1- ¿Si Ud. tuviera que implementar un sistema de retención de clientes, considera que Data Mining puede ser de ayuda en la detección de clientes próximos a abandonar la compañía?

En el presente informe, se propone la aplicación del proceso de KDD (Knowledge Discovery in Databases) con el objetivo de detectar la fuga de clientes en una empresa del sector de telecomunicaciones. Si bien se abordarán las etapas de preprocesamiento y transformación de datos, haremos hincapié en la fase de modelado (Data Mining o análisis de datos) y en la evaluación de los modelos resultantes.

Como trasfondo de este informe subyace la idea de que, partiendo de los datos existentes, es posible detectar patrones de comportamiento con ciertos niveles de exactitud y que esta información constituye evidencia suficiente para orientar la toma de decisiones de la organización en relación a la retención de clientes.

2- ¿Qué tipo de tarea o tareas de *Data Mining* se podrían aplicar?

Una vez realizado el análisis exploratorio y la consecuente limpieza y selección de datos, se han propuesto dos subconjuntos del dataset depurado. A uno de ellos (que representa aproximadamente un 90% del total de los datos y al que llamamos *“subscriptores\_training”*) se lo utilizó para entrenar los modelos, mientras que al segundo, que representa el resto del dataset original y llamamos *“subscriptores\_test”,* se lo ha empleado en la fase de evaluación.

Entre las distintas alternativas de modelos, hemos optado por implementar árboles de decisión. Consideramos que esta es una alternativa factible y eficaz en la detección temprana de las posibles fugas de la empresa.

Con la intención de cumplir los objetivos planteados, esta vez priorizamos un árbol no sólo de capacidad predictiva relevante sino también, uno de baja complejidad: el modelo obtenido debe ser fácilmente interpretado y comunicado ya que, a partir de este, se procura predecir, sin necesidad de aplicar el modelo, el comportamiento de los clientes en función de los atributos que el modelo señale como de mayor relevancia.

Para seleccionar un árbol conveniente, una de las tareas fundamentales es la de construir y comparar una serie de modelos posibles. Estos se distinguen por los valores de sus parámetros y por los resultados que arrojan.

Específicamente, para llevar esto a cabo se ha propuesto generar, a partir de un ciclo iterativo, un conjunto de árboles vinculados a valores que van de 0.05 a 0.95 para el parámetro *Confidence Factor* y de xx a xx para *Minimum object.*

Una vez obtenidos los resultados asociados a cada modelo, se los ha almacenado para facilitar la interpretación y evaluación de las distintas alternativas en base a la información que hemos considerado de mayor interés. En particular, hemos priorizado los índices de *Accuracy*, *Recall*[[1]](#footnote-24118)y *Size*.

Cabe destacar que debido a que los resultados de cada modelo están mediados por las tareas vinculadas a la selección de variables y a su aporte a la predicción de *churners*; con el objetivo principal de incrementar los valores en los índices de *Recall* y *Accuracy*, como también el de disminuir la complejidad del modelo, se ha realizado este procedimiento con tres datasets diferentes.

3- ¿Qué tipo de datos requeriría para poder aplicar las tareas de Data Mining seleccionadas?

Al hablar de clasificación nos referimos a aquellas tareas orientadas al agrupamiento de datos bajo la categoría o clase a la que pertenecen. En este sentido, la construcción de un modelo predictivo de clasificación (?) sólo es posible si se cuenta con un conjunto finito de observaciones que, a su vez, contenga valores asociados a una variable indicadora u objetivo.

En este caso, contamos con un clasificador binario denominado “*Churn*”, cuyos valores lógicos indican si el cliente ha dado de baja o no los servicios de la empresa.

Además de esta variable, el dataset original cuenta con un conjunto de 20 variables sobre 3099 clientes. De este se eliminaron los valores únicos para cada registro (“*Phone number*”) y las variables fuertemente correlacionadas con alguna otra. Tales son los casos de las variables referidas a los cargos monetarios que redundan en información sobre los minutos diurnos, vespertinos, nocturnos e internacionales.

Por otro lado, se conservó el Estado (u provincia) del cliente (variable cualitativa convertida en R a factor); su código de área (numérica convertida a factor de 3 niveles); antigüedad (variable numérica continua); los servicios a los que suscribe y la cantidad de mensajes de voz (ambas cualitativas dicotómicas). También se han conservado los datos asociados a las llamadas realizadas a Servicio al cliente (variable cuantitativa discreta) y los minutos y cantidad de llamadas para cada una de las 4 categorías mencionadas -día, tarde, noche e internacional-, todas ellas variables cuantitativas continuas.

A pesar de que estas han sido las variables a partir de las que se procedió a modelar los árboles propuestos, hubiese sido de interés contar con información acerca del momento en el que se realizaron los consumos.

A partir de esta información hubiese sido posible agregar una nueva variable al dataset original: una que describa la variación en el consumo de minutos en el tiempo (si se ha sostenido de manera regular, ha crecido o disminuido) con el fin detectar si -en qué medida, y en qué ventana temporal- los *churners* alteran su comportamiento previo a la baja.

Asimismo, debido a que estas diferencias podrían producirse debido a la estacionalidad, de construirse una variable como la anteriormente mencionada, habría que considerar el hecho de que el volumen de minutos podría verse afectado en los meses de receso laboral/escolar, durante las festividades, etc. En este sentido, una variación en el consumo constituiría un patrón común entre los que abandonan y permanecen. Como queda en evidencia, a mayor cantidad de variables, mayor complejidad del problema y mayores las consideraciones a tener en cuenta.

Otro trabajo interesante sería el de integrar otras bases de datos referidas a los mismos registros y procedentes de la misma empresa. Estas podrían aportar, por ejemplo, datos transaccionales (clientes morosos, sus métodos de pago y su relación con las clases a predecir) o también información acerca de la asociación de la línea telefónica con un grupo familiar o corporativo.

En este último caso, se podría evaluar si los clientes que forman parte de un plan familiar o corporativo tienen o no mayores probabilidades de abandonar la compañía.

En esta dirección, los esfuerzos deberían dirigirse a la construcción de un conjunto integrado de datos, una suerte de reservorio que sea capaz de explotar la mayor cantidad de datos posible, todo con el objetivo de mejorar la capacidad predictiva de los modelos a generar.

1. Asumiendo que las medidas que deben ser tomadas al detectar a un posible cliente en fuga no representan grandes costos (posiblemente se trate de establecer contacto con el cliente y ofrecerle mejoras en el servicio o algo semejante), detectar los falsos positivos (que afectan el valor de *Precision*) tiene menor importancia (costo) que el hecho de que efectivamente sea un cliente que desea abandonar la compañía y no sea detectado como tal (que sea un falso negativo, vinculado al índice de *Recall*). [↑](#footnote-ref-24118)