $Escenario$ tomará los valores de 0.8, 0.5 y 0.2 según sea positivo, intermedio o negativo respectivamente, lo que indica la proporción de clientes que fueron clasificados como fugados y que se logró retener.

El número de clientes con contratos de televisión, internet y telefonía durante Agosto no es una cifra fácil de calcular, debido a los sesgos de información y selección mencionados en la sección 8.1. Dada la naturaleza de los modelos, los cuales utilizan información de la interacción con el Call Center, es necesario tomar en cuenta clientes que hayan realizado llamadas.

Como ya se explicó, existe una fracción de clientes que llamaron al Call Center de un teléfono que no está asociado a la compañía, por lo tanto no pueden ser detectados.

De esta forma, para el siguiente análisis se utilizará la cantidad de clientes que sí puede ser detectada, que para el mes de Agosto corresponde a 56127 clientes, los cuales llaman al Call Center desde su teléfono fijo o móvil asociado a la compañía.

La inversión en retención se calculará utilizando las medidas Sensitivity y False Positive de cada modelo, así se podrá obtener los costos de retener a clientes que efectivamente se darán de baja y los costos por retener clientes mal clasificados. Las fórmulas para el cálculo de la inversión se muestran a continuación:

$$Inversión_{sensitivity} = \big(Sensitivity * (N°Clientes * TasaChurn)\big) * CostoRetención$$

$$Inversión_{FalsePositive} = (FalsePositive * (N°Clientes * (1 - TasaChurn))) * CostoRetención$$

Los costos asociados a la contención de clientes están formados por una componente fija y otra variable. La componente fija corresponde al costo de

contactar al cliente, lo que incluye llamarlo y atenderlo por lo tanto se incurre en gastos al realizar la llamada y horas hombre, mientras que la componente variable corresponde a la promoción o descuento que se le ofrece. El costo fijo equivale a $4000 y se incurre una sola vez, mientras que el costo variable toma un valor promedio de $15000 a lo largo de 6 meses, que es la duración del descuento o promoción ofrecidos al cliente[4]. De esta forma, se toma como supuesto que el cliente permanecerá afiliado a la compañía por los siguientes 6 meses.

Dado que los costos involucrados son por contactar al cliente y luego por el descuento ofrecido, el análisis se separará en un mes inicial (costo de contactar al cliente) y luego si el cliente decidió quedarse en la compañía, se consideran 6 meses en que se le hace un descuento, por lo que los costos en inversión del segundo mes al último son los mismos. Así se calcularán los Ingresos Netos en un horizonte corto de planificación (6 meses), ya que al séptimo mes se termina la promoción y se toma como supuesto que el cliente vuelve al estado normal donde tiene la misma probabilidad de fugarse y se le ofrecería nuevamente algún descuento.

Lo anterior podría modificarse y alcanzar un horizonte de planificación más allá de los 6 meses si se conociera la probabilidad de un cliente de permanecer en la compañía terminado los 6 meses de descuento y la probabilidad de permanecer en la compañía de un cliente al cual no se le han hecho acciones de retención. De este modo en el séptimo mes se tendrían distintas tasas de retención y se podría seguir con el cálculo. Al no ser este el caso es que se tomaron los supuestos ya mencionados y el cálculo se realizará hasta el mes número 6.

Es posible calcular los Ingresos Netos como sigue:

$$IngresosNetos = FIP - (Inversión_{sensitivity} + Inversión_{FalsePositive})$$

La inversión en retención debido a errores en la clasificación de los clientes será la misma en todos los escenarios, solo presentará variaciones si se trata del primer mes en que se contacta al cliente o luego de ya ofrecido el descuento. Esto ya que se toma como supuesto que ofrecer un descuento a un cliente que no está propenso a fugarse será aceptado por éste. De esta forma los costos por contactar y ofrecer descuentos a clientes

[4] Costos asociados a la contención de clientes fueron entregados por la empresa en estudio.

incorrectamente clasificados como fugados son los siguientes para cada modelo:

Mes 1:

$$Inversión_{FalsePositive}(A) = \left[0.3333 * \left(56127 * (1 - 1.16\%)\right) * \$6500\right] = \$\ 120.185.821{,}62$$

$$Inversión_{FalsePositive}(B) = \left[0.4166 * \left(56127 * (1 - 1.16\%)\right) * \$6500\right] = \$\ 150.223.262{,}18$$

$$Inversión_{FalsePositive}(C) = \left[0.0833 * \left(56127 * (1 - 1.16\%)\right) * \$6500\right] = \$\ 30.037.440{,}57$$

Mes 2 a 6:

$$Inversión_{FalsePositive}(A) = \left[0.3333 * \left(56127 * (1 - 1.16\%)\right) * \$2500\right] = \$\ 46.225.316{,}01$$

$$Inversión_{FalsePositive}(B) = \left[0.4166 * \left(56127 * (1 - 1.16\%)\right) * \$2500\right] = \$\ 57.778.177{,}76$$

$$Inversión_{FalsePositive}(C) = \left[0.0833 * \left(56127 * (1 - 1.16\%)\right) * \$2500\right] = \$\ 11.552.861{,}76$$

Los cálculos para cada escenario se muestran a continuación:

Escenario Positivo: Se logra retener el 80%

Mes 1:

$$FIP_A = \left[\left(0.6363 * (56127 * 1.16\%)\right) * \$32835.22 * 0.8\right]$$

$$FIP_A = \$\ 10.882.324{,}19$$

$$Inversión_{Sensitivity}(A) = \left[\left(0.6363 * (56127 * 1.16\%)\right) * \$4000\right] + [(0.6363 * (56127 * 1.16\%) * 0.8) * \$2500] = \$\ 2.485.667{,}26$$

$$IngresosNetos_A = -\$111.789.164{,}69$$

$$FIP_B = \left[(0.8181 * (56127 * 1.16\%)) * \$32835.22 * 0.8\right]$$

$$FIP_B = \$\ 13.991.559{,}67$$

$$Inversión_{Sensitivity}(B) = \left[\left(0.8181 * (56127 * 1.16\%)\right) * \$4000\right] + [(0.8181 * (56127 * 1.16\%) * 0.8) * \$2500] = \$\ 3.195.857{,}91$$

$$IngresosNetos_B = -\$139.427.560{,}42$$

$$FIP_C = \left[\left(0.5454 * (56127 * 1.16\%)\right) * \$32835.22 * 0.8\right]$$

$$FIP_C = \$\ 9.327.706{,}45$$

$$Inversión_{Sensitivity}(C) = \left[\left(0.5454 * (56127 * 1.16\%)\right) * \$4000\right] + [(0.5454 * (56127 * 1.16\%) * 0.8) * \$2500] = \$\ 2.130.571{,}94$$

$$IngresosNetos_C = -\$22.840.306{,}06$$

Mes 2 al 6:

$$FIP_A = \$\ 10.882.324{,}19$$

$$Inversión_{Sensitivity}(A) = [(0.6363 * (56127 * 1.16\%) * 0.8) * \$2500] = \$\ 828.555{,}75$$

$$IngresosNetos_A = -\$36.171.547{,}57$$

$$FIP_B = \$\ 13.991.559{,}67$$

$$Inversión_{Sensitivity}(B) = [(0.8181 * (56127 * 1.16\%) * 0.8) * \$2500] = \$\ 1.065.285{,}97$$

$$IngresosNetos_B = -\$44.851.904{,}06$$

$$FIP_C = \$\ 9.327.706{,}45$$

$$Inversión_{Sensitivity}(C) = [(0.5454 * (56127 * 1.16\%) * 0.8) * \$2500] = \$\ 710.190{,}65$$

$$IngresosNetos_C = -\$2.935.345{,}95$$

Escenario Intermedio: Se logra retener el 50%

Mes 1:

$$FIP_A = \left[\left(0.6363 * (56127 * 1.16\%)\right) * \$32835.22 * 0.5\right]$$

$$FIP_A = \$\ 6.801.452{,}62$$

$$Inversión_{Sensitivity}(A) = \left[\left(0.6363 * (56127 * 1.16\%)\right) * \$4000\right] + [(0.6363 * (56127 * 1.16\%) * 0.5) * \$2500] = \$\ 2.174.958{,}86$$

$$IngresosNetos_A = -\$115.559.327{,}85$$

$$FIP_B = \left[\left(0.8181 * (56127 * 1.16\%)\right) * \$32835.22 * 0.5\right]$$

