MÁSTER OFICIAL EN EMPRESA Y TECNOLOGÍAS DE LA INFORMACIÓN

CURSO 2019-2020

TRABAJO FIN DE MÁSTER

TÉCNICAS DE MACHINE LEARNING EN EL ANÁLISIS DEL CHURN RATE

MACHINE LEARNING TECHNIQUES IN CHURN RATE ANALYSIS

AUTOR: DIEGO GUTIÉRREZ GONZÁLEZ

TUTOR: FRANCISCO JAVIER LENA ACEBO

Índice

Resumen

Este trabajo tiene como objetivo ofrecer una visión simplificada de la importancia del estudio del Churn Rate por parte de las empresas. A su vez, se explican diferentes técnicas de Data Mining que sirven para medir esta tasa y aportar información de cómo reducirla. La motivación de este análisis viene dada por la necesidad de conocer cuáles son las técnicas más utilizadas en su medición y a la vez averiguar cuáles son más efectivas en diferentes escenarios. Se comienza definiendo el concepto de Churn Rate y su importancia para después continuar con la definición de Data Mining. Mas adelante se explican cuatro métodos que sirven para calcular esta tasa aportando ejemplos prácticos de su efectividad. Todo esto se apoyará en una revisión bibliográfica de diferentes estudios relacionados con este tema. El objetivo de esta revisión es descubrir cuales son los métodos más utilizados en el análisis del Churn Rate en los últimos cinco años. Por otro lado, también se busca encontrar cuales son los más eficaces para este cálculo. Como resultado de este análisis se puede concluir que las técnicas de Data Mining más utilizadas en los últimos años son el Support Vector Machine y las Redes Neuronales Artificiales. Estas dos técnicas son las más relacionadas con la inteligencia artificial ya que se busca crear modelo de aprendizaje automatizado para obtener mejores resultados. Las Regresiones y los Árboles de decisión son técnicas menos usadas en el campo objeto de estudio de este trabajo pero que ofrecen unos resultados más precisos, al menos a corto plazo, quizás debidos a su mayor sencillez de aplicación. El tamaño de la muestra utilizada para el análisis también es importante ya que a mayor tamaño menor precisión, pero más posibilidades de desarrollar un modelo de aprendizaje automatizado que de mejores resultados a largo plazo.

Abstract

This work aims to provide a simplified view of the importance of the study of the Churn Rate by companies. At the same time, different techniques of Data Mining are explained that serve to measure this rate and to contribute information of how to reduce it. The motivation of this analysis is given by the need to know which are the most used techniques in their measurement and at the same time find out which are more effective in different scenarios. It begins by defining the concept of Churn Rate and its importance and then continue with the definition of Data Mining. Later on, four methods are explained that serve to calculate this rate providing practical examples of its effectiveness. All this will be supported by a bibliographic review of different studies related to this topic. The objective of this review is to discover which are the most used methods in the analysis of the Churn Rate in the last five years. On the other hand, it also seeks to find which are the most effective for this calculation. As a result of this analysis it can be concluded that the most used Data Mining techniques in recent years are the Support Vector Machine and Artificial Neural Networks. These two techniques are the most related to artificial intelligence as it seeks to create automated learning model for better results. Regressions and Decision Trees are less used techniques in this field, but they offer more precise results, at least in the short term, perhaps due to their simplicity of application. The size of the sample used for analysis is also important because the larger the sample, the lower the accuracy, but the more likely it is to develop an automated learning model that will yield better long-term results.

1. Introducción

El Churn Rate o tasa de abandono de clientes es el gran enemigo de numerosas empresas especialmente en el sector de las telecomunicaciones y las empresas SaaS. Captar a un cliente nuevo es mucho más costoso que retener a clientes que ya tienen su confianza depositada en ti, por lo que tener controlada esta tasa es clave para el correcto desarrollo de todos los negocios. Empresas proveedoras de soluciones de CRM y servicios de gestión de clientes, estiman que la mayoría de los ingresos de una empresa de estos sectores provienen de las renovaciones de sus clientes y no de la venta nueva. Entre el 70 % y el 95% de los ingresos provienen de renovaciones mientras que solo entre el 5% y el 30% provienen de la venta inicial de los productos (Nirpaz 2018).

En la actualidad, las empresas cada vez se preocupan más por mantener bajas tasas de abandono de clientes ya que esto les aporta una mejora competitiva considerable y aumenta sus tasas de ahorro. Esto solo se puede conseguir si se es capaz de predecir con precisión la rotación de clientes, lo que permite actuar sobre ella de manera eficaz.

En este aspecto, la misión de las empresas es realizar una correcta clasificación de los posibles churners (clientes que se dan de baja) antes de que aparezcan a la vez que se obtiene información de los motivos que llevan a su aparición o desarrollo. Un churner, en la mayoría de los casos, no es un cliente que tome la decisión de abandonar la compañía de un día para otro, sino que es una decisión que toma tiempo. Durante este tiempo, el cliente pueda dar señales que permitan averiguar que se va a marchar, dando lugar a posibles estrategias a realizar para evitar su perdida. Esto se puede lograr gracias a técnicas de Data Mining e Inteligencia Artificial que permiten controlar el Churn Rate y obtener las causas más relevantes de su aparición para poder actuar contra ellas.

El objetivo de este trabajo es tratar de explicar en qué consiste el Churn Rate (tasa de abandono de clientes), el problema que plantea y las posibles maneras de detectarlo y reducirlo. Para ello se explicará a fondo en que consiste y algunas de las técnicas de Data Mining existentes para su estudio. También se realizará una revisión bibliográfica de varios estudios relacionados con el tema que nos permitirá ver la importancia real que este despierta y los métodos más comunes o efectivos para medirlo.

La estructura de este trabajo consiste, en primer lugar, de una introducción en la que se explica brevemente el Churn Rate y se muestras los objetivos del estudio. A continuación, se profundiza más en el concepto y problema que supone esta tasa de abandono de clientes y sus efectos en las empresas actuales. Después se detalla el concepto de Data Mining como un método que engloba varias técnicas para su predicción u obtención. Se detallan tres de las principales técnicas de Data Mining existentes con el fin de hacernos a la idea de en qué consisten y cómo se aplican. En el siguiente apartado se realiza una revisión bibliográfica de varios estudios relevantes sobre la tasa de abandono de clientes y cómo detectarla y combatirla. Por último, se expondrán las conclusiones obtenidas de este trabajo.

2. Concepto de Churn Rate

El término Churn Rate hace referencia a la tasa de cancelación de clientes, es decir, se encarga de medir el porcentaje de clientes que se dan de baja de una empresa en un periodo de tiempo determinado. Es una métrica indispensable a la hora de conocer el motivo de la pérdida de clientes y gracias a él, poder elaborar estrategias de marketing que te permitan obtener una mayor fidelización de clientes.

> *"El Churn de clientes se mide directamente en cuánto tiempo permanece un cliente en una empresa y, a su vez, indica el valor de vida del cliente (CLV) para esa empresa" Neslin et al. (2006).*

El problema del abandono de clientes es una constante en sectores en los que los clientes se tienen que suscribir o abonar a un determinado servicio como por ejemplo el sector de las telecomunicaciones. La insatisfacción del cliente o la alta rivalidad de la competencia se pueden presentar como dos de las principales causas de esta pérdida de clientes. Luchar contra este problema se antoja clave para poder lograr un servicio beneficioso para la compañía. Obtener un cliente nuevo puede ser mucho más complicado que retener a los actuales por lo que un alto abandono de clientes se convierte en un problema importante. Conseguir la fidelidad de los clientes tiene que ser un objetivo para todas las empresas de este tipo de sectores.

> *"Retener a un cliente resulta aproximadamente diez veces más barato que conseguir uno nuevo, por eso debe ser una prioridad saber aplicar estrategias de retención y fidelización que consigan mantener y desarrollar a los clientes rentables y fieles"* (Domínguez y Hermo 2008).

Existen diversos métodos para calcular el Churn Rate de un determinado negocio, aunque si queremos comenzar por la forma más genérica de hacerlo podemos decir que se trata del número de clientes perdidos en un determinado tiempo dividido entre los clientes existentes al comienzo de ese periodo. Por lo tanto, si perdemos 10 clientes y al inicio teníamos 100 obtenemos un Churn Rate del 10%. Pero, y ahora ¿cuál es el porcentaje optimo que se debe permitir? Lo cierto es que este porcentaje no es una cifra concreta debido a que cada negocio es muy diferente de cualquier otro y existe una gran diferencia entre sectores.

Recientemente, según ha ido ganando importancia para las empresas mantener unos niveles de cancelación de clientes bajos, se han ido realizando muchos estudios relacionados con la tasa optima a conseguir. El resultado de estos análisis debe ser meramente orientativo, aunque te permite obtener una tasa de referencia del sector con la que empezar a trabajar. A modo de ejemplo, un análisis realizado por Wordstream (Herramienta que ofrece soluciones de marketing digital a empresas) estableció unos niveles de Churn Rate aproximados para diferentes sectores (Shewan 2019). En el caso de las compañías de telefonía móvil se obtiene entre un 20% y un 38% de abandono de clientes mientras que las empresas SaaS generan solo entre un 5% y un 7%.

Analizando todos estos datos comentados anteriormente podemos observar que una tasa adecuada para un tipo de empresa puede ser desastrosa para otra. Esto nos indica que el objetivo es reducir el Churn Rate en la medida que se pueda sirviéndose de apoyo de los datos existentes, pero si tomarlos como unos límites específicos.

