ISSN 0717-9103
ISSN Online 0718-8307
Universidad del Bío-Bío

https://doi.org/10.22320/S07179103/2017.01

DISEÑO DE UN MODELO PREDICTIVO DE FUGA DE CLIENTES UTILIZANDO ÁRBOLES DE DECISIÓN

DESIGN OF A PREDICTIVE CUSTOMER LEAKAGE MODEL USING DECISION TREES

Evelyn Francisca Contreras Morales[1,♦], Francisca Mercedes Ferreira Correa[1], Mauricio A. Valle[1]

RESUMEN

Este estudio tiene como objetivo diseñar un modelo basado en árboles de decisión, que permita predecir potenciales abandonos voluntarios de clientes de una empresa de telecomunicaciones para servicios post pago (servicio contratado de monto fijo que se paga mensualmente) de televisión digital. Los resultados del modelo permiten actuar proactivamente en la retención de clientes y mejorar los servicios prestados. Se utilizaron 23 variables predictoras que inciden en la fuga de clientes. Se utilizó una variable clase o variable dependiente, que es un identificador que determina si el cliente sigue vigente y activo en la empresa, o ha dejado de consumir servicios. Los resultados con el conjunto de datos de prueba, logran una precisión del 96,5%, indicando que los árboles de decisión resultan ser una atractiva alternativa para elaborar modelos de predicción de fuga de clientes en este tipo de datos, debido a la simplicidad de interpretación de los resultados.

Palabras Clave: Retención de clientes, árboles de decisión, telecomunicaciones, fuga de clientes.

ABSTRACT

This study aims to design a model based on decision trees, which allows predicting potential voluntary abandonment of customers of a telecommunications company for post-paid services (fixed monthly contracted service) of digital television. The results of the model allow to proactively act in the retention of clients and improve the services provided. We used 23 predictor variables that influence the leakage of clients. The dependent variable is an identifier that determines if the client is still active and active in the company, or has stopped consuming services. The results with the test dataset achieve an Accuracy of 96,5% indicating that the decision trees prove to be an attractive alternative to elaborate models of

[1] Universidad de Valparaíso, Facultad de Ingeniería, Departamento de Ingeniería Industrial Santiago de Chile, Chile. orcid.org/0000-0001-6638-069; orcid.org/0000-0003-4390-3173

♦Autor para correspondencia: ef.contreras.m@gmail.com

Recibido: 28.04.2016 Aceptado: 30.01.2017

prediction of customer leakage in this type of data due to the simplicity of Interpretation of results.

Keywords: Customer retention, decision trees, telecommunications, customer leak.

INTRODUCCIÓN

En la actualidad existe una ardua competencia en la industria de las telecomunicaciones en Chile, particularmente, en diversas empresas que ofrecen servicios de Televisión pagada. De acuerdo a datos de la Subsecretaría de Telecomunicaciones (SubTel), hasta marzo 2017 hubo un incremento de un 10,3% de suscriptores de televisión de pago con respecto a 2014 en el mismo periodo. El mercado lo constituye cinco empresas principalmente (VTR, 35%; Telefónica, 21,6%; DIRECTV, 17,5%, Claro, 15%, Entel S.A 3,1% y Otras Compañías, 7,6%) (SubTel, 2017) lo cual sugiere que la competencia es elevada por captar tantos clientes como sea posible, especialmente en un mercado en que la disponibilidad de clientes es limitada. El panorama competitivo de la industria de la TV pagada hace pensar que la administración de la retención de clientes es clave para mantener y/o mejorar la posición de mercado. Desde una perspectiva de inteligencia de negocios, el proceso de gestión de fuga de clientes se considera dentro del marco de la administración de las relaciones con el cliente, la cual se focaliza en dos actividades: La primera, en predecir una potencial fuga, y la segunda, aplicar medidas preventivas para evitar que se produzca la fuga (Hung *et al.,* 2006).

De lo anterior, el presente trabajo se enfoca en la primera actividad, es decir, determinar un modelo predictivo basado en datos históricos que permita al administrador identificar clientes con intención de renuncia de sus servicios contratados con la empresa.

La fuga de clientes se define como el “Movimiento de un cliente de un proveedor de servicio a otro”, mientras que la gestión de fuga de clientes describe el proceso por el cual el operador del servicio intenta evitar la fuga del cliente (Berson *et al.,* 2000). En este sentido, la consecución de modelos predictivos que permitan identificar clientes en situación de riesgo de fuga, permite al operador dirigir esfuerzos en prevenir la salida del cliente e iniciar un movimiento hacia otro operador.

En la entidad en la cual se realiza este estudio, existe un área que analiza distintos reportes para predecir el abandono involuntario de clientes por no pago para el mes siguiente, sin embargo no se realiza ningún diagnóstico para abordar el abandono voluntario de clientes. Debido a lo anterior, la problemática surge debido al desconocimiento de las causas de abandono voluntario de clientes de servicio tipo post pago ya que no existe un modelo, análisis o estudio que permita anticiparlo. Consecuencia de esto, la compañía actúa de manera reactiva a los escenarios que enfrenta, desaprovechando la oportunidad de retener a sus clientes y actuar bajo un proceso estandarizado de seguimiento y fidelización de los mismos. Tras la renuncia de clientes a los servicios de la organización la cartera disminuye, con esto se pierde posicionamiento en el mercado de servicios de telecomunicaciones. Esta situación, en conjunto con la competencia del mercado descrita anteriormente, pone en riesgo la sostenibilidad y posicionamiento del negocio en el largo plazo (Kim *et al*., 2004).