$$FIP_B = \$\,8.744.724{,}80$$

$$Inversión_{Sensitivity}(B) = \left[\left(0.8181 * (56127 * 1.16\%)\right) * \$4000\right] + [(0.8181 * (56127 * 1.16\%) * 0.5) * \$2500] = \$\,2.796.375{,}67$$

$$IngresosNetos_B = -\$144.274.913{,}06$$

$$FIP_C = \left[\left(0.5454 * (56127 * 1.16\%)\right) * \$32835.22 * 0.5\right]$$

$$FIP_C = \$\,5.829.816{,}53$$

$$Inversión_{Sensitivity}(C) = \left[\left(0.5454 * (56127 * 1.16\%)\right) * \$4000\right] + [(0.5454 * (56127 * 1.16\%) * 0.5) * \$2500] = \$\,1.864.250{,}45$$

$$IngresosNetos_C = -\$26.071.874{,}48$$

Mes 2 a 6:

$$FIP_A = \$\,6.801.452{,}62$$

$$Inversión_{Sensitivity}(A) = [(0.6363 * (56127 * 1.16\%) * 0.5) * \$2500] = \$\,517.847{,}35$$

$$IngresosNetos_A = -\$39.941.710{,}73$$

$$FIP_B = \$\,8.744.724{,}80$$

$$Inversión_{Sensitivity}(B) = [(0.8181 * (56127 * 1.16\%) * 0.5) * \$2500] = \$\,665.803{,}73$$

$$IngresosNetos_B = -\$49.699.256{,}70$$

$$FIP_C = \$\,5.829.816{,}53$$

$$Inversión_{Sensitivity}(C) = [(0.5454 * (56127 * 1.16\%) * 0.5) * \$2500] = \$\,443.869{,}15$$

$$IngresosNetos_C = -\$6.166.914{,}38$$

Escenario Negativo: Se logra retener el 20%

Mes 1:

$$FIP_A = \left[\left(0.6363 * (56127 * 1.16\%)\right) * \$32835.22 * 0.2\right]$$

$$FIP_A = \$\ 2.720.581.05$$

$$Inversión_{Sensitivity}(A) = \left[\left(0.6363 * (56127 * 1.16\%)\right) * \$4000\right] + [(0.6363 * (56127 * 1.16\%) * 0.2) * \$2500] = \$\ 1.864.250{,}45$$

$$IngresosNetos_A = -\$119.329.491{,}02$$

$$FIP_B = \left[\left(0.8181 * (56127 * 1.16\%)\right) * \$32835.22 * 0.2\right]$$

$$FIP_B = \$\ 3.497.889{,}92$$

$$Inversión_{Sensitivity}(B) = \left[\left(0.8181 * (56127 * 1.16\%)\right) * \$4000\right] + [(0.8181 * (56127 * 1.16\%) * 0.2) * \$2500] = \$\ 2.396.893{,}43$$

$$IngresosNetos_B = -\$149.122.265{,}70$$

$$FIP_C = \left[\left(0.5454 * (56127 * 1.16\%)\right) * \$32835.22 * 0.2\right]$$

$$FIP_C = \$\ 2.331.926{,}61$$

$$Inversión_{Sensitivity}(C) = \left[\left(0.5454 * (56127 * 1.16\%)\right) * \$4000\right] + [(0.5454 * (56127 * 1.16\%) * 0.2) * \$2500] = \$\ 1.597.928{,}95$$

$$IngresosNetos_C = -\$29.303.442{,}91$$

Mes 2 a 6:

$$FIP_A = \$\ 2.720.581.05$$

$$Inversión_{Sensitivity}(A) = [(0.6363 * (56127 * 1.16\%) * 0.2) * \$2500] = \$\ 207.138{,}94$$

$$IngresosNetos_A = -\$43.711.873{,}90$$

$$FIP_B = \$\ 3.497.889{,}92$$

$$Inversión_{Sensitivity}(B) = [(0.8181 * (56127 * 1.16\%) * 0.2) * \$2500] = \$\ 266.321{,}49$$

$$IngresosNetos_B = -\$54.546.609{,}34$$

$$FIP_C = \$\,2.331.926{,}61$$

$$Inversión_{Sensitivity}(C) = [(0.5454 * (56127 * 1.16\%) * 0.2) * \$2500] = \$\,177.547{,}66$$

$$IngresosNetos_C = -\$9.398.482{,}81$$

Como es posible notar, los ingresos netos calculados para cada modelo son negativos, lo que estaría representando una pérdida para la empresa. Esto ocurre debido a los costos asociados a las campañas de retención, y además a los errores de los modelos al clasificar como fugado a un cliente que no lo es, ya que se gastan recursos innecesarios.

El modelo A cuenta con un false positive de 32.94%, el modelo B con 41.17% y el modelo C con 8.23%, que sería la cantidad de clientes clasificados como fugados incorrectamente, sobre el total de clientes activos. Lo anterior en conjunto con otras métricas permite calcular el porcentaje de clientes clasificados como fugados del total de clientes, el modelo A clasifica al 33.68% como fugado, el modelo B al 42.13% y el modelo C al 8.86%, siendo que la tasa de fuga promedio es de un 1.16%. Por lo tanto es evidente que existe un gran porcentaje de error al clasificar a los clientes, lo que se puede explicar por los sesgos existentes en cuanto a la información disponible y a las limitaciones del modelo mismo.

Debido a los sesgos presentes en el desarrollo de los modelos y en el proyecto en general, es que los cálculos realizados son aproximaciones. Los ingresos netos calculados están dejando de lado una gran cantidad de clientes, se podría hacer un cálculo más exacto si el software desde el cual se extrae la información fuera capaz de identificar a todos los clientes que llaman al Call Center no solo desde el número del cual realizan la llamada sino que también con el RUT. Además de que el software solo guarda el 30% de las llamadas totales registradas.

Asimismo, con esta información disponible se podrían construir modelos con un mejor ajuste y capacidad predictiva disminuyendo el error de clasificación y así también disminuyendo los costos asociados.

Sin embargo, el objetivo de la presente memoria es estimar el impacto del uso de las variables de las interacciones con el Call Center en modelos de predicción de fuga, por lo que lo interesante es analizar las diferencias existentes entre los 3 modelos construidos. De esta forma es posible

determinar si la inclusión de las variables de las llamadas mejora los modelos que no las utilizan y poder cuantificar esta diferencia.

Dado que el modelo A es el modelo básico (que no utiliza variables de llamadas) se comparará el promedio de Ingresos Netos para 6 meses de este modelo versus los modelos B y C.

Escenario Positivo

$$\overline{IngresosNetos_B} - \overline{IngresosNetos_A} = -\$\ 11.840.029{,}70$$

$$\overline{IngresosNetos_C} - \overline{IngresosNetos_A} = +\$\ 42.521.644{,}45$$

Escenario Intermedio

$$\overline{IngresosNetos_B} - \overline{IngresosNetos_A} = -\$\ 12.917.219{,}17$$

$$\overline{IngresosNetos_C} - \overline{IngresosNetos_A} = +\$\ 43.060.239{,}19$$

Escenario Negativo

$$\overline{IngresosNetos_B} - \overline{IngresosNetos_A} = -\$\ 13.994.408{,}65$$

$$\overline{IngresosNetos_C} - \overline{IngresosNetos_A} = +\$\ 43.598.833{,}93$$

Con lo anterior es posible apreciar que el modelo C es superior al modelo A y B, ya que a pesar de contar con ingresos netos negativos, es superior al modelo base en aproximadamente 43 millones de pesos.

En conclusión, agregar variables del contenido de las llamadas a los modelos base (con variables demográficas y transaccionales) aumenta los ingresos netos percibidos, de forma que el impacto que generan los tópicos en los modelos de predicción de fuga es posible estimarlos en aproximadamente 43 millones de pesos. Cabe destacar que este valor no está considerando el costo en horas hombres para la construcción de modelos ni para la implementación de herramientas de text mining, ya que incluir tópicos en los modelos traería consigo un mayor costo en implementación.

Dados los resultados anteriores, es posible realizar un análisis extra y ajustar el punto de corte, de modo de aumentar la exigencia al momento de definir como fugado a un cliente. Realizando el mismo cálculo anterior, se obtiene el promedio de Ingresos Netos para 6 meses con distintos puntos de corte para cada modelo. La Tabla 34 muestra el escenario positivo, la Tabla 35 el escenario intermedio y la Tabla 36 el negativo.