En los mercados altamente competitivos, como es el caso de las telecomunicaciones, cada día está ganando más importancia una estrategia de marketing defensivo centrada en la retención de clientes. En lugar de intentar atraer a nuevos clientes o atraer a los

suscriptores de los competidores, el marketing defensivo se ocupa de reducir la salida de los clientes. Por lo tanto, para tener éxito en mercados maduros y altamente competitivos, el enfoque estratégico de una empresa debe ser pasar de la adquisición de clientes a la retención de clientes mediante la reducción de la rotación de clientes. Para gestionar mejor la rotación de clientes, las empresas deben comprender plenamente el comportamiento de los clientes y los factores relacionados con el Churn Rate de su negocio.

Para conseguir que una empresa crezca es fundamental retener a sus clientes y conseguir que vuelvan a consumir, cuantas más veces mejor. Obviamente, retener clientes tiene un coste elevado por lo que hay que invertir en diferentes factores de tu negocio, pero éste es mucho menor que el coste de conseguir nuevos clientes.

Actualmente existen muchas estrategias de marketing defensivo para proteger a tus clientes. El punto de partida para retener a los consumidores es conocerlos, en la mayoría de las empresas es probable que entre un 50% y un 75% de la facturación se obtenga de un grupo reducido de clientes. Por otro lado, el porcentaje restante es ocupado por una cantidad muy elevada de clientes.

Gastar recursos en conocer a tus clientes es una tarea que merece la pena, una táctica para segmentar los clientes de manera efectiva es el método RFM el cual los clasifica en función de tres variables: Recencia, Frecuencia y Money.

La recencia hace referencia al tiempo que ha transcurrido desde que un cliente ha realizado su última compra. Frecuencia indica el promedio de compras durante un periodo de tiempo (semanales, mensuales…). Por último, Money se refiere al valor de las compras que un cliente ha realizado durante el periodo de tiempo determinado. Cada cliente recibirá una puntuación en cada uno de estos criterios que normalmente va del 1 al 5 hasta clasificarle en un determinado grupo. Este método es especialmente útil para empresas que comercializan productos y servicios ya que les permite orientar sus productos a determinados grupos de consumidores.

En el caso de las empresas SaaS, las cuales están más preocupadas en la retención de sus clientes ya que presentan tasas de retención más bajas, estas deberían plantearse preguntas como cada cuánto tiempo hacen login sus usuarios o qué funcionalidades son las más utilizadas.

Una vez que es conocido el comportamiento de los clientes será más fácil realizar estrategias efectivas de retención. Personalizar los mensajes y las ofertas resultara mucho más sencillo al tener segmentados a tus clientes. Al realizar campañas personalizadas se conseguirá lograr una mayor retención de clientes por segmento. Proporcionar una atención al cliente de calidad también fomenta que los clientes se sientan a gusto con el servicio y permanezcan en la empresa. Otro punto importante es preguntar a los clientes por sus necesidades para así poder ofrecer características que se adapten a ellas y generar una sensación de que son escuchados por la compañía. Esto último es algo muy valorado por los clientes recientemente debido a la gran interacción existente a través de redes sociales entre el público y las marcas.

2.1 Importancia del Churn Rate:

Ahora que ya sabemos qué es el Churn Rate hay que remarcar lo importante que es para las empresas tenerlo en cuenta. En numerosas ocasiones se han visto empresas de nueva creación que comienzas sus primeros meses con un incremento de clientes impresionante pero que acaban fracasando estrepitosamente. Esto se debe a que no

han tenido en cuenta el Churn Rate y solo han pensado en atraer nuevos clientes de cualquier manera posible. Es posible que detecten que están perdiendo clientes, pero no lo consideren importante ya que están ganando clientes por otro lado, pero llega un momento en que esto no es sostenible porque no se va a ganar una cantidad grande de clientes durante mucho tiempo. Si hubiesen tenido en cuenta la tasa de abandono de clientes podrían haber tomado medidas para retener a sus clientes en vez de centrarse solo en ganar nuevos.

Por eso, es vital contar con un buen protocolo de retención de clientes. Si la causa de la baja es una incidencia o contratiempo, una propuesta de valor puntual y un buen trato de atención al cliente pueden salvar la situación. Pero si la causa de abandono es la falta de uso o el desinterés, tendremos que ponernos manos a la obra para facilitar una guía de uso o una serie de acciones que motiven la utilización del servicio. Se trata de lograr que el cliente establezca como hábito la utilización de nuestro servicio, o que recurra a nosotros siempre para una de sus necesidades periódicas. Debemos estar presentes siempre en su memoria como la mejor opción.

3. Inteligencia artificial para calcular el Churn Rate

3.1 Data Mining:

¿Qué es Data Mining?

El Data Mining es un proceso que utiliza una gran cantidad de datos con el fin de detectar relaciones comerciales y patrones significativos que permitan comprender el comportamiento de esos datos. Existen diferentes métodos de Data Mining que son capaces de analizar grandes bases de datos de una manera automática o semiautomática. Una vez que los datos son analizados podemos obtener unos resultados que nos aporten ventajas competitivas respecto al resto del sector (Martin 2018).

> *"Proceso de Seleccionar (Selecting), Explorar (Exploring), Modificar (Modifying), Modelizar (Modeling) y Valorar (Assessment) grandes cantidades de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores"* (SAS Institute, 2019).

Desde hace mucho tiempo ya se utilizan herramientas estadísticas para analizar muestras representativas con una cantidad considerable de datos. Son útiles para encontrar relaciones entre los datos, pero en el momento en el que la cantidad de datos empieza a ser muy grande estos análisis se ven superados y aquí, es donde entra el Data Mining. En un análisis estadístico, cuando nos encontramos con un número muy alto de variables lo más normal es descartar alguna de ellas o reducir la cantidad de datos incluidos en ellas. Al hacer esto estamos desechando información que puede llegar a ser importante para nuestro propósito llevándonos a imprecisiones en el resultado final. El Data Mining no tiene este inconveniente porque profundiza en los datos empleando algoritmos de aprendizaje automático que permiten explicar lo que está sucediendo dentro de esos datos.

Puede parecer que esta tecnología solo es usada por determinados sectores o que es de reciente creación, pero lo cierto es que se lleva realizando desde hace más de 30 años y cada vez se utiliza en más sectores (Vidal, 2016). Una de las aplicaciones del análisis de datos es permitir la extracción de información para entender mejor las necesidades de nuestros clientes y orientar una determinada promoción o campaña

publicitaria hacia ellas. Obtener nuevos clientes o mantener los ya existentes es mucho más fácil gracias a estos métodos, lo que no pasa nada desapercibido en casi ningún sector de la economía.

Cada análisis realizado con metodología de Data Mining puede ser totalmente distinto al anterior, pero por término general, la mayoría de los análisis suelen seguir un proceso similar que está formado por cuatro etapas (Sinexus, 2019):

1. *Determinación de los objetivos:* en esta fase se fijan los objetivos que se desean analizar a través de Data Mining.
2. *Procesamiento de los datos:* esta tarea consiste en la selección y limpieza de las bases de datos objetivo de análisis con el fin de enriquecerlas y reducir su contenido para su análisis. Este proceso suele ocupar la mayor parte de tiempo del proceso de Data Mining.
3. *Determinación del modelo*: a la hora de determinar el modelo se comienza realizando un análisis estadístico de los datos tras el cual se elaborará un análisis gráfico de los resultados obtenidos. Esto nos permitirá tener una primera visión de los resultados de nuestro análisis. Según el tipo de estudio que se esté realizando y sus objetivos se podrán emplear diferentes técnicas de Inteligencia Artificial para su elaboración.
4. *Análisis de los resultados:* como paso final, se deberá comprobar que los resultados obtenidos son coherentes con el análisis gráfico y estadístico realizados anteriormente.

Siguiendo estos pasos podemos descubrir patrones que nos resulten de interés sobre una gran cantidad de datos los cuales nos ofrecen conocimiento sobre un tema determinado. No siempre es sencillo encontrar estos patrones debido a que, al analizar una cantidad muy elevada de datos, estos pueden contener errores o estar incompletos. Esto puede provocar que la minería no sea del todo precisa y los patrones obtenidos puedan no ser los deseados o no resulten del todo interesantes. Por esto las técnicas de Data Mining tienen que implementar métodos de limpieza de datos, realizar un correcto procesado de los datos y ser capaces de detectar outliers que afecten al resultado final.

Los patrones obtenidos a través de técnicas de Data Mining pueden no ser siempre interesantes, lo que les hace interesantes puede variar entre los diferentes usuarios. Por eso estos métodos tienen que aprender a valorar la información obtenida en base a patrones más subjetivos de cada individuo. Cuando el método sea capaz de integrar medidas basadas en el interés de la propia persona debido a unos valores dados previamente o a resultados anteriores será más efectivo. Esto puede permitir reducir el campo de búsqueda inicial para dotar de mayor precisión al análisis.

A modo de ejemplo sencillo podemos comentar el Data Mining realizado por la cadena de televisión inglesa BBC. Dicha cadena utilizó técnicas como los árboles de decisión y las redes neuronales con el fin de identificar los patrones de visualización de sus programas. Obteniendo datos de una gran cantidad de programas emitidos en televisión consiguió predecir el tiempo optimo que debía durar un programa o la hora a la que era apropiado emitirse para conseguir la mayor audiencia posible. Gracias al aprendizaje automatizado esta especie de base de datos se va completando continuamente con nuevos datos para conseguir mejores resultados (BBC, 1996).

Existen muchas técnicas de Data Mining diferentes, algunas con más popularidad que otras. Estas técnicas funcionan de distinta manera según el tipo de datos analizado o el tamaño de muestra, por ejemplo. A continuación, basándonos en los estudios

realizados sobre este tema en los últimos años se van a explicar las técnicas más populares y que ofrecen mejores resultados.

3.2 Técnicas y algoritmos de Data Mining:

Tras definirnos el problema que ocasiona el Churn Rate se definen algunos de los principales métodos de Data Mining mostrando sus ventajas, desventajas y posibles casos de aplicación. Esto se realiza con el fin de darnos una visión de cuales se adaptarían más al tipo de análisis que queramos realizar. Tras esta descripción se clasifican diferentes métodos mediante una serie de características divididas en dos tablas. Los métodos y algoritmos han sido escogidos analizando los estudios relacionados con este tema de los últimos 3 años, eligiendo los más utilizados.