El problema de fuga de clientes ha sido objeto de una nutrida área de investigación, tanto desde el punto de vista del marketing, enfocándose en el comportamiento del consumidor

y los componentes emocionales que la gatillan y desde el punto de vista de la minería de datos, enfocándose en la aplicación de algoritmos de aprendizaje de máquinas para predecir el abandono de clientes (Hung *et al*., 2006). Del mismo modo, se han comparado diversas técnicas de minería de datos para elaborar un puntaje de «propensión a la fuga» de clientes de telefonía celular. Estos investigadores han reportado que los árboles de decisión y las redes neuronales son las técnicas que ofrecen la mejor precisión para la muestra de datos que utilizaron. En el mismo sentido, una desventaja de estas técnicas es que no es posible observar la cadena de condiciones que se deben cumplir para que un cliente tenga el potencial de abandonar el servicio. (Lemmens & Croux, 2006). Dicho de otra forma, se trata de modelos de caja negra, en los cuales se hace difícil entender qué es lo que gatilla o motiva una fuga. De esta forma, se limita la aplicación práctica del modelo. Cimpoeru & Andreescu (2014) han aplicado exitosamente árboles de decisión y una red bayesiana Naive para predecir la fuga de clientes de una compañía de fabricación de softwares. Ellos han reportado que los árboles de decisión logran un desempeño ligeramente superior a la red bayesiana. También se ha logrado predecir la fuga de clientes utilizando modelos de sobrevivencia, con una mejor comprensión de las variables que tienen significancia en explicar la permanencia de un cliente en la organización (Lu, 2002).

Más recientemente y en una línea similar Hung *et al.* (2006) comparan diversas modelos de clasificación, tales como árboles de decisión, redes neuronales, máquinas de soporte vectorial y k-ésimo vecino más cercano para comparar el desempeño de clasificadores en el problema de predicción de fuga de clientes. Estos autores logran establecer un modelo híbrido que logra mejorar el desempeño de predicción, particularmente el de precisión y exactitud. En este punto cabe aclarar que si bien, diversos algoritmos han probado tener buen desempeño en la predicción de fuga de clientes, muchos de ellos tienen la desventaja de ser difíciles de interpretar (Feelders, 2000). Los modelos de caja negra tales como el de redes neuronales y máquinas de soporte vectorial suelen tener desempeños en las actividades de clasificación superiores a la de sus contrapartes menos robustas como árboles de decisión, pero que son más transparentes en la interpretación de los resultados. Un árbol de clasificación permite establecer visualmente las condiciones que un cliente debe tener para entrar en una condición de abandono, mientras que una red neuronal sólo eventualmente reporta el resultado de la predicción. Este tipo de diferencias es crucial a la hora de implementar acciones preventivas que permitan evitar la fuga. En otras palabras, el modelo predictivo debe no sólo tener un desempeño adecuado en la clasificación, sino también que sea interpretable para identificar acciones en pro de la permanencia del cliente en la organización. Es necesario entender y comprender el o los factores que permiten predecir la conducta de los clientes con el propósito de satisfacer sus necesidades. Para ello, es recomendable utilizar la información disponible (base de datos) en las organizaciones a través de las herramientas de minería de datos, que extraen información con el objetivo de generar conocimiento y en particular para el desarrollo de este estudio, ayudar a predecir una serie de factores como las tendencias del abandono de clientes.

La técnica de minería de datos seleccionada para el diseño del modelo predictivo propuesto, es el árbol de decisión de tipo clasificación, debido a que es una técnica predictiva y los datos disponibles para el desarrollo del modelo corresponden a variables categóricas y discretas, que se ajustan a las características de un árbol de decisión, además por la factibilidad de interpretar la información obtenida de forma gráfica (Hernández *et al*., 2004).

Estos algoritmos de partición recursiva, pueden ser utilizados como medios para establecer un modelo de clasificación de abandono voluntario de clientes. El resultado de la aplicación del algoritmo, permite construir un árbol de decisión el cual tiene la ventaja de ser fácil de interpretar y permite de manera rápida y fácil al usuario, determinar si un cliente, dado un

ISSN 0717-9103
ISSN Online 0718-8307
Universidad del Bío-Bío

conjunto de atributos que define su comportamiento histórico, está bajo riesgo de abandonar el servicio (Cimpoeru & Andreescu, 2014)

Una condición para que el modelo sea "operacionalmente" útil a la organización, es que éste debe ser fácil de interpretar y de evaluar. De esta manera, el modelo tiene sentido práctico como herramienta de gestión dentro del contexto de la administración de relaciones con el cliente.

En tal sentido, este trabajo tiene como principal objetivo, establecer un modelo de clasificación que permita predecir las condiciones que se cumplen para el abandono voluntario de clientes, basado en antecedentes de comportamiento histórico.

Árboles de decisión

Los árboles de decisión se definen como un procedimiento recursivo, en el cual un número 'N' de instancias se dividen progresivamente en grupos, de acuerdo a una regla de división que permita maximizar la homogeneidad o pureza de la variable de respuesta o variable clase (Giudici & Figni, 2009). Los árboles de decisión pueden ser de regresión o de clasificación. En el primer caso, la variable de respuesta es continua, mientras que en el segundo caso la variable de respuesta o variable clase es discreta. Una ventaja de los árboles de decisión es su fácil interpretación, debido al modelo gráfico que se puede rescatar del resultado de la partición recursiva.

En la Figura 1 se muestra un ejemplo de árbol de decisión. El nodo raíz se encuentra por encima del árbol. Los nodos internos (nodos de decisión) corresponden a particiones sobre atributos particulares. Los arcos que emanan de un nodo corresponden a los posibles valores del atributo considerado en ese nodo. Cada arco conduce a otro nodo de decisión o a un nodo hoja. Los nodos hoja representan la predicción (clase) del problema para todas aquellas instancias que la alcanzan. Para clasificar una instancia desconocida se recorre el árbol de arriba hacia abajo de acuerdo a los valores de los atributos probados en cada nodo y cuando se llega a una hoja, la instancia se clasifica con la clase indicada por esa hoja (Hernández *et al*., 2004).