Tabla 34. Promedio de Ingresos Netos para 6 meses en escenario Positivo, se logra retener el 80%.

Punto de Corte	Promedio de Ingresos Netos		
	Modelo A	Modelo B	Modelo C
0.5	-$48.774.483,76	-$60.614.513,45	-$6.252.839,30
0.6	-$36.934.454,06	-$50.171.281,35	-$7.649.636,90
0.7	-$22.283.261,79	+$6.983.987,99	-$7.649.636,90
0.8	-$7.649.636,90	+$6.983.987,99	-$7.649.636,90
0.9	-$10.443.232,10	+$6.983.987,99	-$9.046.434,50

Tabla 35. Promedio de Ingresos Netos para 6 meses en escenario Intermedio, se logra retener el 50%.

Punto de Corte	Promedio de Ingresos Netos		
	Modelo A	Modelo B	Modelo C
0.5	-$52.544.646,92	-$65.461.866,09	-$9.484.407,73
0.6	-$39.627.427,75	-$53.402.849,78	-$10.342.610,59
0.7	-$24.976.235,48	+$4.291.014,30	-$10.342.610,59
0.8	-$10.342.610,59	+$4.291.014,30	-$10.342.610,59
0.9	-$12.059.016,31	+$4.291.014,30	-$11.200.813,45

Tabla 36. Promedio de Ingresos Netos para 6 meses en escenario Negativo, se logra retener el 20%.

Punto de Corte	Promedio de Ingresos Netos		
	Modelo A	Modelo B	Modelo C
0.5	-$56.314.810,08	-$70.309.218,73	-$12.715.976,16
0.6	-$42.320.401,44	-$56.634.418,21	-$13.035.584,28
0.7	-$27.669.209,17	+$1.598.040,61	-$13.035.584,28
0.8	-$13.035.584,28	+$1.598.040,61	-$13.035.584,28
0.9	-$13.674.800,52	+$1.598.040,61	-$13.355.192,40

Como es posible observar, el modelo B presenta Ingresos Netos positivos cuando el punto de corte es superior a 0.6. Del punto de corte 0.6 a 0.7 pasa de ser el peor modelo a ser el mejor (en términos de ingresos netos). Esto ocurre debido a que al ajustar el punto de corte en 0.7, la medida False Positive del modelo B es cero. Es decir, no se equivoca al clasificar como fugado a un cliente, lo que disminuye considerablemente los gastos de inversión en retención de clientes.

Tomando en cuenta el punto de corte 0.8, el modelo B es superior en aproximadamente 14 millones al modelo A y C. Lo que quiere decir que un modelo de predicción de fuga que utilice solo las variables de las llamadas conlleva a un aumento promedio de ingresos netos de 14 millones de pesos.

Comparando los modelos A y C, en todos los puntos de corte el modelo C es igual o superior al modelo A. Por lo tanto, haciendo el análisis con puntos de cortes desde 0.5 a 0.9, los modelos que entregan los mejores Ingresos Netos promedios son los modelos B y C, es decir, lo que cuentan con variables de las llamadas.

Así es posible reafirmar la conclusión anterior y decir que las variables de las llamadas son un buen aporte a los modelos de predicción de fuga, ya que mejoran la capacidad predictiva de éstos.

10. CONCLUSIONES

10.1. Conclusiones Generales

En la mayoría de las empresas se cuenta con más información de la que utilizan, o no saben cómo utilizarla. Específicamente la empresa en la cual se basa este estudio cuenta con una fuente de información que no ha sido utilizada, la cual corresponde a las llamadas que realizan los clientes al Call Center de la compañía.

Con el desarrollo de las tecnologías el cliente cada vez tiene más acceso a información sobre los productos o servicios que desea comprar, lo que al mismo tiempo provoca que sea más exigente. Es por esto que la voz de los usuarios es cada vez más importante para mantenerlos satisfechos y evitar que se pierdan en la competencia.

Sin embargo, los clientes no son los únicos que se ven beneficiados con el desarrollo de las tecnologías, las empresas también están invirtiendo más en capturar información valiosa sobre sus clientes. Esto con el objetivo de crear modelos de predicción de fuga cada vez más sofisticados y campañas de retención que logren conservar a sus clientes.

El Call Center podría ser una buena herramienta para lograr lo anterior, ya que permite conocer a los clientes e identificar sus problemas con los servicios entregados. Es por esto que esta memoria pretende estimar el impacto que generan las llamadas en el Call Center en los modelos de fuga, para así definir si esta fuente de información es relevante en cuanto a la predicción de fuga de los clientes.

Tras realizar tres modelos de predicción de fuga utilizando distintos tipos de variables, el primero utilizando solo variables que describen al cliente, el segundo solo de las llamadas, y el último con ambos tipos de variables, se llegó a la conclusión que la información de las llamadas registradas en el Call Center efectivamente mejora los modelos de predicción de fuga y por lo tanto, es información que podría ser aprovechada por la empresa estudiada.

Si se utilizan ambos tipos de variables en la predicción de fuga, el modelo aumenta su exactitud en un 8.7% con respecto a un modelo que no utilice la información de las llamadas. Además, en términos de ingresos netos, un

modelo con ambos tipos de variables es superior en aproximadamente 43 millones de pesos a uno que no utilice esta nueva información.

Sin embargo, los tres modelos construidos cuentan con altos grados de error en lo que respecta a la incorrecta clasificación de clientes fugados, lo que se ve reflejado en costos innecesarios en retención de clientes. Esto se debe principalmente a los sesgos existentes en la información al momento de extraer la muestra y a las limitaciones con la disponibilidad de la misma.

Aun así, es posible concluir que el contenido de las llamadas sí agrega valor a los modelos de predicción de fuga, por lo que conocer lo que se habla en ellas es una fuente de información valiosa y de gran importancia en lo que respecta a la retención de los clientes.

Dados los análisis que se realizaron, se pudo detectar aquellos tópicos que tienen una mayor probabilidad de estar asociado a una fuga, que corresponden principalmente a llamadas específicas que necesitan ser transferidas a un área determinada o a un ejecutivo especialista. Lo anterior se puede explicar con que si un cliente no está satisfecho con su producto o servicio y desea darlo de baja, llamará al Call Center con un objetivo específico y tendrá que ser transferido al área que se encarga de las bajas. Un posible inconveniente de este tópico es que al momento de ser detectado puede ser muy tarde para tratar de retener al cliente. Por ser un tópico tan específico es que se encuentra en el primer lugar de los temas más frecuentes en las llamadas de los clientes que se fugan.

Otro tópico asociado a la fuga son los problemas con la conexión a internet, donde el cliente expresa claramente que presenta problemas con su conexión o con el funcionamiento del módem. Finalmente se tiene el tópico asociado a cambios, ya que generalmente los productos que ofrece la compañía se venden en planes, promociones o packs, por lo que si un cliente no está satisfecho y quiere cambiar su pack, puede conllevar a una baja de algún producto.

En conclusión, con los análisis realizados ha sido posible encontrar aquellos tópicos asociados a la baja de algún producto, y determinar su impacto en lo que respecta a modelos de predicción de fuga.

10.2. Recomendaciones y Propuestas de Trabajo Futuro

Dado el análisis realizado, y que ha sido posible establecer la importancia de las interacciones con el Call Center en lo que respecta a modelos de fuga de forma positiva, una de las recomendaciones esenciales es efectivamente utilizar la información respectiva al contenido de las llamadas en el futuro.

Sin embargo, es necesario eliminar los sesgos existentes en cuanto a la obtención de la información, por lo que para utilizar esta herramienta se recomienda contar con toda la información de las llamadas (no solo el 30%) y además reconocer en el sistema el RUT del cliente, y no solo del número desde el cual realiza la llamada. Por lo que sería necesaria una inversión en el desarrollo del software de Speech Analytics con el que cuentan para poder tener toda la información necesaria y así eliminar los sesgos presentes en este estudio.

Teniendo libre acceso a esta nueva fuente de información, los modelos de predicción de fuga podrían ser mejorados. Además se tendría conocimiento de los temas más frecuentes en las llamadas al Call Center, lo que permitiría detectar aquellos tópicos asociados a la posible fuga, para así contar con un factor que logre explicar el comportamiento del cliente y por lo tanto, que mejore el desempeño del modelo.

De esta forma, se podría generar un número estándar de tópicos y definir sus etiquetas de forma mensual, de modo de ir actualizando los temas tratados según sean las contingencias del momento.