En primer lugar, se expone una tabla a modo de resumen de diferentes técnicas de Data Mining usadas en los últimos años para mostrar los diferentes algoritmos y el resultado que produce cada modelo.

Tabla 1. Agrupación de algoritmos por Tipo y Resultado.

Algoritmo	Tipo de algoritmo	Resultado
Regresión Logística	Estadístico	Fórmula para calcular una probabilidad
Arboles de decisión	Clasificación	Reglas con una única conclusión (atributo o target), expresadas mediante un árbol descriptivo
RNA	IA	Función con valor de salida de la neurona en base al estado de activación de la misma
SVM	IA	Fórmula para calcular una probabilidad
Naive Bayes	Estadístico / Clasificación	Árbol de ponderación de dependencias significativas
Reglas de asociación	Asociación	Reglas con conclusiones diferentes por asociación de atributos
SNA	IA / Clasificación	Árbol de la estructura de la red

Fuente: *Análisis de algoritmos aplicados al Churn Analysis, (Fabbro y Deroche 2019).*

En la Tabla 1 se puede ver el tipo/metodología utilizada por cada algoritmo, así como el resultado que obtendremos al aplicarlo para poder ver cual resultaría más efectivo en función del tipo de análisis que queramos realizar. Esta clasificación nos permite tener una mayor precisión al escoger el método que queremos aplicar en función de los datos que se tienen y el resultado que se quiere obtener.

A continuación, en la Tabla 2, se muestra una clasificación que indica el grado de dificultad de aplicación de cada algoritmo remarcando el nivel de experiencia necesario

para poder desarrollarlo correctamente. También indica el tipo de datos que son necesarios para su aplicación distinguiendo entre numéricos o discretos. Por último, permite ver si es posible sacar información extra del análisis a parte de la que buscamos al realizarlo.

Tabla 2. Agrupación de algoritmos por tipo de datos, información y experiencia de equipo.

Algoritmo	Experiencia del equipo	Información extra	Tipo de datos
Regresión Logística	Alta	Si	Numéricos
Arboles de decisión	Baja	Si	Todos
RNA	Media	No	Numéricos
SVM	Alta	No	Numéricos
Naive Bayes	Media	Si	Discretos
Reglas de asociación	Baja	Si	Todos
SNA	Baja	Si	Todos

Fuente: *Análisis de algoritmos aplicados al Churn Analysis, (Fabbro y Deroche, 2019)*

Como observamos en la Tabla 2 existen algunos algoritmos que solo son válidos si utilizamos datos numéricos mientras que otros se pueden usar con cualquier tipo de dato. Esto puede afectar a la precisión de los algoritmos, por un lado, la regresión logística a pesar de ser un método sencillo, si lo llevamos al extremo puede complicarse más dándonos unos mejores resultados. Por otro lado, realizar arboles de decisión es una tarea sencilla que nos puede servir para tener una visión más general del negocio. Métodos como las redes neuronales o el Support Vector Machine (SVM) pueden resultar más complicados en su aplicación, pero gracias a la Inteligencia Artificial (IA) pueden llegar a ser automatizados para ser aplicados con mayor facilidad y obtener unos buenos resultados.

3.2.1 Regresiones:

Este método de Data Mining es, probablemente, el más sencillo que existe para poder realizar predicciones de acciones futuras en base a datos ya existentes. Al ser un método sencillo de aplicar puede que no muestre la precisión que se podría esperar de un modelo más complejo. Las regresiones pueden ser lineales o no lineales. En el caso de las primeras, como su propio nombre indica, consisten en trazar una línea que se asemeje lo más posible a una nube de puntos (datos) que tenemos. El problema de este modelo es que al ajustarse tanto a los datos existentes puede ser impreciso a la hora de introducir nuevos datos que difieran un poco de los previos. Por lo tanto, no ofrecen resultados demasiado precisos ante sucesos más complejos o diferentes.

El modelo de regresión lineal se basa en obtener la relación entre dos variables de una manera consistente. Una vez que conocemos la relación entre las variables podemos estimar los valores de una determinada variable en función de los datos existentes de la otra variable. Los resultados obtenidos aplicando este método no son totalmente precisos ya que se trata de una aproximación. Debido a esto, toda regresión realizada tiene un término de error o residuo que no podemos eliminar.

La regresión lineal múltiple tiene en cuenta todas las variables predictoras necesarias

para explicar un hecho al mismo tiempo. Al tener en cuenta un mayor número de variables ya no solo podemos obtener el nivel de influencia de una variable en otra sino que podemos establecer un orden jerárquico entre ellas y obtener un subconjunto relevante.

El método de regresión nos permite comprobar los efectos que tienen variables medidas en diferentes escalas lo que permite desarrollar modelos predictivos con una mejor evaluación.

Avanzando un poco más en el modelo de regresiones encontramos la regresión logística la cual nos permite introducir variables que se miden con una escala de 1 y 0 (verdadero o falso). Este tipo de regresión no requiere una relación lineal entre las variables. Se suele usar frecuentemente en problemas de clasificación. Estas variables que se pueden clasificar con un 1 o un 0 son muy frecuentes por lo que si no se les puede aplicar una regresión lineal hay que aplicar la logística convirtiendo estos valores en probabilidades.

La principal ventaja del método de regresiones es que es muy sencillo de entender y permite ver la relación existente entre los datos de una manera muy rápida. Esto a su vez puede convertirse en desventaja si nuestro objetivo es descubrir relaciones complejas entre variables. Para esta necesidad este modelo se nos puede quedar un poco corto y es necesario desarrollar modelos más complejos.

Este modelo se puede aplicar para la identificación de variables explicativas que nos resultan útiles o son más influyentes en nuestro estudio a la vez que descartamos las que nos son de menos utilidad o no tienen influencia alguna en la respuesta. También nos permite ver las relaciones que se producen entre las variables influyentes del modelo y que afectan a la variable de respuesta. Por último, nos permite identificar las variables de confusión que afectan a la relación entre el resto de las variables de nuestro modelo.

Este método se puede automatizar a través de algoritmos de Machine Learning con el fin de desarrollar los modelos de una manera mucho más rápida y efectiva mediante la inteligencia artificial.

Con el fin de ilustrar más este método se plantea un ejemplo de un caso de estudio en el sector bancario. Customer churn analysis – a case study (Teemu Mutanen, 2006).

El sector de la banca presenta, por lo general, una tasa de abandono menor que el de las telecomunicaciones. Normalmente cuando un cliente elige un banco no suele cambiar a no ser que ocurra alguna circunstancia drástica que lo provoque. El problema es que si esto pasa el efecto puede ser mucho más grande que la pérdida de un cliente en otro sector diferente.

El método utilizado para realizar el análisis al que nos referimos es el de regresiones logísticas. Los datos fueron obtenidos de una base de 251.000 clientes de los cuales se escogieron aleatoriamente 115.000 para realizar la regresión. Se tuvieron en cuenta 75 variables en el modelo las cuales estaban relacionadas con las transacciones de la cuenta, datos personales, datos de cliente e indicadores de servicio. El análisis se realizó para un periodo de 3 años en intervalos de 3 meses considerando churners a los que se daban de baja durante ese periodo obteniendo 12 puntos clave a lo largo de la muestra. El problema, al igual que en el caso anterior, es que la cantidad de churners existente es muy pequeña, solo 4.275 clientes son considerados churners de los 115.000 (3,7%).

Para solucionar este problema se puede aplicar el sobremuestreo o el de reducción de la muestra mayoritaria hasta llegar a un nivel decente en la minoritaria. En este estudio se ha empleado el segundo método llegando a una situación de churner/no churner de 1/1 y de 2/3 para realizar las predicciones. Aun así, la validación del modelo se realizó con el conjunto total de datos.

Para obtener los resultados se utilizan matrices de confusión. Estas matrices sirven para comparar el número de casos reales que se obtienen frente a los previstos, obteniendo tasas de verdaderos y falsos positivos y por otro lado verdaderos y falsos negativos.

Se estimaron 6 modelos distintos con diferentes datos generados de manera aleatoria obteniendo los siguientes resultados mediante una matriz de confusión que nos muestra los churners identificados correctamente:

Tabla 3 Eficacia de los modelos de regresión en el caso estudiado

	Nº de predicciones correctas	% de predicciones correctas	% de churners	% de churners identificados correctamente
Modelo 1	69670	62	0,8	75,6
Modelo 2	81361	72	0,9	60,5
Modelo 3	66346	59	0,8	79,5
Modelo 4	72654	65	0,8	73,4
Modelo 5	15384	14	0,5	97,5
Modelo 6	81701	73	0,9	61,3

Fuente: *Customer churn analysis – a case study, (Teemu Mutanen 2006).*

En la Tabla 3 podemos observar como el % de churners detectados correctamente varia significativamente entre los diferentes modelos. El dato clave de esta tabla se trata del % de churners identificados correctamente. Se puede ver que los modelos más eficaces a la hora de predecir correctamente presentan un porcentaje de churners correctos más bajo. Esto se debe al desequilibrio entre el total de la muestra y los churners (115.000 – 4.275) que ya hemos comentado anteriormente. De todos modos, el modelo detecta los churners de una manera correcta ya que supera el 60% en todos los modelos.