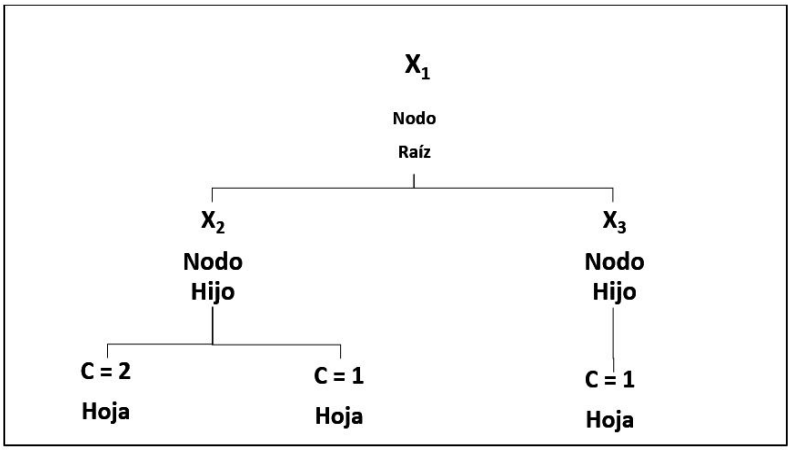

Figura 1. Ejemplo de árbol de decisión.

Al recorrer cada nodo del árbol, se llega eventualmente a las hojas, que representan el resultado final del cumplimiento de todas las condiciones y que clasifican a una instancia en

alguno de los estados de la variable clase.

La única condición que hay que exigir es que las particiones separen ejemplos en distintos hijos, con lo que la cardinalidad de los nodos va disminuyendo a medida que se desciende en el árbol (Hernández *et al.*, 2004).

ALGORITMO Partición

Input: D conjunto de N ejemplos etiquetados, cada uno de los cuales está caracterizado por n variables predictoras X_1,….., X_n y la variable clase C.

Output: Árbol de clasificación.

SI todos los ejemplos de D son de la misma clase c => Asignar la clase c al nodo N.

SALIR: # Esta rama es pura, ya no hay que seguir partiendo. N es hoja.

SI NO: -

Particiones: = generar posibles particiones.

Mejor partición:= seleccionar la mejor partición según el criterio de partición.

1. Seleccionar la variable que se adecúa mejor al criterio de partición X_r con valores x_r^1,…., x_r^{rn}.
2. Particionar D de acorde con los n_r valores de x_r en D_1,….., D_{nr}
3. Construir n_r subárboles T_1,…., T_{nr} con los valores x_r^1,……., x_r^{nr}

FIN

FIN SI

FIN ALGORITMO

Figura 2. Algoritmo de aprendizaje de árboles de decisión por “partición” (divide y vencerás).

ISSN 0717-9103
ISSN Online 0718-8307
Universidad del Bío-Bío

La construcción del árbol de decisión, se realiza a través del algoritmo de partición que se explica en la Figura 2.

Los dos puntos importantes para que el algoritmo anterior funcione correctamente, son las particiones a considerar y el criterio de selección de particiones (Hernández *et al*., 2004).

Basándose en la idea de buscar particiones que discriminen o consigan nodos más puros, se han presentado numerosos criterios de partición, tales como: Criterio de error esperado, Criterio de Gini y Entropía (Witten & Frank, 2005). Estos criterios de partición buscan la partición S que minimice la función I(s) definida de la siguiente manera (Hernández *et al*., 2004):

$$I(s) = \sum_{j=1\ldots n} Pj*(P_j^1, P_j^2, \ldots\ldots, P_j^c) \qquad \textbf{Ecuación 1}$$

En la Ecuación 1 (Ecuación general de Impureza), n es el número de nodos hijos de la partición (número de condiciones de la partición), Pj es la probabilidad de "caer" en el nodo j, Pj1 es la proporción de elementos de la clase 1 en el nodo j, Pj2 es la proporción de elementos de la clase 2 en el nodo j y así para las c clases. Bajo esta fórmula general, cada criterio de partición implementa una función f distinta, como se muestra en la Tabla 1 (Hernández et al., 2004):

Tabla 1. Función de impureza para cada criterio de partición.

Criterio	$\boldsymbol{f(P^1, P^2, \ldots., P^c)}$
Error esperado	$min(P^1, P^2, \ldots., P^c)$
GINI	$1 - \sum (P_j)^2$
Entropía	$\sum P_j * log(P_j)$

Las funciones de la Tabla 1, se denominan funciones de impureza y la función I(s), calcula la media ponderada (dependiendo de la cardinalidad de cada hijo) de la impureza de los hijos de una partición.

El criterio de partición utilizado en esta investigación corresponde al Índice de Gini, debido a que se encuentra incorporado por defecto en el algoritmo Rpart del software R (Therneau *et al.*, 2010), programa empleado para diseñar el modelo predictivo de este estudio. Este criterio compara la heterogeneidad o impureza del nodo padre con la suma de las impurezas de los nodos hijos (Ramírez *et al.*, 2009).

Las particiones son un conjunto de condiciones exhaustivas y excluyentes. Cuantas más particiones se permitan más expresivos y precisos son los árboles de decisión. No obstante, cuantas más particiones la complejidad del algoritmo es mayor. (Hernández *et al.*, 2004)