En el caso que no sea posible contar con toda esta información, una solución podría ser generar manualmente tópicos estándar en el software de Speech Analytics. Así podrán ser incluidos en los modelos de predicción de fuga pero utilizando altos puntos de corte, de modo de evitar errores de predicción. Con esto podría ser posible detectar grupos de clientes, aunque sean pequeños, que estén propensos a fugarse y realizar acciones de retención.

La generación de tópicos y el conocimiento de estos es de gran importancia para la empresa, no solo en lo que respecta a la performance de los modelos de predicción de fuga. Detectando los tópicos con mayor presencia en las llamadas de clientes que deseen dar de baja un producto, la empresa podrá tener conciencia de cuáles son los temas que están provocando la baja. De

este modo, podrán enfocarse específicamente en estos problemas y tratar de solucionarlos para evitar seguir perdiendo clientes o dejándolos insatisfechos.

Según este trabajo uno de los tópicos asociados a la fuga es la conexión a internet y problemas con el funcionamiento del módem, pero este tópico podría estar sesgado por factores de estacionalidad o temporalidad. Esto se debe a que la muestra que se utilizó corresponde a llamadas realizadas en un mes, por lo que podría ocurrir que en el mes estudiado los clientes tuvieron mayores problemas con la conexión a internet en sus casas y decidieron darla de baja. Por lo que utilizando toda la información disponible los tópicos asociados a la fuga podrían cambiar, y se podría generar nuevo conocimiento en lo que respecta a las llamadas que terminan en la baja de un producto.

Finalmente, dado que esta memoria está basada en clientes *personas* que se fugan por razones voluntarias o por portabilidad, también sería interesante conocer cómo se comportan las empresas y si los tópicos presentes en sus llamadas son similares a los de las personas naturales. Del mismo modo, es interesante analizar los tópicos presentes en aquellas fugas por motivos morosos, y si estos clientes se comportan de forma distinta en cuanto a las interacciones con el Call Center.

11. BIBLIOGRAFÍA

[1] C. Ranaweera y J. Prabhu, «On the relative importance of customer satisfaction and trust as determinants of customer retention and positive word of mouth,» *Journal of Targeting, Measurement and Analysis for Marketing,* vol. 12, pp. 82-90, 2003.

[2] A. Nadali, E. N. Kakhky y E. H. Nosratabadi, «Evaluating the Success Level of Data Mining Projects Based on CRISP-DM Methodology by a Fuzzy Expert System,» *Electronics Computer Technology (ICECT),* vol. 6, pp. 161-165, 2011.

[3] Z. G. O. &. B. S. Bosnjak, «CRISP-DM as a Framework for Discovering Knowledge in Small and Medium Sized Enterprises' Data,» *Applied Computational Intelligence and Informatics,* pp. 509-514, 2009.

[4] D. M. Blei, A. Ng y M. Jordan, «Latent Dirichlet Allocation,» *Journal of Machine Learning Research,* pp. 993-1022, 2003.

[5] C. Reed, *Latent Dirichlet Allocation: Towards a Deeper Understanding,* 2012.

[6] T. L. Griffiths y S. Mark, «Finding scientific topics,» *PNAS,* vol. 101, pp. 5228-5235, 2004.

[7] Y. Xie y C. F. Manski, «The Logit Model and Response-Based Samples,» *Socilogical Methods & Research,* pp. 283-302, 1989.

[8] C. F. Manski y S. R. Lerman, «The Estimation of Choice Probabilities from Choice Based Samples,» *Econometrica,* vol. 45, nº 8, pp. 1977-1988, 1977.

[9] A. Scott y C. Wild, «Fitting Logistic Models under Case-Control or Choice Based Sampling,» *Royal Statistical Society,* vol. 48, nº 2, pp. 170-182, 1986.

[10] P. Zhang, «Model Selection via Multifold Cross Validation,» *The Annals of Statistics,* vol. 21, nº 1, pp. 299-313, 1993.

[11] D. Blei y L. John, «Topic Models,» *Text Mining: Classification, Clustering, and Applications,* 2009.

[12] B. Grün y K. Hornik, «topicmodels: An R package for Fitting Topic Models,» *Journal of Statistical Software,* vol. 40, nº 13, 2011.

[13] M. A. Taddy, «On Estimation and Selection for Topic Models,» pp. 1184-1193.

[14] X. Qin, «Service quality measure of B2C company based on customer complaints,» *Service Systems and Service Management (ICSSSM), 2012 9th International Conference,* pp. 41-44, 2012.

[15] Y. Xie y C. F. Manski, «The Logit Model, The Probit, and Response-Based Samples».

12. ANEXOS

Anexo 1. Sesgos de Selección para clientes catalogados como fugados.

Casos	Cantidad	% del total
Total de clientes Fugados en Agosto que dieron de baja alguno de los servicios de Hogar	18418	
Clientes que llaman al CC en Julio por teléfono fijo de la compañía y dan de baja alguno de los servicios de hogar	766	4.15%
Clientes que llaman al CC en Julio por celular de la compañía y dan de baja alguno de los servicios de hogar	268	1.45%
Clientes que llaman al CC en Julio por número de otra compañía, y dan de baja alguno de los servicios de hogar	?	-
Clientes que no llaman al CC durante Julio y dan de baja alguno de los servicios de hogar	?	-

Anexo 2. Sesgos de Selección para clientes catalogados como activos.

Casos	Cantidad	% del total
Total de clientes con al menos un servicio de hogar contratado en Agosto	1092013	
Clientes que llaman al CC en Julio por teléfono fijo de la compañía y en Agosto permanecen con los mismos servicios contratados	44787	4.10%
Clientes que llaman al CC en Julio por celular de la compañía y en Agosto permanecen con los mismos servicios contratados	12142	1.11%
Clientes que llaman al CC en Julio por número de otra compañía y en Agosto permanecen con los mismos servicios contratados	?	-
Clientes que no llaman al CC durante Julio y en Agosto siguen con los mismos servicios contratados	?	-

Anexo 3. Matrices de confusión dentro de la muestra para los modelos elegidos considerando distintas proporciones y puntos de corte.

Proporción 80/20

Punto de corte 0.5:

Modelo			Matriz de Confusión				
1	AIC	137.1			Pronosticado		
					0	1	Error
	BIC	172.74	Observado	0	29	11	27.5
				1	6	46	11.5384615
2	AIC	137.5			Pronosticado		
					0	1	Error
	BIC	170.47	Observado	0	29	11	27.5
				1	5	47	9.61538462
3	AIC	132.7			Pronosticado		
					0	1	Error
	BIC	171.13	Observado	0	31	9	22.5
				1	8	44	15.3846154
4	AIC	133			Pronosticado		
					0	1	Error
	BIC	168.64	Observado	0	32	8	20
				1	8	44	15.3846154

Punto de corte 0.7:

Modelo			Matriz de Confusión				
1	AIC	137.1			Pronosticado		
					0	1	Error
	BIC	172.74	Observado	0	35	5	9.43396226
				1	17	35	29.3103448
2	AIC	137.5			Pronosticado		
					0	1	Error
	BIC	170.47	Observado	0	35	5	9.43396226
				1	16	36	27.5862069
3	AIC	132.7			Pronosticado		
					0	1	Error
	BIC	171.13	Observado	0	35	5	9.43396226
				1	13	39	22.4137931
4	AIC	133			Pronosticado		
					0	1	Error
	BIC	168.64	Observado	0	34	6	11.5384615
				1	15	37	25.862069

<u>Proporción 90/10</u>

Punto de corte 0.5:

Modelo			Matriz de Confusión				
1	AIC	137.1		Pronosticado			
					0	1	Error
	BIC	172.74	Observado	0	35	11	23.9130435
				1	7	50	12.2807018
2	AIC	137.5		Pronosticado			
					0	1	Error
	BIC	170.47	Observado	0	35	11	23.9130435
				1	7	50	12.2807018
3	AIC	132.7		Pronosticado			
					0	1	Error
	BIC	171.13	Observado	0	35	11	23.9130435
				1	10	47	17.5438596
4	AIC	133		Pronosticado			
					0	1	Error
	BIC	168.64	Observado	0	36	10	21.7391304
				1	10	47	17.5438596

Punto de corte 0.7:

Modelo			Matriz de Confusión				
1	AIC	137.1		Pronosticado			
					0	1	Error
	BIC	172.74	Observado	0	41	5	10.8695652
				1	18	39	31.5789474
2	AIC	137.5		Pronosticado			
					0	1	Error
	BIC	170.47	Observado	0	40	6	13.0434783
				1	18	39	31.5789474
3	AIC	132.7		Pronosticado			
					0	1	Error
	BIC	171.13	Observado	0	40	6	13.0434783
				1	16	41	28.0701754
4	AIC	133		Pronosticado			
					0	1	Error
	BIC	168.64	Observado	0	40	6	13.0434783
				1	18	39	31.5789474

<u>Proporción 95/5</u>

Punto de corte 0.5:

Modelo			Matriz de Confusión			
1	AIC	137.1		Pronosticado		
				0	1	Error
	BIC	172.74	Observado 0	35	13	27.0833333
			1	9	52	14.7540984
2	AIC	137.5		Pronosticado		
				0	1	Error
	BIC	170.47	Observado 0	36	12	25
			1	8	53	13.1147541
3	AIC	132.7		Pronosticado		
				0	1	Error
	BIC	171.13	Observado 0	35	13	27.0833333
			1	10	51	16.3934426
4	AIC	133		Pronosticado		
				0	1	Error
	BIC	168.64	Observado 0	36	12	25
			1	10	51	16.3934426

Punto de corte 0.7:

Modelo			Matriz de Confusión			
1	AIC	137.1		Pronosticado		
				0	1	Error
	BIC	172.74	Observado 0	43	5	10.4166667
			1	22	39	36.0655738
2	AIC	137.5		Pronosticado		
				0	1	Error
	BIC	170.47	Observado 0	42	6	12.5
			1	21	40	34.4262295
3	AIC	132.7		Pronosticado		
				0	1	Error
	BIC	171.13	Observado 0	42	6	12.5
			1	20	41	32.7868852
4	AIC	133		Pronosticado		
				0	1	Error
	BIC	168.64	Observado 0	42	6	12.5
			1	21	40	34.4262295

Anexo 4. Curvas ROC y área bajo la curva (AUC) en muestra de entrenamiento.

Proporción 80/20

Modelos		AUC (ROC Curve)
1		
AIC	137.1	0.8832
BIC	172.74	
2		
AIC	137.5	0.8774
BIC	170.47	
3		
AIC	132.7	0.9149
BIC	171.13	
4		
AIC	133	0.9125
BIC	168.64	

Proporción 90/10

Modelos		AUC (ROC Curve)
1		
AIC	137.1	0.886
BIC	172.74	
2		
AIC	137.5	0.8825
BIC	170.47	
3		
AIC	132.7	0.9039
BIC	171.13	
4		
AIC	133	0.9008
BIC	168.64	

Proporción 95/5

Modelos		AUC (ROC Curve)
1		
AIC	137.1	0.8651
BIC	172.74	
2		
AIC	137.5	0.8644
BIC	170.47	
3		
AIC	132.7	0.8811
BIC	171.13	
4		
AIC	133	0.8801
BIC	168.64	

Anexo 5. Matrices de confusión dentro de la muestra para modelo utilizando variables que describen a las llamadas, considerando distintas particiones y puntos de corte.

Proporción 80/20

Punto de corte 0.5:

Modelo		Matriz de Confusión			
AIC	159.91		Pronosticado		
			0	1	Error
BIC	168.14	Observado 0	10	30	75
		1	5	47	9.61538462

Punto de corte 0.7:

Modelo		Matriz de Confusión			
AIC	159.91		Pronosticado		
			0	1	Error
BIC	168.14	Observado 0	40	0	0
		1	52	0	100

Proporción 90/10

Punto de corte 0.5:

Modelo		Matriz de Confusión			
AIC	159.91		Pronosticado		
			0	1	Error
BIC	168.14	Observado 0	12	34	73.9130435
		1	8	49	14.0350877

Punto de corte 0.7:

Modelo		Matriz de Confusión			
AIC	159.91		Pronosticado		
			0	1	Error
BIC	168.14	Observado 0	46	0	0
		1	57	0	100

Proporción 95/5

Punto de corte 0.5:

Modelo		Matriz de Confusión			
AIC	159.91		Pronosticado		
			0	1	Error
BIC	168.14	Observado 0	14	34	70.8333333
		1	8	53	13.1147541

Punto de corte 0.7:

Modelo		Matriz de Confusión			
AIC	159.91		Pronosticado		
		Observado	0	1	Error
BIC	168.14	0	48	0	0
		1	61	0	100

Anexo 6. Matrices de confusión dentro de la muestra para los modelos utilizando variables de las llamadas elegidas considerando distintas proporciones y puntos de corte.

Proporción 80/20

Punto de corte 0.5:

Modelo			Matriz de Confusión			
1	4 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.6	Observado 0	15	25	62.5
	BIC	157.09	1	12	40	23.0769231
2	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.5	Observado 0	15	25	62.5
	BIC	156.99	1	9	43	17.3076923
3	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	146.3	Observado 0	28	12	30
	BIC	195.72	1	8	44	15.3846154
4	6 Tópicos			Pronosticado		
				0	1	Error
	AIC	150.1	Observado 0	22	18	45
	BIC	161.03	1	18	34	34.6153846
5	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	145.1	Observado 0	24	16	40
	BIC	186.29	1	11	41	21.1538462

Punto de corte 0.7:

Modelo			Matriz de Confusión			
1	4 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.6	Observado 0	39	1	2.5
	BIC	157.09	1	33	19	63.4615385
2	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.5	Observado 0	36	4	10
	BIC	156.99	1	38	14	73.0769231
3	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	146.3	Observado 0	36	4	10
	BIC	195.72	1	21	31	40.3846154
4	6 Tópicos			Pronosticado		
				0	1	Error
	AIC	150.1	Observado 0	36	4	10
	BIC	161.03	1	36	16	69.2307692
5	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	145.1	Observado 0	36	4	10
	BIC	186.29	1	24	28	46.1538462

Proporción 90/10:

Punto de corte 0.5:

Modelo			Matriz de Confusión			
1	4 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.6	Observado 0	20	26	56.5217391
	BIC	157.09	1	15	42	26.3157895
2	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.5	Observado 0	21	25	54.3478261
	BIC	156.99	1	13	44	22.8070175
3	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	146.3	Observado 0	33	13	28.2608696
	BIC	195.72	1	10	47	17.5438596
4	6 Tópicos			Pronosticado		
				0	1	Error
	AIC	150.1	Observado 0	28	18	39.1304348
	BIC	161.03	1	19	38	33.3333333
5	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	145.1	Observado 0	29	17	36.9565217
	BIC	186.29	1	10	47	17.5438596

Punto de corte 0.7:

Modelo			Matriz de Confusión			
1	4 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.6	Observado 0	45	1	2.17391304
	BIC	157.09	1	36	21	63.1578947
2	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.5	Observado 0	43	3	6.52173913
	BIC	156.99	1	43	14	75.4385965
3	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	146.3	Observado 0	43	3	6.52173913
	BIC	195.72	1	25	32	43.8596491
4	6 Tópicos			Pronosticado		
				0	1	Error
	AIC	150.1	Observado 0	42	4	8.69565217
	BIC	161.03	1	38	19	66.6666667
5	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	145.1	Observado 0	42	4	6.77966102
	BIC	186.29	1	27	30	40.9090909

Proporción 95/5

Punto de corte 0.5:

Modelo			Matriz de Confusión			
1	4 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.6	Observado 0	19	29	60.4166667
	BIC	157.09	1	16	45	26.2295082
2	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.5	Observado 0	20	28	58.3333333
	BIC	156.99	1	12	49	19.6721311
3	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	146.3	Observado 0	32	16	33.3333333
	BIC	195.72	1	11	50	18.0327869
4	6 Tópicos			Pronosticado		
				0	1	Error
	AIC	150.1	Observado 0	28	20	41.6666667
	BIC	161.03	1	21	40	34.4262295
5	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	145.1	Observado 0	30	18	37.5
	BIC	186.29	1	10	51	16.3934426

Punto de corte 0.7:

Modelo			Matriz de Confusión			
1	4 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.6	Observado 0	46	2	4.16666667
	BIC	157.09	1	42	19	68.852459
2	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	151.5	Observado 0	45	3	6.25
	BIC	156.99	1	46	15	75.4098361
3	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	146.3	Observado 0	45	3	6.25
	BIC	195.72	1	28	33	45.9016393
4	6 Tópicos			Pronosticado		
				0	1	Error
	AIC	150.1	Observado 0	44	4	8.33333333
	BIC	161.03	1	42	19	68.852459
5	8 Tópicos			Pronosticado		
				0	1	Error
	AIC	145.1	Observado 0	44	4	8.33333333
	BIC	186.29	1	28	33	45.9016393

Anexo 7. Área bajo la curva (AUC) de la curva ROC para distintas proporciones de entrenamiento para modelos utilizando variables de llamadas.