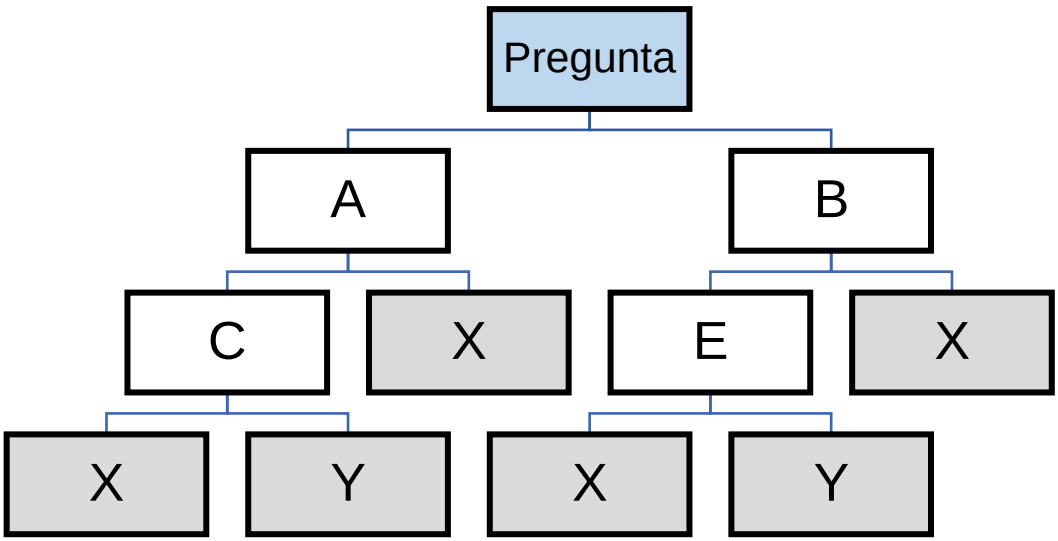
Como conclusión de este estudio podemos afirmar que la diferencia entre los datos de cada cliente registrados en los modelos nos indica la naturaleza dinámica de los clientes churners de este sector. Esto provoca que la elaboración de de un modelo estándar para realizar predicciones en este ámbito es una tarea costosa debido al dinamismo de los datos. Comprobamos que si se quiere predecir datos utilizando modelos de regresión es necesario actualizar constantemente los datos del modelo para poder obtener predicciones más precisas. Por otro lado, también es importante definir correctamente la variable objeto de estudio, en este caso se utilizó como variable para determinar si el usuario puede clasificarse como churner o no la posesión de una cuenta corriente al final de cada periodo. Si se hubiesen tenido en cuenta otro tipo de variables los resultados podrían haber sido distintos.

3.2.2 Arboles de decisión:

Cuando nuestro conjunto de datos tiene una gran cantidad de variables explicativas, las cuales interactúan entre sí, puede ser complicado utilizar un modelo de regresión para analizar todos esos datos. Los árboles de decisión nos permiten apreciar las relaciones existentes entre estos datos mediante la creación de categorías secundarias o

diferentes grupos. Son un método estadístico orientado a problemas de decisión supervisados. Estos métodos se basan en obtener un determinado resultado a partir de unas variables conocidas como inputs. Se les conoce así porque pueden ser representados en forma de árbol:

Gráfico 1: Árbol de decisión



Fuente: Elaboración propia

A cada casilla le vamos a llamar nodo. En este caso la casilla de la parte superior del gráfico es el nodo raíz porque es el primero que encontramos en el árbol, las de color blanco las conocemos como nodos padres debido a que de ellos nace algún nodo más. Por último, tenemos los nodos de color gris que son los nodos terminales porque ya no tienen continuación y representan el final de un subgrupo resultante de todas las divisiones anteriores. Así una vez que llegues a un nodo terminal obtendrás una respuesta a la cuestión que se está tratando.

Este mecanismo se puede utilizar por ejemplo en una entidad de crédito a la hora de conceder o no un crédito a una persona. Cada nodo blanco lo podríamos interpretar como una pregunta a la que la respuesta puede ser sí o no (o el valor de una variable) y entonces, dependiendo del resultado, avanzaríamos a uno u otro nodo hasta llegar a un nodo terminal que indica si conceder el crédito o no.

Los árboles de decisión pueden servir de gran ayuda en los casos en las que las alternativas de acción están bien definidas y pueden ser cuantificadas con una probabilidad de éxito.

Este método, aplicado a la inteligencia artificial utilizando algoritmos nos puede proporcionar unos resultados que permitan mejorar el rendimiento de la empresa. Con el aprendizaje automático se puede ajustar el porcentaje de respuesta a cada nodo a través de la experiencia con el fin de ser cada vez más precisos en nuestros resultados. En el caso anterior nos permitiría tener un mayor éxito a la hora de conceder créditos a personas que sean capaces de pagarlos y reducir así el riesgo.

Dentro de los algoritmos aplicados a los árboles de decisión destaca el ID3 (Quinlan 1983) el cual se usa para crear los árboles de decisión a través de la Teoría de la Información de Shannon (Weaver y Shannon 1940). Este método construye el árbol de

arriba hacia debajo de forma directa, usando el concepto de ganancia de información para encontrar el atributo más útil en cada caso. La ganancia de información se calcula mediante la entropía la cual permite calcular el grado de incertidumbre de una muestra. El objetivo es reducir la entropía para poder obtener una ganancia de información mayor. Al llegar a un nivel de entropía de 0 obtenemos un nodo terminal que ya nos da una decisión. Este proceso se sigue repitiendo con todos los nodos que no son terminales, en el caso de que falle se divide el nodo hasta llegar a uno terminal para poder clasificarlo mejor. Este algoritmo nos proporciona una serie de beneficios:

-Suministra un conjunto de reglas comprensibles para un conjunto de datos
-Es un algoritmo muy rápido
-Elabora un árbol minimizado
-Cuando encuentra un nodo terminal se detiene el algoritmo por lo que reduce el número de comprobaciones necesarias.

Para ver la aplicación de este modelo a un caso práctico podemos considerar los resultados de un estudio de la aplicación de Arboles de decisión para el cálculo del Churn Rate en entornos B2B desmarcándose de los casos más comunes que serían de tipo B2C. “Managing B2B customer churn, retention and profitability” Tamaddoni Jahromi et al. (2014).

Por lo general la tasa de abandono de clientes se identifica con las empresas que pierden a los clientes que consumen sus productos, pero para ofrecer estos productos también necesitan de la ayuda de otras empresas que actúan como proveedores u ofreciéndoles sus servicios. El efecto que supone la pérdida en este caso puede ser muy grande para la empresa por lo que al igual que en los entornos B2C hay que tenerlo muy en cuenta.

> *“Hoy en día, debido a la mejora del acceso a la información, los clientes son más transitorios y es más fácil y menos costoso para ellos cambiar entre competidores” (Wiersema, 2013).*

Los estudios realizados en este sector son muy escasos por lo tanto el objetivo de este estudio es mostrar la importancia de la retención de clientes en este ámbito utilizando varias técnicas de Data Mining. La pérdida de un cliente comercial es mucho mayor ya que estos son más escasos y valiosos para las empresas.

Para realizar el análisis se ha hecho una revisión de la literatura para encontrar los métodos más relevantes históricamente y de estos se ha elegido utilizar los árboles de decisión debido a la robustez de sus resultados y su fácil entendimiento. Los datos elegidos se corresponden con 11.021 clientes empresariales de un importante minorista de bienes de consumo australiano. El periodo considerado es de 1 año y se ha dividido en 2 etapas de 183 días considerando la primera como la de calibración y la segunda sobre la que se quiere hacer la predicción. Se considera churner si no ha realizado ninguna actividad al final de ese periodo. La tasa de no churners real es del 72% y la de churners del 28%. Se va a utilizar matrices de confusión para mostrar los datos como en los casos anteriores. Los resultados obtenidos por el modelo son los siguientes:

Tabla 4: Matriz de confusión B2B

		Previsto	
Modelo	Real	No churner	Churner
Árbol de decisión	No churner	88%	12%
	Churner	8%	92%

Fuente: Elaboración Propia a partir de datos de "Managing B2B customer churn, retention and profitability" Tamaddoni Jahromi et al. (2014).

En la Tabla 4 se puede observar como el modelo de árboles de decisión es bastante efectivo en este caso de estudio llegando a identificar en torno a un 90% de los clientes correctamente. Esto indica que los modelos de predicción funcionan correctamente en entornos B2B llegando a ser muy útiles a la hora de identificar los posibles churners.

3.2.3 Redes neuronales artificiales:

Un sistema de procesamiento de información formado por una elevada cantidad de elementos de procesamiento (neuronas), conectados entre sí a través de canales de comunicación se conoce como Redes Neuronales Artificiales (RNA) (Reguero 1995). Estas conexiones establecen una estructura jerárquica siendo capaces de interactuar con los objetos del mundo real intentando imitar al sistema nervioso de las personas. Mientras que la computación tradicional se basa en algoritmos predecibles, la computación neuronal es capaz de generar sistemas que solucionen problemas más complejos cuyo desarrollo matemático tiene una dificultad muy elevada.

Las redes neuronales artificiales tienen como objetivo imitar la manera en la que el cerebro humano aprende. De esto se encargan las neuronas que reciben estímulos y emiten respuestas a nuestro cerebro. Las RNA están formadas por neuronas que obtienen entradas de un determinado tamaño y posteriormente emiten una salida.

Las neuronas de nuestro cerebro están conectadas entre sí, trabajan en equipo. Cada vez que reciben un estímulo del exterior se activan procesando la información recibida y buscan la forma de procesar una respuesta de la manera más precisa posible interconectándose entre sí. Las neuronas van aprendiendo con cada estimulo y consiguen generar respuestas más rápidas y precisas gracias a este aprendizaje. En resumen, recibido un determinado estimulo, existe una forma de obtener el resultado deseado. Este proceso es el que tratan de imitar las Redes Neuronales.