ISSN 0717-9103
ISSN Online 0718-8307
Universidad del Bío-Bío

https://doi.org/10.22320/S07179103/2017.01

METODOLOGÍA

El desarrollo del modelo de árbol de decisión se construyó a partir de 3 bases de datos de la compañía de telecomunicaciones correspondientes a los periodos de 2012 y 2013, una de ellas son las tipificaciones de llamados de los clientes (registro de los motivos por los cuales los clientes llaman), dentro de estos motivos esta la cancelación permanente de los servicios contratados, otra base de datos utilizada es la base de todos los clientes activos y cancelados desde 2012 a 2013, lo que permitió identificar el segmento de cliente, productos contratados, cantidad de productos entre otros y una tercera base de datos que corresponde al estatus del cliente hasta diciembre 2012, es decir, si presenta el servicio activo, suspendido por no pago, entre otros. Finalmente de estas tres bases de datos se consolidó en una sola donde se especifican las 23 variables predictivas y la variable clase. El procesamiento, desarrollo y prueba del modelo se realizó en 6 etapas siguiendo la metodología CRISP (Cross Industry Standard Process), debido a que es la metodología utilizada para los modelos de minería de datos, donde se permite tener una comprensión de los datos y prepararlos para el modelado, sustentando su uso en artículos como (Valle *et al.*, 2012). Esta metodología es un proceso jerárquico que consiste en un grupo de tareas descrita en niveles de abstracción (general a particular), algunas de estas fases son bidireccionales, lo que significa que algunas fases permiten revisar parcial o totalmente las fases anteriores (Goicochea, 2009). Los objetivos son:

1. Aprender nuevas técnicas para comprender y aplicar la minería de datos
2. Desarrollar proyectos de minería de datos, mediante un proceso estandarizado.

A continuación se describe en mayor detalle los pasos llevados a cabo al seguir las fases sugeridas por la metodología CRISP (Figura 3).

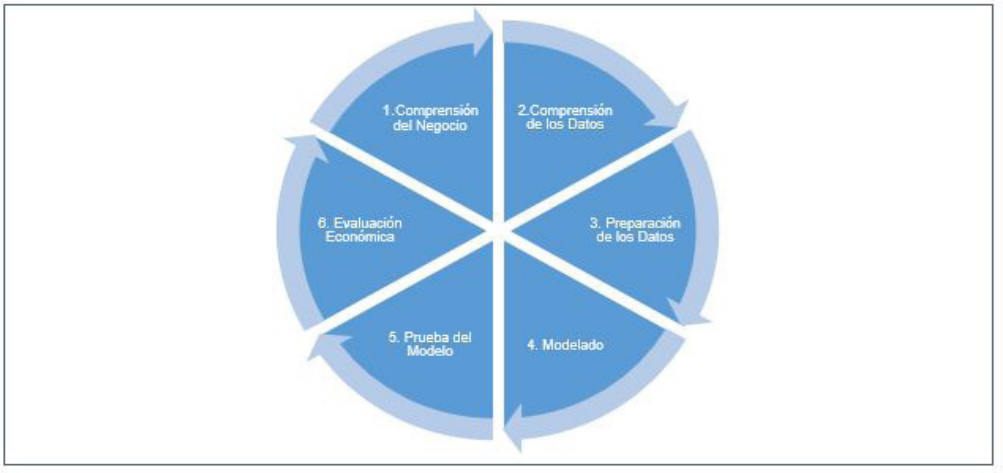

Figura 3. Fases de Metodología Crisp.

Primera Etapa: Comprensión de la situación actual de la empresa.
Para llevar a cabo esta etapa se visita las dependencias de la empresa y conocen los procesos de trabajo en cada uno de los departamentos relacionados con el área de abandono de clientes con el objetivo de comprender la situación actual de la empresa y realizar una evaluación inicial.

ISSN 0717-9103
ISSN Online 0718-8307
Universidad del Bío-Bío

Segunda Etapa: Conocimiento de los datos.
Se accede a las base de datos requeridas con el propósito de dimensionar la cantidad de clientes y servicios que ellos contratan.
Se considera una muestra de 326886 llamados de clientes desde enero de 2012 a julio de 2013.

Tercera Etapa: Preparación y Análisis de los datos.
Es necesario verificar que la muestra tenga una cantidad equilibrada de clases, es decir, que no exista una proporción muy pequeña de clientes en calidad de abandono, versus la cantidad de clientes que no han abandonado el servicio. Esto es importante porque las particiones son sensibles a la cantidad de observaciones que hay en cada una de las clases de la muestra (Giudici & Figni, 2009). Además se debe definir y filtrar las variables a utilizar que influyen en el abandono de clientes, las cuales serán utilizadas en el modelo predictivo.

Las variables definidas para el modelo son establecidas en base a distintas investigaciones que se han realizado con respecto al abandono de clientes en empresas de telecomunicaciones, que forman parte de la bibliografía de este estudio y en base a la información disponible en la compañía (Hadden *et.al*., 2006) bases de datos nombradas anteriormente Luego de agrupar ambas fuentes de información, se discute y define con los analistas de la empresa las posibles variables que pueden afectar al abandono de clientes en primera instancia (Hernández *et al.*, 2004).

Al definir las variables que afectan al abandono de clientes, en referencia a la literatura y discutida con los analistas, se procede a consolidar y crear la base de estudio. La base de datos consolidada, parte desde la base de clientes únicos (Base de datos total de clientes de la compañía), de ella se cruza con la base de datos de tipificaciones (Base de datos de motivos por los cuales llaman los clientes) con el objetivo de conocer la cantidad de veces que el cliente llamo por “X” motivo definido como variable predictora y finalmente se cruza con la base de estatus del clientes, para conocer el estado de su cuenta en el momento del estudio.

Luego, en la fase de limpieza se eliminan datos atípicos como datos incoherentes, logrando un almacén de información puro para el posterior entrenamiento del algoritmo de partición recursiva. Las observaciones incompletas o anómalas son descartadas debido a que estos casos podrían distorsionar severamente el resultado del árbol de decisión.

Definición de Variables:

Variable Clase
La variable clase corresponde a la variable objetivo que se quiere predecir. Es aquella que indica los resultados del modelo, en este caso, el abandono voluntario de clientes y se define como CV (Clase Voluntaria). Esta variable es de carácter dicotómica, de clase asociada 0 ó 1 (0: cliente permanece con sus servicios y/o 1: clientes abandona voluntariamente los servicios).

Variables Predictoras
Las variables predictivas son aquellas que permiten predecir la variable clase del modelo, en este caso es el abandono voluntario de clientes (Hernández *et al.*, 2004).
Algunas de las variables predictoras para el desarrollo del modelo, son las siguientes:

- Tipo de Cliente
- Medio de pago
- Solicitud de cancelación

- Cancelación permanente
- Estado de cuenta, entre otras.