Proporción 80/20

Modelos		AUC (ROC Curve)
Modelo 1		0.6712
AIC	151.6	
BIC	157.09	
Modelo 2		0.6654
AIC	151.5	
BIC	156.99	
Modelo 3		0.8481
AIC	146.3	
BIC	195.72	
Modelo 4		0.6712
AIC	150.1	
BIC	161.03	
Modelo 5		0.8077
AIC	145.1	
BIC	186.29	

Proporción 90/10

Modelos		AUC (ROC Curve)
Modelo 1		
AIC	151.6	0.6728
BIC	157.09	
Modelo 2		
AIC	151.5	0.6693
BIC	156.99	
Modelo 3		
AIC	146.3	0.86
BIC	195.72	
Modelo 4		
AIC	150.1	0.6968
BIC	161.03	
Modelo 5		
AIC	145.1	0.8257
BIC	186.29	

Proporción 95/5

Modelos		AUC (ROC Curve)
Modelo 1		
AIC	151.6	0.6557
BIC	157.09	
Modelo 2		
AIC	151.5	0.6626
BIC	156.99	
Modelo 3		
AIC	146.3	0.8381
BIC	195.72	
Modelo 4		
AIC	150.1	0.6745
BIC	161.03	
Modelo 5		
AIC	145.1	0.8139
BIC	186.29	

Anexo 8. Primeras 10 palabras más frecuentes en cada tópico.

Tópico 1	Tópico 2	Tópico 3	Tópico 4
Perfecto	Internet	Reclamo	Punto
Puk	Modem	Mañana	Aparece
Perdón	Conexión	Llama	Página
Información	Red	Revisar	Claro
Verificar	Computador	Dirección	Configuración
Rut	Cable	Sistema	Correcto
Indicar	Directamente	Tema	Clave
Sim	Conectado	Contrato	Ejemplo
Bloqueo	Conecta	Máximo	Ingresar
Pin	Mes	Ayer	Contraseña

Tópico 5	Tópico 6	Tópico 7	Tópico 8
Correcto	Aparece	Claro	Habla
Escucha	Señal	Tenía	Cuanto
Espera	Pantalla	Mire	Comercial
Disculpe	Sistema	Casa	Indica
Llamada	Mensaje	Dijeron	Tendría
Costo	Tarjeta	Dar	Transferir
Ningún	Televisión	Cambiar	Pesos
Amable	Decodificador	Baja	Exactamente
Inconveniente	Codificador	Pagar	Telefónico
Fija	Nuevo	Comercial	Cambio

Anexo 9. Matrices de confusión dentro de la muestra para los modelos elegidos considerando distintas proporciones y puntos de corte.

Proporción 80/20

Punto de corte 0.5:

Modelo			Matriz de Confusión			
1	AIC	126		Pronosticado		
				0	1	Error
	BIC	172.64	Observado 0	35	5	12.5
			1	6	46	11.5384615
2	AIC	131.9		Pronosticado		
				0	1	Error
	BIC	170.36	Observado 0	30	10	25
			1	7	45	13.4615385

Punto de corte 0.7:

Modelo			Matriz de Confusión			
1	AIC	126		Pronosticado		
				0	1	Error
	BIC	172.64	Observado 0	36	4	10
			1	13	39	25
2	AIC	131.9		Pronosticado		
				0	1	Error
	BIC	170.36	Observado 0	37	3	7.5
			1	13	39	25

Proporción 90/10

Punto de corte 0.5:

Modelo			Matriz de Confusión			
1	AIC	126		Pronosticado		
					0	1	Error
	BIC	172.64	Observado	0	40	6	13.0434783
				1	7	50	12.2807018
2	AIC	131.9		Pronosticado		
					0	1	Error
	BIC	170.36	Observado	0	35	11	23.9130435
				1	8	49	14.0350877

Punto de corte 0.7:

Modelo			Matriz de Confusión			
1	AIC	126		Pronosticado		
					0	1	Error
	BIC	172.64	Observado	0	42	4	8.69565217
				1	13	44	22.8070175
2	AIC	131.9		Pronosticado		
					0	1	Error
	BIC	170.36	Observado	0	40	6	13.0434783
				1	17	40	29.8245614

Proporción 95/5

Punto de corte 0.5:

Modelo			Matriz de Confusión			
1	AIC	126		Pronosticado		
					0	1	Error
	BIC	172.64	Observado	0	41	7	14.5833333
				1	10	51	16.3934426
2	AIC	131.9		Pronosticado		
					0	1	Error
	BIC	170.36	Observado	0	37	11	22.9166667
				1	10	51	16.3934426

Punto de corte 0.7:

Modelo			Matriz de Confusión			
1	AIC	126		Pronosticado		
					0	1	Error
	BIC	172.64	Observado	0	43	5	10.4166667
				1	18	43	29.5081967
2	AIC	131.9		Pronosticado		
					0	1	Error
	BIC	170.36	Observado	0	42	6	12.5
				1	22	39	36.0655738

Anexo 10. Curvas ROC y área bajo la curva en el conjunto de entrenamiento para distintas particiones.

Partición 80/20

Modelos		AUC (ROC Curve)
1		
AIC	126	0.9361
BIC	172.64	
2		
AIC	131.9	0.9111
BIC	170.36	