Se basan en el concepto de que, dados una serie de parámetros, siempre existe una forma de combinarlos para ser capaces de predecir un determinado resultado. Encontrar esta combinación es el proceso de aprendizaje anteriormente comentado. En el momento que una red neuronal esta ya entrenada se puede emplear para realizar predicciones aplicando la combinación establecida previamente.

La misión de las RNA es minimizar el error que se produce al desarrollarse cualquier tipo de proceso gracias al aprendizaje. Su trabajo consiste en encontrar relaciones entre las entradas que van llegando con el fin de poder procesar la mejor respuesta posible. Este análisis nos permite realizar predicciones de lo que va a suceder con unos determinados datos y cómo actuar ante ellos obteniendo mejores cada vez que lo realizamos.

Esto se logra gracias a los principios de funcionamiento de las RNA, de los cuales citamos a continuación los cinco más importantes (Hilera, 1995).

-Aprendizaje adaptativo: probablemente la característica más importante de las RNA, ya que pueden comportarse en función de un entrenamiento con una serie de ejemplos ilustrativos. De esta forma, no es necesario elaborar un modelo a priori, ni establecer funciones probabilísticas. Una RNA es adaptativa porque puede modificarse constantemente si necesita adaptarse a cualquier condición de trabajo nueva.

- Autoorganización: mientras que el aprendizaje es un proceso donde se modifica la información interna de la RNA, la autoorganización consiste en la modificación de toda la red completa con el fin de llevar a cabo un objetivo específico. Autoorganización significa generalización, de esta forma una red puede responder a datos o situaciones que no ha experimentado antes, pero que puede inferir en base a su entrenamiento. Esta característica es muy útil sobre todo cuando la información de entrada es poco clara o se encuentra incompleta.

- Tolerancia a fallos: en la computación tradicional la pérdida de un fragmento pequeño de información puede acarrear comúnmente la inutilización del sistema. Las RNA poseen una alta capacidad de tolerancia a fallos. La tolerancia a fallos se entiende aquí en dos sentidos: primero, las redes pueden reconocer patrones de información con ruido, distorsión o incompletos (tolerancia de fallos respecto de los datos); y segundo, pueden seguir trabajando (con cierta degradación) aunque se destruya parte de la red (tolerancia a fallos respecto de la estructura). La explicación de este fenómeno se encuentra en que, mientras la computación tradicional almacena la información en espacios únicos, localizados y direccionables, las redes neuronales lo hacen de forma distribuida y con un alto grado de redundancia.

- Operación en tiempo real: de todos los métodos existentes, la RNA son las más indicadas para el reconocimiento de patrones en tiempo real, debido a que trabajan en paralelo actualizando todas sus instancias simultáneamente. Es importante destacar que esta característica solo se aprecia cuando se implementan redes con hardware especialmente diseñado para el procesamiento en paralelo.

- Fácil inserción en la tecnología existente: es relativamente sencillo obtener chips especializados para redes neuronales que mejoran su capacidad en ciertas tareas. Ello facilita la integración modular en los sistemas existentes.

El modelo más sencillo de Redes Neuronales es el perceptrón.

> "*Un perceptrón es un elemento que tiene varias entradas con un cierto peso cada una. Si la suma de esas entradas por cada peso es mayor que un determinado número, la salida del perceptrón es un uno. Si es menor, la salida es un cero. Además de como una unidad básica, también puede entenderse como una red neuronal artificial en sí misma" (Ruiz, 2015).*

Esto podría hacer referencia a una neurona que obtiene información y la procesa, pero en las RNA existen varias capas que procesan la información recibida y que están interconectadas entre sí. Mediante este proceso cada capa puede encontrar características que le ayuden a clasificar todos los datos recibidos con una mayor precisión cada vez. Lo que se hace añadiendo más capas es añadir información de la que no disponíamos antes teniendo una sola capa. Se escogen los datos que más nos ayuden a comprender lo que estamos buscando y obtenemos sus características.

Durante el proceso de aprendizaje cada capa empieza a comprender cuáles son esas

características que más nos sirven a la hora de conseguir nuestro objetivo y las detecta en próximas ocasiones.

Con el fin de comprobar la aplicación de este modelo en el análisis del Churn Rate vamos a analizar un estudio sobre el mercado de las telecomunicaciones en Brasil. "Data Mining Techniques on the Evaluation of Wireless Churn" Ferreira et al. (2004*).*

En este trabajo se habla sobre la importancia que tiene el Churn Rate en este sector y como poder desarrollar métodos para reducirlo a través de técnicas de Inteligencia Artificial y Data Mining. Para poder realizar este análisis disponían de una muestra de 100.000 usuarios de los cuales 1.125 se daban de baja después del periodo estimado. Dado el pequeño porcentaje de "churners" (clientes que se san de baja) se realizó un sobre muestreo obteniendo 3.500 observaciones de las cuales el 28% eran churners con el fin de comprobar mejor la detección de estos clientes por parte de los algoritmos.

Para identificar estos clientes se usaron hasta 37 variables en los modelos entre las que podemos destacar:

- Datos de facturación (ingresos mensuales, coste de la itinerancia, etc.);
- Datos de uso (tiempo, tipos de servicios utilizados, etc.);
- Datos demográficos de los clientes (edad, sexo, región, etc.);
- Datos de la relación con el cliente (plan tarifario, tecnología utilizada, edad del teléfono, etc.);
- Datos de mercado (tarifas de la competencia, costes de publicidad, etc.).

Después de realizar las estimaciones correspondientes con el modelo se utilizaron matrices de confusión para comprobar su eficacia. Los resultados son los siguientes:

Tabla 5: Matriz de confusión

		Previsto	
Modelo	Real	No churner	Churner
Red neuronal	No churner	70%	30%
	Churner	35%	65%

Fuente: "Data Mining Techniques on the Evaluation of Wireless Churn" Ferreira et al. (2004).

En la tabla 5 observamos como el modelo de redes neuronales es capaz de identificar al 70% de los clientes que van a permanecer en la compañía y al 65% de los clientes que se van a dar de baja. De este modo comprobamos en que en este caso el modelo de redes neuronales es eficaz a la hora de identificar los clientes.

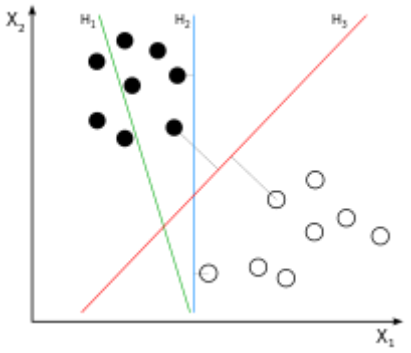
Este estudio obtiene unas conclusiones que indican que la selección de variables para medir el Churn Rate es fundamental para tener éxito y poder desarrollar las medidas necesarias para poder reducirlo. En concreto señalan que para este sector las más importantes son el uso nocturno y el tiempo que usan datos/roaming. Por lo tanto, estas técnicas nos permiten detectar en que puntos nos tenemos que basar para poder desarrollar estrategias que favorezcan la permanencia de los clientes y los puntos débiles que tenemos que reforzar.

3.2.4 Support Vector Machine (SVM):

El Support Vector Machine (SVM) es un modelo supervisado de aprendizaje con algoritmos asociados que analizan los datos y reconocen patrones, que se utiliza para la clasificación y el análisis de regresión en la Inteligencia de Negocios. El SVM básico toma un conjunto de datos de entrada y predice, para cada entrada dada, a cuál de las dos clases de salida pertenece, por lo que es un clasificador no-probabilístico lineal binario (solo escoge entre 2 opciones). Dado un conjunto de ejemplos de entrenamiento, cada uno marcado como perteneciente a una de dos categorías, un algoritmo de entrenamiento construye un modelo que asigna nuevos ejemplos en una categoría u otra. Este tipo de modelos es muy utilizado para el análisis de modelos en los cuales se tiene que clasificar un conjunto de datos a solo dos categorías, fraude o no-fraude, sí o no al crédito, etc. Un modelo de SVM es una representación de los ejemplos (base de datos con la cual se realizó la estimación) como puntos en el espacio, de modo que asignan los ejemplos de las categorías separadas que generalmente quedan divididas por un espacio definido, espacio que tiene que ser tan amplio como sea posible. Los nuevos datos de entrada serán clasificados en el mismo espacio y para predecir a que categoría pertenece (Matías Riquelme, 2013).

Este podría considerarse un modelo de Support Vector Machine perfecto en el que ambos tipos de datos se encuentran perfectamente divididos. (Modelo lineal, Gráfico 2)

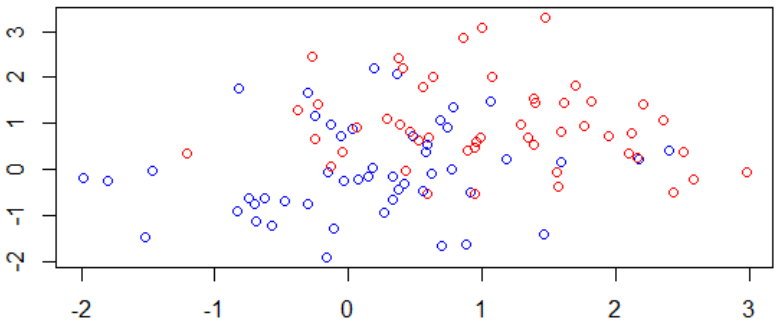
Gráfico 2: Modelo Lineal



Fuente: (Zack Wienberg, 2012)

En la realidad esto no suele pasar nunca, pues los datos normalmente están mezclados entre sí, siendo el objetivo de este modelo clasificarlos con la mayor tasa de acierto posible (Modelo no lineal, Gráfico 3).