La definición de cada una de las variables utilizadas para el modelo, están disponibles en Tabla 2 ubicada en anexo del documento.

Cuarta Etapa: Modelado.
Tras analizar cada una de las técnicas de minería de datos, se determina la técnica a trabajar para diseñar el modelo predictivo, en este caso árboles de decisión de tipo clasificación, debido a que es una técnica predictiva y los datos disponibles para el desarrollo del modelo corresponden a variables categóricas y discretas, que se ajustan a las características de un árbol de decisión, además por la facilidad de interpretar la información obtenida en forma gráfica. (Hernández *et al.*, 2004). Luego de definir las posibles variables se trabaja en base a la metodología KDD (Knowledge Discovery in Databases), específicamente en la etapa de recopilación, limpieza y transformación y finalmente exploración y selección (Hernández *et al.*, 2004). En la fase de recolección, se recopilan 3 base de datos, una de todos los clientes, otra con los llamados que realizan y la tercera con la morosidad de estos. Las dos últimas fases del proceso KDD, se realizan de forma paralela, posteriormente se seleccionan los datos influyentes en el abandono de clientes de acuerdo a las variables establecidas. (Hernández *et al.*, 2004). Finalmente se agrupan los datos seleccionados en una sola base de datos. En la fase de limpieza se eliminan datos atípicos, logrando un almacén de información puro para el posterior diseño del modelo.

Finalmente se diseña el modelo a partir de las variables predictoras, estas variables son aquellas que permiten predecir la variable objetivo del modelo, dando a conocer las condiciones que se deben cumplir para el abandono voluntario de clientes.

Quinta Etapa: Evaluación y Prueba del Modelo.
La Evaluación y Prueba del modelo diseñado, se realiza utilizando técnicas de recolección de datos, considerando una fracción de la base total de los clientes de la empresa, en este caso el 75% de la información y finalmente se prueba el modelo con la proporción restante.

Sexta Etapa: Evaluación Económica del Modelo.
En la Evaluación Económica del modelo, se compara situación actual v/s mejora a partir del modelo.

RESULTADOS

Para el Modelo de abandono voluntario de Clientes se desarrollan dos Modelos, Modelo A y Modelo B respectivamente. El primero se construye con todas las variables predictoras mencionadas en Anexo (Tabla 3).

En el segundo, se elimina la variable CP (cancelación permanente) con el propósito de construir un nuevo Modelo, ya que la variable mencionada se encuentra altamente correlacionada con la variable clase.

El Modelo se construye a partir de dos conjuntos de datos, uno para el entrenamiento y otro conjunto para la prueba del mismo. En el entrenamiento del Modelo, se utiliza el 75% del total

de datos, mientras que para la prueba, se emplea el 25% restante, estos porcentajes son utilizados de manera estándar. Siendo un total de 326886 datos (entrenamiento + prueba) con observaciones de 23 variables predictoras.

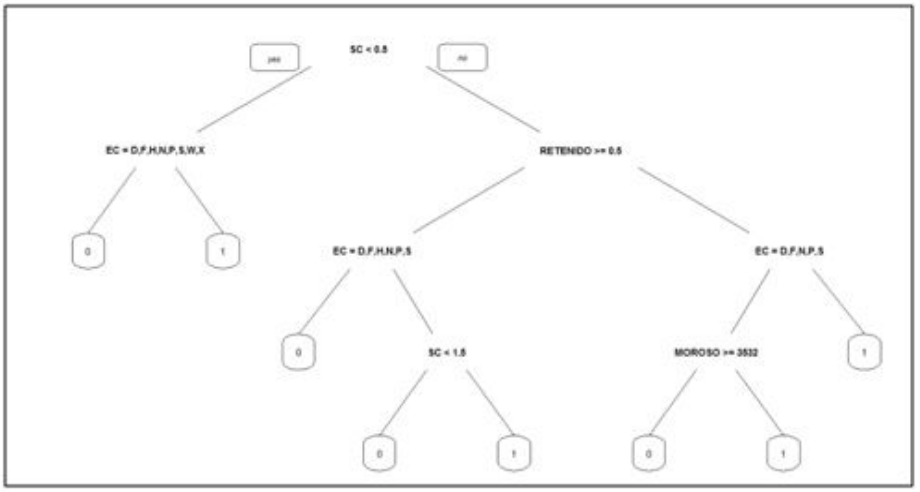

Figura 4. Modelo A

Modelo A
Luego de las simulaciones realizadas con los datos correspondientes en el software R, se obtiene el Modelo representado en la Figura 4.

El Modelo de la Figura 4, permite predecir si un Cliente va hacer abandono de los servicios de manera voluntaria, dadas las condiciones establecidas por las variables nombradas en Anexo (Tabla 3)

En concreto, el Modelo genera 8 reglas de decisión. De estas, 4 conducen a abandono voluntario de servicios, una de estas reglas se describe a continuación:

Por ejemplo, Si CP > 0,5 Λ EC ≠ D, F, N, P, S => 1 (cumple abandono voluntario de Cliente)

Es decir, si un Cliente llama para solicitar cancelar permanentemente sus servicios (CP) y tiene un estado de cuenta (EC) distinto a los mencionados en la regla de decisión, el Cliente abandona los servicios voluntariamente.

Modelo B
El Modelo de entrenamiento para el abandono voluntario de Clientes para el segundo caso, es el siguiente:

El Modelo B al igual que el anterior genera 8 reglas de decisión, 4 para el abandono voluntario de Clientes. Se diferencian en los atributos que desencadenan dichas reglas y el atributo asignado al nodo raíz, que en este caso es solicitud de cancelación (SC).
Una de las reglas de decisión para el abandono voluntario de Clientes del Modelo B, se muestra a continuación:
Si SC > 0,5 Λ RETENIDO < 0,5 Λ EC ≠ D, F, N, P, S => 1 (cumple abandono voluntario de

ISSN 0717-9103
ISSN Online 0718-8307
Universidad del Bío-Bío

https://doi.org/10.22320/S07179103/2017.01

Cliente).