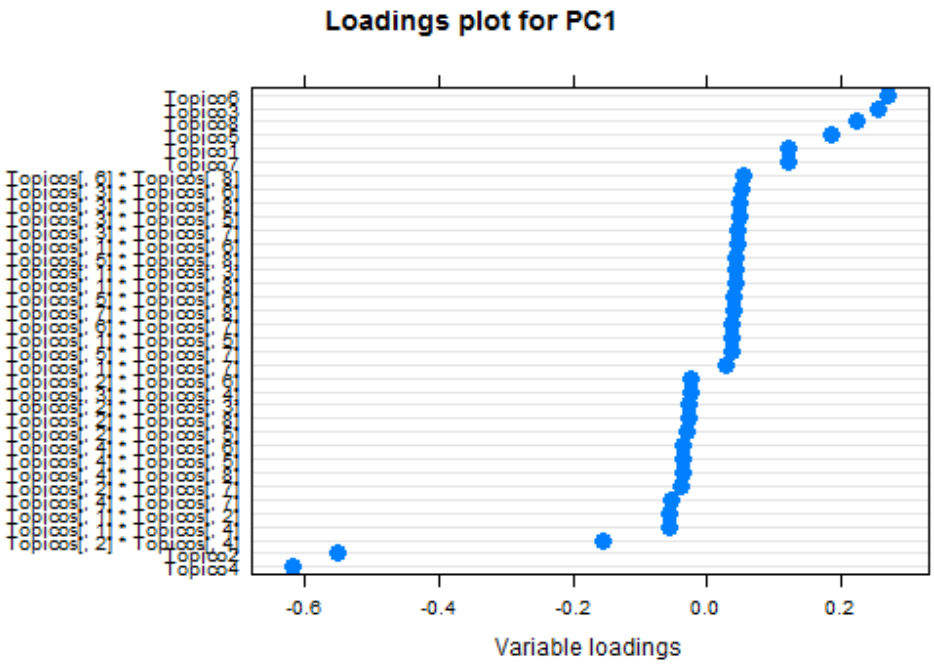
Partición 90/10

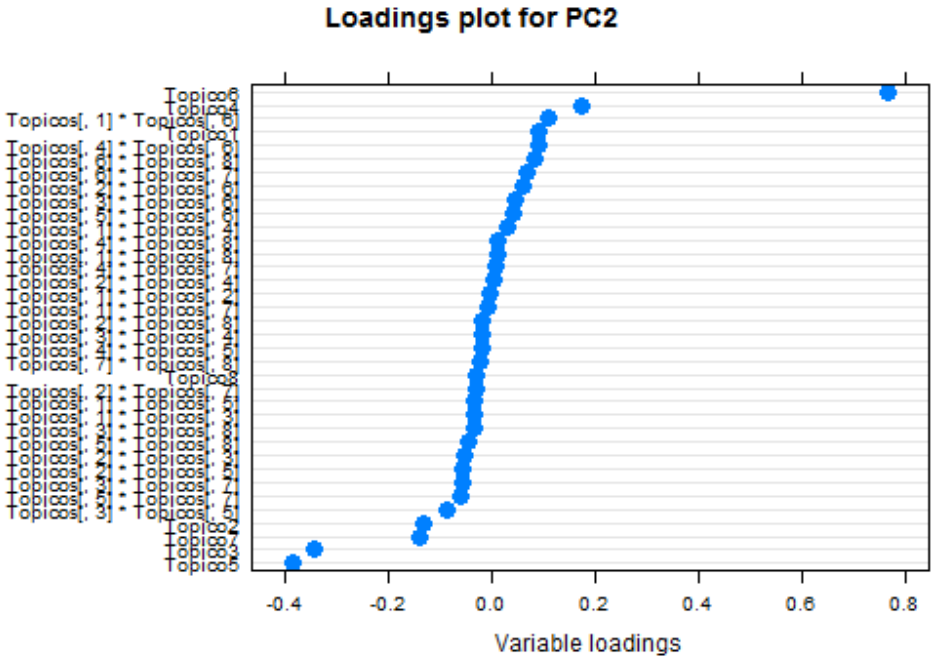
Modelos		AUC (ROC Curve)
1		
AIC	126	0.9451
BIC	172.64	
2		
AIC	131.9	0.897
BIC	170.36	

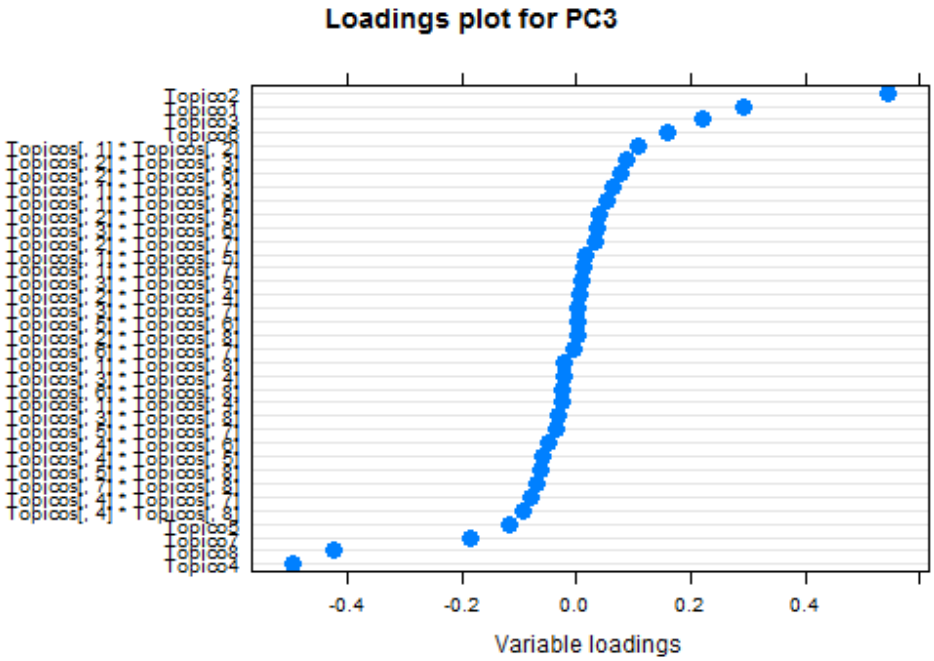
Partición 95/5

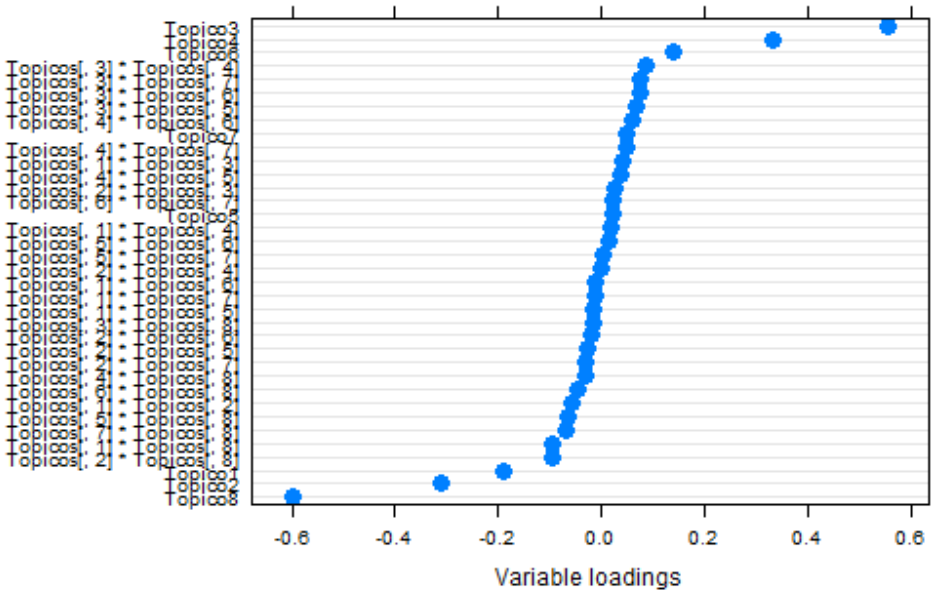
Modelos		AUC (ROC Curve)
1		
AIC	126	0.9143
BIC	172.64	
2		
AIC	131.9	0.8805
BIC	170.36	

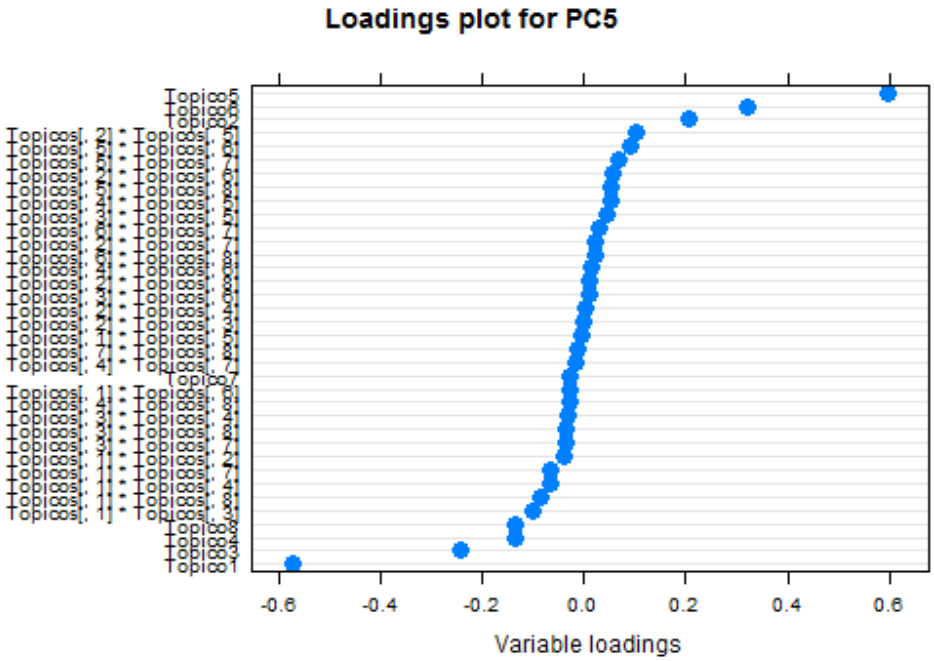
Anexo 11. Coeficientes dentro de cada componente principal.