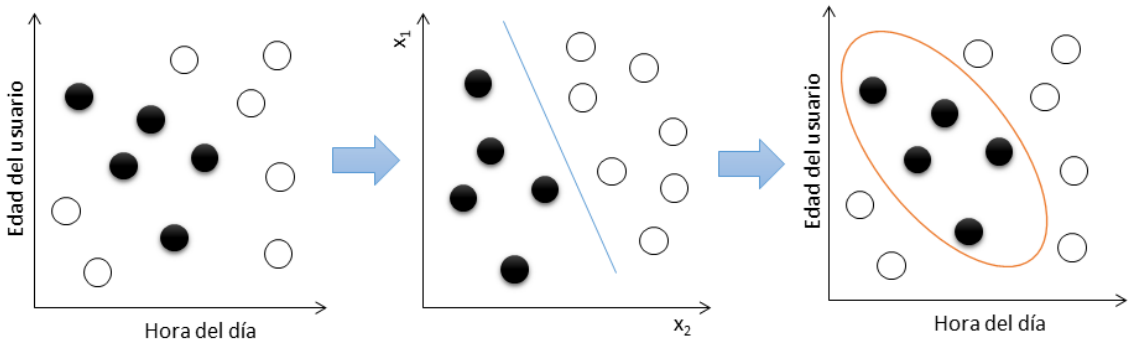
Gráfico 3; Modelo no lineal



Fuente: (Legorreta, 2015)

En el caso de que la función no sea lineal SVM permite utilizar las llamadas funciones Kernel (no lineales). Estas funciones resuelven el problema de clasificación trasladando los datos a un espacio donde el hiperplano solución es lineal y, por tanto, más sencillo de obtener (Álvarez 2016). Una vez conseguido, la solución se transforma, de nuevo, al espacio original:

Gráfico 4: Gráficos no lineales



Fuente: (Analítica Web, 2016).

En algunos casos de clasificación, tratamos de encontrar un hiper-plano óptimo que separe dos clases. Cuando hay un número similar o no muy diferente de datos de ambas categorías (como en las imágenes anteriores) utilizar este modelo general de Support Vector Machine es útil. El Churn Rate de clientes es un problema de clasificación entre "churn" y "no churn". Pero cuando el número de los ejemplos negativos es demasiado pequeño (cosa que suele ser habitual en la tasa de abandono de clientes), el rendimiento de la generalización del clasificador SVM debe ser débil, y los índices de error son insatisfactorios (Zhao, 2005).

Uno de los ejemplos más utilizados para explicar esta técnica es el de las flores Iris (Fisher 1936). La flor Iris tiene tres variantes: Setosa, Versicolor y Virginica. El objetivo es implementar un modelo de Support Vector Machine el cual, a través de las características de estas flores, sea capaz de clasificarlas y usar el modelo entrenado para predecir el tipo de especie del Iris. Así, con datos como pueden ser el ancho o la longitud del pétalo sea capaz de identificar correctamente el tipo de flor. A través del entrenamiento el modelo cada vez será capaz de identificar una mayor cantidad de flores correctamente.

Por último, se procede a exponer un ejemplo práctico para comprobar la eficiencia de este modelo. Se analizará el caso de una compañía crediticia. “An Application of Support Vector Machines for Customer Churn Analysis: Credit Card Case” Kim et al. (2005).

Para el propósito de este estudio se obtienen los datos de los clientes que conservan su tarjeta de crédito desde abril.1997 a octubre de 2000 y los clientes que cierran su cuenta durante el mismo período. El conjunto de datos cubre las variables demográficas y las variables sobre la tarjeta de crédito en uso. Después de filtrar los datos con valores faltantes, se seleccionan 4.650 muestras para cada uno de ellos.

Para comprobar su eficacia se estiman cinco modelos con 800 datos seleccionados aleatoriamente:

Tabla 6. Eficacia de los modelos SVM

	Nº de predicciones correctas	% de predicciones correctas
Modelo 1	764	95.5%
Modelo 2	758	94.8%
Modelo 3	755	94.4%
Modelo 4	748	93.5%
Modelo 5	764	95.5%

Fuente: An Application of Support Vector Machines for Customer Churn Analysis: Credit Card Case*,* Kim et al. (2005).

Como vemos en la Tabla 6 la precisión del modelo supera el 93% en todos los casos demostrando su efectividad con este conjunto de datos.

3.2.5 Otros modelos y algoritmos

A parte de los cuatro métodos analizados anteriormente también existen otra serie de algoritmos para calcular el Churn Rate. Estos algoritmos no son tan usados en los estudios analizados, pero también tienen presencia en ellos:

Naïve Bayes:

> "Este algoritmo calcula la probabilidad condicional entre atributos de entrada y de predicción y supone que las relaciones de dependencias entre los atributos del conjunto de datos son condicionalmente independientes entre sí dado un atributo clase o target. De esta manera, se suele utilizar cuando se desea estimar la probabilidad de que ocurra un suceso determinado". *Análisis de algoritmos aplicados al Churn Analysis, (Fabbro y Deroche, 2019).*

SNA:

Hace referencia a Social Network Analysis y está surgiendo como un importante algoritmo en la sociedad moderna cada vez más presente en este campo. Es utilizado para observar las relaciones sociales en términos de la teoría de redes.

> "Permite reconocer las relaciones entre la gente para plasmarlas en un mapa que facilite la identificación del flujo de conocimiento: de quién toma la gente información y conocimiento, con quiénes lo comparten o quién conoce a quién; a diferencia de los organigramas que sólo muestren relaciones formales." *Análisis de algoritmos aplicados al Churn Analysis, (Fabbro y Deroche, 2019).*

Reglas de asociación:

> "Las reglas de asociación encuentran relaciones de interés entre los valores de los atributos en una base de datos. Su objetivo se centra en identificar patrones de asociación de ítems que aparecen juntos en un determinado conjunto de datos, representando esas asociaciones de dependencia mediante reglas. A través de éstas es posible conocer la probabilidad de que la ocurrencia de un conjunto de ítems implique la ocurrencia de otro conjunto de ítems". *Análisis de algoritmos aplicados al Churn Analysis, (Fabbro y Deroche, 2019).*

4. Revisión Bibliográfica

El objetivo de esta revisión bibliográfica es comparar los diferentes métodos de Data Mining utilizados en la predicción del Churn Rate en los últimos años con el fin de descubrir cuales son los más fiables en función del caso analizado. Para ello se expondrán varios estudios comparando sus porcentajes de acierto y efectividad, así como, el tipo y tamaño de muestra analizada.

Para realizar esta revisión se ha utilizado la base de datos Scopus, en la cual se han buscado artículos y trabajos utilizando palabas clave como “Churn Rate prediction”, “Churn análisis” o “Machine learning”. Se han utilizado artículos de los últimos tres años obteniendo una gran cantidad de trabajos relacionados con estos temas, de los cuales se ha ido seleccionando los más apropiados en base a una serie de criterios que se explicaran a continuación:

1º- En primer lugar, se han establecido unas palabras clave de búsqueda las cuales son “Churn prediction” y “Churn Rate” con el fin de obtener los resultados deseados

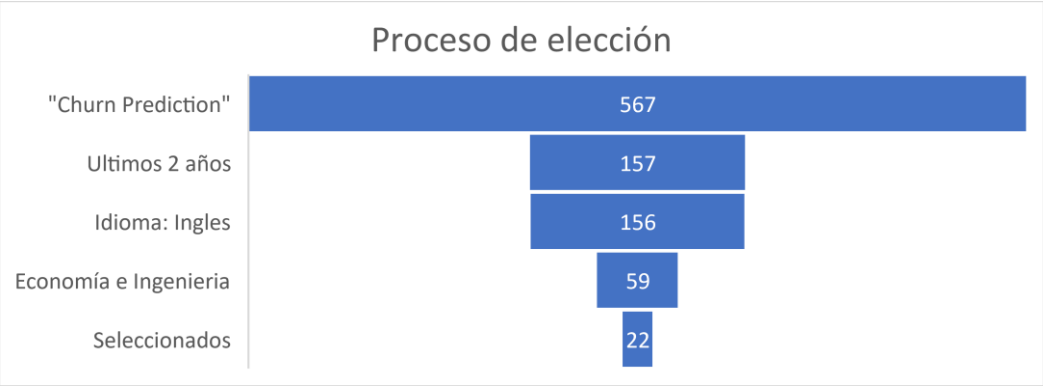
2º- A continuación, se escogen los resultados pertenecientes a los 3 últimos años (2017,2018 y 2019) para observar los estudios mas recientes en la materia.

3º- Se seleccionan solo los resultados en ingles

4º- Por último, se eligen los estudios de temas relacionados con la economía y la ingeniería ya que son dos de los campos en los que más útil es este análisis

Tras la aplicación de estos criterios se han seleccionado los cuatro métodos más utilizados en los últimos tres años, los cuales son: Regresiones, Redes Neuronales Artificiales (RNA), Decision Trees y Support Vector Machine (SVM).

A continuación, se muestra un gráfico que indica el proceso de selección trasladado a números.