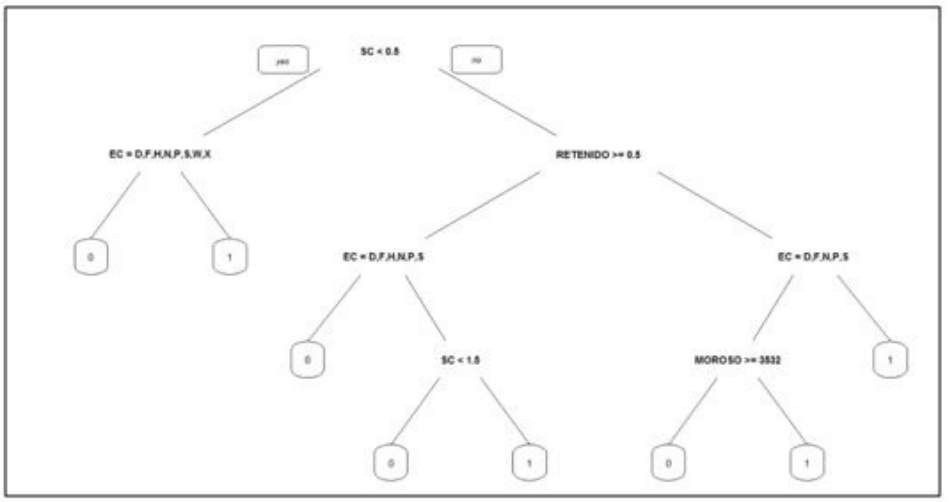

Figura 5. Modelo B

Comparación de Modelos:
Para decidir cuál de los dos Modelos anteriores es el que se utilizará para proponer políticas que permita evitar el abandono voluntario de Clientes, se hace una comparación de los distintos índices, que se pueden extraer de la matriz de confusión de cada Modelo, la cual se expone a continuación:

Tabla 2. Comparación de Modelos

Indicadores	Modelo A	Modelo B
Acierto	0,9673	0,9652
Error	0,0327	0,0348
TPR	0,886	0,884
SPC	0,989	0,973
PPV	0,788	0,768
NPV	0,989	0,988
FPR	0,024	0,026
FNR	0,114	0,115

A partir de los resultados obtenidos de la Tabla 2, se observa que no existe una gran diferencia entre los indicadores del Modelo A y B, es decir, ambos modelos son buenos predictores ya que su precisión es sobre el 75%, como lo indica el PPV. Por lo tanto el criterio de selección del Modelo más apropiado será en términos prácticos.

El Modelo A restringe las acciones preventivas para evitar el abandono de Clientes ya que la variable cancelación permanente (CP) indica que el Cliente ya está decidido a abandonar

ISSN 0717-9103
ISSN Online 0718-8307
Universidad del Bío-Bío

https://doi.org/10.22320/S07179103/2017.01

los servicios por lo que las acciones de retención sobre este no son efectivas. En cambio en el Modelo B, el campo de acción para retener a los Clientes es mayor que el del Modelo A, ya que el Cliente aún no decide cancelar permanentemente los servicios teniendo la posibilidad revertir la solicitud de cancelación que presenta.

Modelo Escogido

Tras el entrenamiento del algoritmo, se aplica el modelo entrenado sobre la muestra de prueba. El modelo entrenado y los resultados de la aplicación sobre la muestra de prueba se observa en la Figura 6.

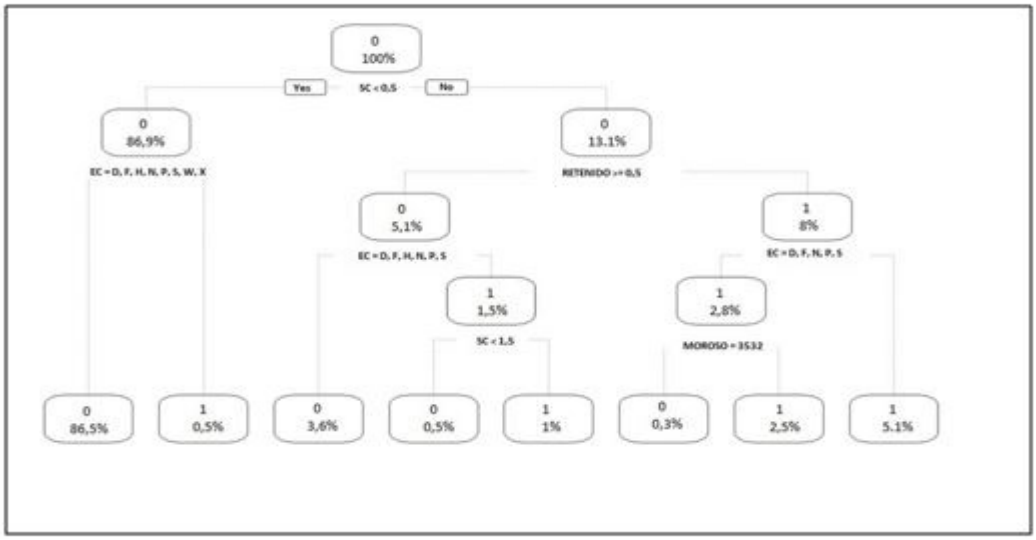

Figura 6. Árbol de decisión obtenido a partir de la muestra de entrenamiento.

El Modelo genera 8 reglas de decisión, de las cuales 4 representan el abandono voluntario de clientes.