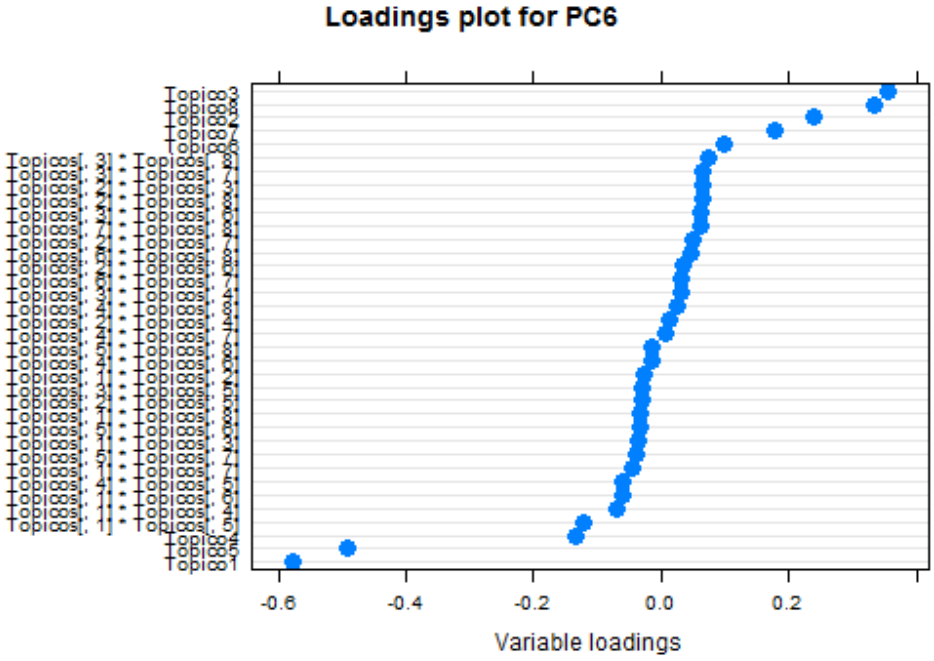





Loadings plot for PC3
Topico2
Topico1
Topico3
Topico6
Topicos[, 1] * Topicos[, 2]
Topicos[, 2] * Topicos[, 3]
Topicos[, 2] * Topicos[, 6]
Topicos[, 3] * Topicos[, 3]
Topicos[, 1] * Topicos[, 6]
Topicos[, 2] * Topicos[, 5]
Topicos[, 3] * Topicos[, 6]
Topicos[, 2] * Topicos[, 7]
Topicos[, 1] * Topicos[, 5]
Topicos[, 1] * Topicos[, 7]
Topicos[, 3] * Topicos[, 5]
Topicos[, 4] * Topicos[, 4]
Topicos[, 3] * Topicos[, 7]
Topicos[, 6] * Topicos[, 6]
Topicos[, 4] * Topicos[, 8]
Topicos[, 6] * Topicos[, 7]
Topicos[, 1] * Topicos[, 8]
Topicos[, 3] * Topicos[, 4]
Topicos[, 6] * Topicos[, 8]
Topicos[, 3] * Topicos[, 8]
Topicos[, 5] * Topicos[, 7]
Topicos[, 4] * Topicos[, 6]
Topicos[, 4] * Topicos[, 5]
Topicos[, 5] * Topicos[, 8]
Topicos[, 7] * Topicos[, 8]
Topicos[, 4] * Topicos[, 7]
Topicos[, 4] * Topicos[, 8]
Topico5
Topico7
Topico8
Topico4
-0.4
-0.2
0.0
0.2
0.4
Variable loadings


Loadings plot for PC4
Topico3
Topico4
Topico6
Topicos[, 3] * Topicos[, 4]
Topicos[, 3] * Topicos[, 7]
Topicos[, 3] * Topicos[, 6]
Topicos[, 3] * Topicos[, 5]
Topicos[, 4] * Topicos[, 6]
Topico7
Topicos[, 4] * Topicos[, 7]
Topicos[, 1] * Topicos[, 3]
Topicos[, 4] * Topicos[, 5]
Topicos[, 2] * Topicos[, 3]
Topicos[, 6] * Topicos[, 7]
Topico5
Topicos[, 1] * Topicos[, 4]
Topicos[, 5] * Topicos[, 6]
Topicos[, 5] * Topicos[, 7]
Topicos[, 2] * Topicos[, 4]
Topicos[, 1] * Topicos[, 6]
Topicos[, 1] * Topicos[, 7]
Topicos[, 1] * Topicos[, 5]
Topicos[, 3] * Topicos[, 8]
Topicos[, 2] * Topicos[, 5]
Topicos[, 5] * Topicos[, 7]
Topicos[, 4] * Topicos[, 8]
Topicos[, 6] * Topicos[, 8]
Topicos[, 1] * Topicos[, 2]
Topicos[, 5] * Topicos[, 8]
Topicos[, 7] * Topicos[, 8]
Topicos[, 1] * Topicos[, 8]
Topicos[, 2] * Topicos[, 8]
Topico1
Topico2
Topico8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
Variable loadings


Loadings plot for PC5
Topics[, 2] * Topics[, 5]
Topics[, 5] * Topics[, 8]
Topics[, 5] * Topics[, 7]
Topics[, 2] * Topics[, 6]
Topics[, 5] * Topics[, 8]
Topics[, 4] * Topics[, 5]
Topics[, 3] * Topics[, 5]
Topics[, 6] * Topics[, 7]
Topics[, 2] * Topics[, 7]
Topics[, 6] * Topics[, 8]
Topics[, 4] * Topics[, 6]
Topics[, 2] * Topics[, 8]
Topics[, 3] * Topics[, 6]
Topics[, 4] * Topics[, 4]
Topics[, 4] * Topics[, 2]
Topics[, 2] * Topics[, 4]
Topics[, 1] * Topics[, 6]
Topics[, 7] * Topics[, 8]
Topics[, 4] * Topics[, 7]
Topics[, 1] * Topics[, 6]
Topics[, 3] * Topics[, 4]
Topics[, 3] * Topics[, 8]
Topics[, 3] * Topics[, 7]
Topics[, 1] * Topics[, 2]
Topics[, 1] * Topics[, 7]
Topics[, 1] * Topics[, 4]
Topics[, 1] * Topics[, 8]
Topics[, 1] * Topics[, 5]
Topic5
Topic8
Topic6
Topic2
Topic7
Topic4
Topic3
Topic1
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
Variable loadings


Loadings plot for PC6
Topics[, 3] * Topics[, 8]
Topics[, 3] * Topics[, 7]
Topics[, 2] * Topics[, 3]
Topics[, 2] * Topics[, 8]
Topics[, 3] * Topics[, 6]
Topics[, 7] * Topics[, 8]
Topics[, 2] * Topics[, 7]
Topics[, 6] * Topics[, 8]
Topics[, 5] * Topics[, 6]
Topics[, 6] * Topics[, 7]
Topics[, 3] * Topics[, 4]
Topics[, 4] * Topics[, 8]
Topics[, 2] * Topics[, 4]
Topics[, 4] * Topics[, 7]
Topics[, 5] * Topics[, 8]
Topics[, 4] * Topics[, 5]
Topics[, 1] * Topics[, 3]
Topics[, 3] * Topics[, 5]
Topics[, 5] * Topics[, 8]
Topics[, 1] * Topics[, 8]
Topics[, 5] * Topics[, 6]
Topics[, 1] * Topics[, 2]
Topics[, 5] * Topics[, 7]
Topics[, 1] * Topics[, 7]
Topics[, 4] * Topics[, 5]
Topics[, 1] * Topics[, 6]
Topics[, 1] * Topics[, 4]
Topics[, 1] * Topics[, 5]
Topic3
Topic8
Topic2
Topic7
Topic6
Topic4
Topic5
Topic1
-0.6
-0.4
-0.2
0.0
0.2
Variable loadings

Anexo 12. Matriz de confusión en muestra de entrenamiento para modelo 4 con interacción entre tópicos.

Proporción 80/20

Punto de corte 0.5:

Modelo			Matriz de Confusión			
				Pronosticado		
4	AIC	125.6		0	1	Error
	BIC	169.48	Observado 0	34	6	15
			1	6	46	11.5384615

Punto de corte 0.7:

Modelo			Matriz de Confusión			
				Pronosticado		
4	AIC	125.6		0	1	Error
	BIC	169.48	Observado 0	35	5	12.5
			1	11	41	21.1538462

Proporción 90/10

Punto de corte 0.5:

Modelo			Matriz de Confusión			
				Pronosticado		
4	AIC	125.6		0	1	Error
	BIC	169.48	Observado 0	38	8	17.3913043
			1	8	49	14.0350877

Punto de corte 0.7:

Modelo			Matriz de Confusión			
				Pronosticado		
4	AIC	125.6		0	1	Error
	BIC	169.48	Observado 0	41	5	10.8695652
			1	15	42	26.3157895

Proporción 95/5

Punto de corte 0.5:

Modelo			Matriz de Confusión			
4	AIC	125.6		Pronosticado		
					0	1	Error
	BIC	169.48	Observado	0	39	9	18.75
				1	7	54	11.4754098

Punto de corte 0.7:

Modelo			Matriz de Confusión			
4	AIC	125.6		Pronosticado		
					0	1	Error
	BIC	169.48	Observado	0	41	7	14.5833333
				1	18	43	29.5081967