Autor	Estudio
Abdi y Abolmakarem (2018)	Customer Behavior Mining Framework (CBMF) using clustering and classification techniques
Awang et al. (2017)	Improving accuracy and performance of customer churn prediction using feature reduction algorithm
Azeem et al. (2017)	A churn prediction model for prepaid customers in telecom using fuzzy classifiers
Chouiekh et al. (2017)	Machine Learning techniques applied to prepaid subscribers: case study on the telecom industry of Morocco
Gordini y Veglio (2017)	Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry
Gun et al. (2017)	Predicting customer churn in mobile industry using data mining technology
Jamil y Khan (2017)	Churn comprehension analysis for telecommunication industry using ALBA
Mishra y Reddy (2017)	A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers
Mishra y Reddy (2017)	A Novel Approach for Churn Prediction Using Deep Learning
Pijahari et al. (2017)	Evaluation of Machine Learning Models for Employee Churn Prediction
Abdi y Abolmakarem (2018)	Customer Behavior Mining Framework (CBMF) using clustering and classification techniques
Adhirai et al. (2018)	Customer Churn Prediction in Mobile Networks using Logistic Regression and Multilayer Perceptron (MLP)
Agrawal et al. (2018)	Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning
Alamsyah y Salma (2018)	A Comparative Study of Employee Churn Prediction Model
Arifin y Samopa (2018)	Analysis of Churn Rate Significantly Factors in Telecommunication Industry Using Support Vector Machines Method
Chen et al. (2018)	Promotion Recommendation Method and System Based on Random Forest
Chen y Yanfang (2018)	Research on E-commerce User Churn Prediction Based on Logistic Regression
Jiang et al. (2018)	Application of customer churn prediction based on weighted selective ensembles
Khalid et al. (2018)	Data Classification using Active Learning based Data Modification: An Application to Churn Prediction
Sabbeh (2018)	Machine-Learning Techniques for Customer Retention: A Comparative Study
Vijaya y Sivasankar (2018)	Improved churn prediction based on supervised and unsupervised hybrid data mining system

Chumwatana (2019)	Using Classification Technique for Customer Relationship Management based on Thai Social Media Data
Kumar y Kumar (2019)	Predicting Customer Churn Using Artificial Neural Network
Sarode et al. (2019)	Customers Churn Prediction with Rfm Model and Building a Recommendation System using Semi-Supervised Learning in Retail Sector

A continuación, se muestra el análisis de los estudios incluyendo los datos más relevantes. Se expone una breve descripción de trabajo, el método utilizado, el tamaño de muestra y el resultado. De este análisis se van a obtener los resultados que posteriormente se compararan entre si para extraer las conclusiones. La columna de "método" nos va a permitir identificar cual es el mas utilizado en estos últimos años para así poder observar las técnicas mas adecuadas para analizar estos temas. El tamaño muestral nos permite ver la eficacia de los métodos estudiados en diferentes tamaños muestrales. Por ultimo el resultado aporta información sobre la precisión de los métodos.

Autor	Estudio	Descripción	Método	Muestra	Resultado
Awang et al. (2017)	Improving accuracy and performance of customer churn prediction using feature reduction algorithm	Estudio de atributos que influyen al calcular el CR	Regresión logística	n = 3.333	Precisión Regresión: 86%
Azeem et al. (2017)	A churn prediction model for prepaid customers in telecom using fuzzy classifiers	Estudio de pérdida de clientes de una empresa de telecomunicaciones	Support Vector Machine, Redes Neuronales Artificiales, Arboles de decisión y Regresión lineal	n = 480.000	Precisión Regresión: 60% Precisión RNA: 64% Precisión Árbol: 63% Precisión SVM: 59%
Chouiekh et al. (2017)	Machine Learning techniques applied to prepaid subscribers: case study on the telecom industry of Morocco	Estudio de pérdida de clientes de una empresa de telecomunicaciones	Support Vector Machine, Arboles de decisión y Regresión lineal	n = 635.997	Precisión Regresión: 90.2% Precisión Árbol: 82% Precisión SVM: 90.1%
Gordini y Veglio (2017)	Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry	Técnicas que permiten la retención de clientes en empresas B2B	Support Vector Machine y Regresión lineal	n = 80.000	Precisión Regresión: 83.8% Precisión SVM: 89.6%
Gun et al. (2017)	Predicting customer churn in mobile industry using data mining technology	Analisis Churn Rate en la industria telefónica	Regresión logística, Arboles de decisión y Redes Neuronales Artificiales	n = 26.408	Precisión RNA: 77% Precisión Árbol: 69% Precisión Regresión: 64%
Jamil y Khan (2017)	Churn comprehension analysis for telecommunication industry using ALBA	Algoritmo ALBA bajo diferentes métodos	Support Vector Machine y Arboles de decisión	n = 3.333	Precisión Árbol: 94.7% Precisión SVM: 88.9%
Mishra y Reddy (2017)	A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers	Estudio comparative de modelos	Support Vector Machine y Arboles de decisión	n = 3.333	Precisión Árbol: 90.9% Precisión SVM: 90.1%
Mishra y Reddy (2017)	A Novel Approach for Churn Prediction Using Deep Learning	Evaluciacion del RNA para la predicción de Churn Rate	Redes Neuronales Artificiales	n = 3.333	Precisión RNA: 86.8%

Estudio	Autor	Descripción	Método	Muestra	Resultado
Pijahari et al. (2017)	Evaluation of Machine Learning Models for Employee Churn Prediction	Estudio de la tasa de abandono de los empleados	Support Vector Machine y Arboles de decisión	n = 15.000	Verdaderos positivos: 98.9% (Tree), 80.3% (SVM) Verdaderos negativos: 93.6% (Tree), 59.7% (SVM)
Abdi y Abolmakarem (2018)	Customer Behavior Mining Framework (CBMF) using clustering and classification techniques	Revisión de varios autores sobre el Churn Rate	Redes Neuronales Artificiales y Arboles de decisión	n = 1.000	Precisión RNA: 75.6% Precisión Árbol: 74.2%
Adhirai et al. (2018)	Customer Churn Prediction in Mobile Networks using Logistic Regression and Multilayer Perceptron (MLP)	Predicción de abandono de clientes en el mercado de móviles	Redes Neuronales Artificiales y Regresión logística	n = 3.500	Precisión Regresión: 85.7% Precisión RNA: 94.1%
Agrawal et al. (2018)	Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning	Cálculo del Chrun Rate con Deep Learning (RNA Muticapa)	Redes Neuronales Artificiales	n = 7.190	Precisión RNA: 80%
Alamsyah y Salma (2018)	A Comparative Study of Employee Churn Prediction Model	Estudio comparativo de tecnicas de Data Mining	Arboles de decisión	n = 16.649	Precisión Árbol: 88.7%
Arifin y Samopa (2018)	Analysis of Churn Rate Significantly Factors in Telecommunication Industry Using Support Vector Machines Method	Análisis de factores clave en el estudio del Churn Rate	Support Vector Machine		
Chen et al. (2018)	Promotion Recommendation Method and System Based on Random Forest	Estudio de datos de los clientes con el fin de recomendar la oferta correcta	Support Vector Machine, Arboles de decisión y Redes Neuronales Artificiales	n = 500.000	Precisión RNA: 90.7% Precisión Árbol: 93.3% Precisión SVM: 68.7%
Chen y Yanfang (2018)	Research on E-commerce User Churn Prediction Based on Logistic Regression	Análisis de factores que indican si un comprador esté interesado en consumir	Regresión logística	n = 6.000	Precisión Regresión: 96.6%

Estudio	Autor	Descripción	Método	Muestra	Resultado
Jiang et al. (2018)	Application of customer churn prediction based on weighted selective ensembles	Comparativa de varios modelos para analizar un dataset	Support Vector Machine y Redes Neuronales Artificiales	n = 5.617	Precisión RNA: 55.6% Precisión SVM: 69.9%
Khalid et al. (2018)	Data Classification using Active Learning based Data Modification: An Application to Churn Prediction	Aplicación de SVM a un caso práctico	Support Vector Machine	n = 14.814	Precisión SVM: 92%
Sabbeh (2018)	Machine-Learning Techniques for Customer Retention: A Comparative Study	Estudio de pérdida de clientes de una empresa de telecomunicaciones	Support Vector Machine y Arboles de decisión	n = 3.333	Precisión Árbol: 95.2% Precisión SVM: 89.4%
Vijaya y Sivasankar (2018)	Improved churn prediction based on supervised and unsupervised hybrid data mining system	Comparativa de varios metodos	Support Vector Machine y Arboles de decisión	n = 50.000	Precisión Árbol: 87.6% Precisión SVM: 92.5%

Tras realizar la revisión bibliográfica se exponen las conclusiones a las que se ha llegado:

En primer lugar, como se puede ver en la Tabla 6 se muestra el número de veces que se ha usado cada método en la serie de trabajos analizada para tratar de ver cuál es el más utilizado.

Tabla 6: Frecuencia de uso

Método	Nº de veces utilizado
Support Vector Machine	12
Redes Neuronales Artificiales	8
Arboles de decisión	11
Regresiones	7

Aquí podemos ver que el modelo más utilizado en estudio es el Support Vector Machine, aunque sin presentar una clara diferencia con el resto de los métodos. De estos datos no podemos sacar una conclusión clara de que método puede ofrecer mejores resultados. También podemos destacar que si analizamos solo los estudios publicados en 2018, las Regresiones solo son utilizadas en 2 ocasiones situándose claramente por detrás del resto de métodos.

A continuación, en la Tabla 7 se muestra una media del porcentaje de acierto de los cuatro métodos.

Tabla 7: Porcentaje de acierto

Método	Porcentaje de acierto
Support Vector Machine	81.83%
Redes Neuronales Artificiales	77.9%
Arboles de decisión	84.8%
Regresiones	80.9%

En este caso se puede ver que el modelo con más precisión es claramente el de árboles de decisión aventajando en casi un 3% al segundo que es el de Support Vector Machine. Por detrás se quedan las Regresiones y las Redes Neuronales Artificiales, en este orden.

De este análisis podemos extraer dos conclusiones, la primera es que los dos métodos más usados anteriormente son los que menor porcentaje de acierto. Esto es algo que puede entenderse debido a tener una mayor cantidad de muestras. Por otro lado, vemos que los métodos que desempeñan una menor complejidad de ejecución tienen un porcentaje de acierto mayor.

Por último, se muestra la Tabla 8 en la que vemos el porcentaje de éxito en función del tamaño de la muestra.