Una de las reglas de decisión para abandono voluntario de clientes que arroja el modelo, se explica a continuación:

Si el cliente realiza una solicitud de cancelación (SC > 0,5), y el operador no logra revertir dicha cancelación (RETENIDO <0,5), y el estado de cuenta del cliente es D, F, N, P o S (EC = D, F, N, P, S) entonces el proceso de reversión de la cancelación no logra ser exitoso y el cliente abandona (CV = 1). De la muestra de prueba utilizada, esta situación representa el 5,1% de los casos.

Donde:
EC = Estado de Cuenta.
SC= Solicitud de Cancelación.
D= Desconectado.
F= Firstremainder, cliente está con primer aviso de no pago.
N= Normal, servicio activo.
P= P-Churn No Pago.
S= Secondremainder, cliente está con segundo OSD.
Se observa que el nodo raíz del árbol toma el valor de la Variable SC (Solicitud de cancelación).

Esta variable indica el número de llamados que el cliente ha efectuado para solicitar cancelar sus servicios. Por lo tanto la condición que se evalúa para la partición del nodo es SC<0,5, es decir si un cliente ha llamado para solicitar cancelar sus servicios o no. Si esta condición no se cumple (el cliente llama), la rama continúa su camino hacia abajo del nodo raíz al lado derecho del árbol y se evalúa el nodo hijo siguiente que corresponde a la variable "Retenido". En este punto, la condición que se evalúa es Retenido >= 0,5, si esta se cumple, la ramificación sigue hacia el lado izquierdo del nodo y así sucesivamente se continúa la misma lógica para cada uno de los nodos del árbol.

Por lo tanto, los resultados que generan un mayor impacto en el abandono voluntario de clientes, cumple las siguientes condiciones:
Del universo de clientes, un 13,1% de ellos solicita cancelar sus servicios De estos un 5,1% son transferidos al área de retención y el 8 % restante no está siendo transferido al área mencionada. De esta cantidad de clientes se genera un 2,5% de abandono voluntario por la rama moroso y un 5,1% por la rama EC (Estado de Cuenta), siendo un total de 7,6% de abandono voluntario de clientes, como se puede ver en la Figura 4.

Estas últimas condiciones son las que generan mayor abandono de clientes según el árbol de decisión, por lo tanto estas serán consideradas para realizar un plan de acción que permita evitar de manera proactiva el abandono voluntario de clientes.

Para determinar el desempeño del modelo anterior, se utiliza una matriz de confusión (Fawcett, 2006). Los resultados de la matriz de confusión sobre los datos de prueba se observan en la Figura 7.

		Condición Real	
		Abandono	No abandono
Predicción	Abandono	6.565 (Verdaderos Positivos)	1.980 (Falsos Positivos)
	No abandono	860 (Falsos Negativos)	72.269 (Verdaderos Negativos)

Figura 7. Matriz de Confusión.

Es posible observar que la tasa de verdaderos positivos (TPR=88,4%) y de verdaderos negativos (TNR=97,3%) son mayores que las tasas de falsos positivos (FPR=2,7%) y falsos negativos (FNT=11,6%). Esta es una condición equivalente a decir que el TPR y TNR son mayores que el error tipo I y II, respectivamente (Fawcett, 2006).
El acierto resulta ser del 96,5%, lo cual representa un muy buen desempeño del modelo para este tipo de datos, y en consecuencia un modelo confiable para predecir el abandono de clientes.

CONCLUSIÓN

Este trabajo ha propuesto establecer un modelo de predicción de abandono de clientes de una empresa de la industria chilena de TV pagada. Específicamente, se ha podido diseñar exitosamente un modelo de clasificación basado en árbol de decisión que permite clasificar a un suscriptor como un cliente en potencial abandono voluntario.

Las variables predictoras del modelo de clasificación permitieron establecer reglas de decisión desde las cuales es posible encontrar qué condiciones se debieran cumplir para que un cliente decida voluntariamente abandonar la empresa.

El modelo de árbol de decisión permite explicar en forma satisfactoria las condiciones que se deben cumplir para que ocurra un abandono voluntario de clientes. Esta información permite a la compañía actuar de manera proactiva ante este escenario y así evitar la disminución de la cartera de clientes. En este sentido, los resultados de este trabajo parecen prometedores. Por un lado, la simplicidad de interpretación de los resultados utilizando un árbol de decisión hace factible la gestión de fuga de clientes utilizando este tipo de modelos de clasificación. Esta es una condición importante para la aceptación de este tipo de herramientas en la gestión. Por otro lado, se observa que los resultados del modelo en cuanto a desempeño son satisfactorios, lo cual sugiere que las variables utilizadas tienen suficiente poder predictivo para este tipo de problemas.
Las reglas que emanan del árbol de decisión han mostrado que constituyen un reflejo del comportamiento de los suscriptores en relación a su decisión de seguir como clientes de la empresa o dejar de serlo, aún incluso, sin tener información sociodemográfica de los clientes. Esto implica que la construcción de modelos de fuga de clientes es posible a partir de datos transaccionales y operacionales.

Finalmente, los resultados del modelo muestran una falta de eficiencia en la operación misma de la retención. En otras palabras, el modelo deja en evidencia que existe una oportunidad de mejora en el área de retención. Aunque este resultado no era esperado, se pone de manifiesto que este tipo de modelos tienen el potencial adicional de mejorar procesos operativos de la propia empresa.

REFERENCIAS

AHN, J.H., HAN, S.P. and LEE, Y.S. Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications policy* [En línea]. 2006, **30** (10), 552-568. [cit. 12 de febrero de 2016]. DOI: https://doi.org/10.1016/j.telpol.2006.09.006. Disponible en: http://www.sciencedirect.com/science/article/pii/S0308596106000760

GOICOCHEA, A. CRISP-DM, Una metodología para proyectos de Minería de Datos. 2009 [En línea]. [Citado el: 12 de febrero de 2016.]. Disponible en: https://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de-datos/

BERSON, A., SMITH, S.Y and THEARLING, K. *Building data mining applications for CRM.* New York: McGraw-Hill, 2000.

CIMPOERU,C and ANDREESCU,A. Predicting Customers Churn in a Relational Database.

Informatica Economica [En línea] 2014, **18**(3), 5-16. [cit. 12 de febrero de 2016]. Disponible en:https://search.proquest.com/openview/2b1c7bbe95f3543cd613111094959dce/1?pq-origsite=gscholar&cbl=55108

FAWCETT, T. An introduction to ROC Analysis. *Pattern Recognition* [En línea]. 2006. **27**(8), 861-874. [cit. 12 de febrero de 2016]. DOI: https://doi.org/10.1016/j.patrec.2005.10.010 Disponible en: http://www.sciencedirect.com/science/article/pii/S016786550500303X

FEELDERS, A., DANIELS, H and HOLSHEIMER. M. Methodological and practical aspects of data mining. *Information & Management.* [En línea]. 2000, **37**(5), 271-281. [cit. 12 de febrero de 2016]. DOI: https://doi.org/10.1016/S0378-7206(99)00051-8 http://www.sciencedirect.com/science/article/pii/S0378720699000518

GIUDICI, P AND FIGNI, S. *Applied Data Mining for Business and Industry.* s.l. : Wiley, DOI: https://doi.org/10.1002/9780470745830.index , 2009.

HADDEN, J., TIWARI, A., ROY, R and RUTA, D. Churn Prediction: Does Technology Matter? *International Journal of Intelligent Technology* [En línea]. 2006,1(2), 104-110. [Citado el: 05 de febrero de 2016.] Disponible: https://www.waset.org/Publications/churn-prediction-does-technology-matter-/1793

HERNÁNDEZ O, J., RAMIREZ QUINTANA, MJ., FERRI RAMIREZ, C. *Introducción a la Minería de Datos*. Madrid : Pearson Prentice Hall, 2004.

HUNG, SY., YEN, D.C., WANG, H.UY. Applying data mining to telecom churn management. *Expert Systems with Applications.* [En línea]. 2006. **31**(3), 515-524. [cit. 12 de febrero de 2016]. DOI: https://doi.org/10.1016/j.eswa.2005.09.080 Disponible en: http://www.sciencedirect.com/science/article/pii/S0957417405002654

ISSN 0717-9103
ISSN Online 0718-8307
Universidad del Bío-Bío

Anexos

Tabla 3. Variables utilizadas para el Modelo

	Simbología		**Variables Predictoras**	**Tipo de Variable**	**Valores que puede tomar la Variable**
1.	TC		Tipo de Cliente	Discreta	A, C, D, E, G, H, J, K, N, Q, R, S, T, U, V, W, X, Y, Z, ESC, INCO
2.	MP		Medio de Pago	Discreta	F, A, C, D, G, H, I, J, K, L, M, N, O, P, Q, R, S,T, U, V, W, X, Y, Z
3.	SC		Solicitud de Cancelación	Discreta	Nº de veces que el cliente hace un llamado para cancelar los servicios.
4.	EC		Estado de Cuenta	Nominal	D, F, H, N, O, P, Q, R, S, T, W, X
5.	STL		Servicio Técnico el línea	Discreta	Nº de veces que el cliente hace un llamado para solicitar información de conexión de cables, configuración de control remoto, problemas con la señal, etc.
6.	WO		Workorder (Orden de Trabajo)	Discreta	Nº de veces que el cliente hace un llamado para generar solicitudes de visitas técnicas por problemas técnicos, reclamos por alguna visita técnica mal hecha o no realizada, etc.
7.	VC		Ventas y cancelación	Discreta	Nº de veces que el cliente hace un llamado para activar o cancelar canales, cambiar de plan, entre otros.
8.	PR		Proceso de retención	Nominal	Nº de veces que el cliente recibe un llamado para aceptar o rechazar las ofertas realizadas por la compañía para mantenerlos dentro de esta.
9.	CT		Cobertura Técnica	Discreta	Nº de veces que el cliente hace un llamado con respecto a la a la señal de televisión o problemas de esta índole.
10.	CR		Consulta/ Reclama	Discreta	Nº de veces que el cliente hace un llamado para reclamar o consultar cualquier consulta con respecto a los servicios prestados.
11.	RF		Reclamos Financieros	Discreta	Nº de veces que el cliente hace un llamado para reclamar por devolución saldo a favor, ajuste cargo facturado, reclama devolución, entre otros.
12.	RS		Retención del Servicio	Discreta	Nº de veces que un cliente ha sido retenido, luego de solicitar cancelar los servicios.
13.	ST		Servicio Técnico	Discreta	Nº de veces que un cliente llama para solicitar una asesoría técnica acerca de su servicio en general.
14.	REGIÓN		Región	Nominal	ZN, ZS, ZC, RM

15.	PRODUCTO		Tipo de producto	Discreta	PA, PB, PC, PD, PE, PF, PG
16.	CP		Cancelación Permanente	Discreta	Nº de veces que el cliente hace un llamado para cancelar definitivamente los servicios.
17.	MOROSO		Moroso	Continua	Cantidad de dinero que debe el cliente a julio de 2013.
18.	MORA		Mora	Discreta	Cantidad de tiempo de deuda que tiene el cliente a julio de 2013.
19.	COBRANZA		Cobranza	Discreta	Nº de veces que el cliente hace un llamado con respecto a acreditación de pago, reclamos por envío de cobranza, etc.
20.	FACTURACION		Facturación	Discreta	Nº de veces que el cliente hace un llamado para consultar saldo, explicación del cargo mensual, reclamos de aumento de precio, entre otros.
21.	RETENIDO		Retenido	Discreta	Nº de veces que el cliente hace o recibe un llamado para aceptar beneficios ofrecidos por la compañía para mantenerse dentro de esta, debido a una previa solicitud de cancelación de servicio por parte del cliente.
23.	RECLAMOS		Tipo de Reclamos	Discreta	Nº de veces que el cliente hace un llamado para reclamar por el servicio en general.