Tabla 8: Media del acierto de los métodos

Tamaño de la muestra	Método	Porcentaje de acierto	Media
Entre 0 y 5.000	Support Vector Machine	89.4%	87.37%
	Redes Neuronales Artificiales	85.5%	
	Arboles de decisión	88.75%	
	Regresiones	85.85%	
Entre 5.000 y 15.000	Support Vector Machine	77.7%	84.6%
	Redes Neuronales Artificiales	67.8%	
	Arboles de decisión	96%	
	Regresiones	96.90%	
Mas de 15.000	Support Vector Machine	79.98%	78.07%
	Redes Neuronales Artificiales	77.23%	
	Arboles de decisión	80.6%	
	Regresiones	74.5%	

A partir de esta tabla se puede ver que el % de acierto de los métodos es mayor cuanto más pequeño es el tamaño de la muestra. Igual que en las dos tablas anteriores los dos modelos que más porcentaje de acierto muestran son los árboles de decisión y las Support Vector Machine. Los árboles de decisión son los más eficaces tanto en muestras pequeñas como en muestras grandes. Las Redes Neuronales Artificiales son el método que menor efectividad tiene.

5. Conclusión

Como se ha podido comprobar durante el estudio, la tasa de abandono de clientes es una métrica que hay que tener muy en cuenta a la hora de realizar la estrategia empresarial de la compañía, especialmente en el sector de empresas que trabajan con modelos de suscripción. Si a la vez que se ganan clientes no existe una preocupación por retener a los clientes actuales llegara el momento en el que las nuevas entradas sean cada vez más bajas y el abandono se mantenga o incremente.

Para solucionar este problema existen varias técnicas de Data Mining que permiten analizar el Churn Rate para así ver su porcentaje y sus posibles causas. A través del aprendizaje automatizado podemos perfeccionar estos métodos para que cada vez proporcionen resultados más precisos.

Tras el análisis de los estudios realizados sobre este tema en los últimos años podemos concluir que los métodos de Data Mining más utilizados son: Support Vector Machine, Redes Neuronales Artificiales, Arboles de decisión y las Regresiones. Entre estas cuatro opciones la que ofrece mejores resultados en los estudios analizados son los Árboles de decisión seguido de las regresiones. Estas dos técnicas superan al 80% de precisión media. A pesar de esto son menos usadas que Support Vector Machine y Redes Neuronales Artificiales. Esto se puede deber a que dependiendo del tipo de datos analizado unas sean más eficaces que otras, así se ha comprobado que en términos de Churn Rate las regresiones y los Árboles son las más eficaces.

Otro factor a tener en cuenta es el tamaño de la muestra, cuanto mayor sea menor

precisión demuestran los métodos. Un factor interesante en este análisis es que la precisión de los Arboles y las regresiones disminuye cuanto mayor es la base de datos estudiada. Mientras tanto el porcentaje de acierto de Support Vectot Machine y de Redes Neuronales Artificiales se reduce poco comparado con sus resultados en tamaños de muestra más pequeños. De aquí podemos concluir que cuanto mayor es el tamaño de muestra más se acerca la precisión entre todos los métodos por lo que puede resultar interesante aplicar Redes Neuronales y Support Vector Machine en tamaños de muestra más grandes

Bibliografía

ABDI.F y ABOLMAKAREM.S, 2018. Customer Behavior Mining Framework (CBMF) using clustering and classification techniques [Consulta: 25 de agosto de 2019] Disponible en: Scopus Database

ADHIRAJ.P, SHRISHA.B y SHARATH.K, 2018. Customer Churn Prediction in Mobile Networks using Logistic Regression and Multilayer Perceptron (MLP) [Consulta: 25 de agosto de 2019] Disponible en: Scopus Database

AGRAWAL.S, DAS.A Y DHAGE.S, 2018. Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning [Consulta: 24 de agosto de 2019] Disponible en: Scopus Database

ALAMSYAH Y SALMA , 2018. A Comparative Study of Employee Churn Prediction Model [Consulta: 24 de agosto de 2019] Disponible en: Scopus Database

ARIFIN.S y SAMOPA.F, 2018. Analysis of Churn Rate Significantly Factors in Telecommunication Industry Using Support Vector Machines Method [Consulta: 27 de agosto de 2019] Disponible en: Scopus Database

ARRIAGA.I. 2018. *El churn rate: entender y luchar contra el supervillano del SaaS* Medium [Consulta 10 de julio de 2019] Disponible en: https://medium.com/@ignacio_arriaga/el-churn-rate-entender-y-luchar-contra-el-supervillano-del-saas-3b6d3710ea2d

AWANG.M, MAKHTAR.M y RAHMAN.M, 2017. Improving accuracy and performance of customer churn prediction using feature reduction algorithms [Consulta: 10 de noviembre de 2019] Disponible en: Scopus Database

AZEEM.M. FONG.M y USMAN.M, 2017. A churn prediction model for prepaid customers in telecom using fuzzy classifiers [Consulta: 27 de agosto de 2019] Disponible en: Scopus Database

CHEN.Y, HSU.W, HU.W, TANG.S y YU.C, 2018. Promotion Recommendation Method and System Based on Random Forest [Consulta: 25 de agosto de 2019] Disponible en: Scopus Database

CHEN.L y YANFANG.Q, 2018. Research on E-commerce User Churn Prediction Based on Logistic Regression [Consulta: 24 de agosto de 2019] Disponible en: Scopus Database

CHOUIEKH.A, 2017. Machine Learning techniques applied to prepaid subscribers: case study on the telecom industry of Morocco [Consulta: 26 de agosto de 2019] Disponible en: Scopus Database

CHUMWATANA.T, 2019. Using Classification Technique for Customer Relationship Management based on Thai Social Media Data [Consulta: 27 de agosto de 2019] Disponible en: Scopus Database

DOMINGUEZ.D y HERNO.G. 2018. *Retención y 'Churn Rate'* ESIC [Consulta 14 de julio de 2019] Disponible en: https://www.esic.edu/documentos/editorial/resenas/9788473567183_Esic%20Alumni_01-04-08.pdf

FABBRO Y DEROCHE 2018. *Análisis de algoritmos aplicados al Churn Analysis,* [Consulta 22 de julio de 2019] Disponible en: http://brazilianjournals.com/index.php/BRJD/article/view/1425

GORDINI.N y VEGLIO.V, 2017. Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry [Consulta: 27 de agosto de 2019] Disponible en: Scopus Database

GUN.S, LEE.E y JINWHA.K, 2017. Predicting customer churn in mobile industry using data mining technology [Consulta: 27 de agosto de 2019] Disponible en: Scopus Database

JIANG.Y, WANG.H y XIA.G, 2018. Application of customer churn prediction based on weighted selective ensembles [Consulta: 27 de agosto de 2019] Disponible en: Scopus Database

KUMAR.M y KUMAR.S, 2019. Predicting Customer Churn Using Artificial Neural Network [Consulta: 24 de agosto de 2019] Disponible en: Scopus Database

MARTIN.S. 2018. *El churn rate o tasa de cancelación de clientes* Cyberclick [Consulta 10 de julio de 2019] Disponible en: https://www.cyberclick.es/numerical-blog/el-churn-rate-o-tasa-de-cancelacion-de-clientes

MISHRA.A y REDDY.S, 2017. A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers [Consulta: 25 de agosto de 2019] Disponible en: Scopus Database

MISHRA.A y REDDY.S, 2017. A Novel Approach for Churn Prediction Using Deep Learning [Consulta: 26 de agosto de 2019] Disponible en: Scopus Database

NESLIN ET AL. 2006. *Las 6 estrategias de marketing más eficaces para retener a tus clientes* [Consulta 26 de julio de 2019] Disponible en: http://www.venkyshankar.com/download/Neslin-et-al.-JSR-2006.pdf

NIRPAZ.G. 2018. *Managing Customer Success to Reduce Churn*Holded [Consulta 20 de julio de 2019] Disponible en: https://www.totango.com/

PIJAHARI.A, SISSODIA.D y VISHWAKARMA.S, 2017. Evaluation of Machine Learning Models for Employee Churn Prediction [Consulta: 27 de agosto de 2019] Disponible en: Scopus Database

RAYÓN.A. 2017. *GUÍA PARA COMENZAR CON ALGORITMOS DE MACHINE LEARNING* Deusto Data [Consulta 18 de julio de 2019] Disponible en: https://blogs.deusto.es/bigdata/guia-para-comenzar-con-algoritmos-de-machine-learning/

RIVAS.E. 2018. *¿Qué es el Data Mining o minería de datos?* IEBS School [Consulta 13 de julio de 2019] Disponible en: https://www.iebschool.com/blog/data-mining-mineria-datos-big-data/

SAS INSTITUTE. 2019 *Software y Soluciones de Analítica* [Consulta 19 de julio de 2019] Disponible en: https://www.sas.com/es_es/home.html

SABBEH.F, 2018. Machine-Learning Techniques for Customer Retention: A

Comparative Study [Consulta: 24 de agosto de 2019] Disponible en: Scopus Database

SARODE.S, SHETTY.P y VARSHA.C, 2019. Customers Churn Prediction with Rfm Model and Building a Recommendation System using Semi-Supervised Learning in Retail Sector [Consulta: 25 de agosto de 2019] Disponible en: Scopus Database

SINEXUS 2016. *Business Intelligence* [Consulta 19 de julio de 2019] Disponible en: https://www.sinnexus.com/empresa/index.aspx

SHEWAN.D. 2018. *How to Calculate (And Lower!) Your Customer Churn Rate* Wordstream [Consulta 14 de julio de 2019] Disponible en: https://www.wordstream.com/blog/ws/2014/05/12/customer-churn

TENA.R. 2018. *Las 6 estrategias de marketing más eficaces para retener a tus clientes* Holded [Consulta 19 de julio de 2019] Disponible en: https://www.holded.com/es/blog/6-estrategias-marketing-mas-eficaces-retener-clientes/

VIJAYA.J y SIVASANKAR.E, 2018. Improved churn prediction based on supervised and unsupervised hybrid data mining system [Consulta: 25 de agosto de 2019] Disponible en: Scopus Database