Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza

ISSN: 2525-1333
Vol.:4-Nro.1 (Julio-2019)

Comunicación en congreso

Aplicación de Algoritmos de Aprendizaje Automático al Análisis del Churn en Planes de Ahorro

Application of Automatic Learning Algorithms to the Churn Analysis in Savings Plans

Ariel Fernando Deroche[(1)], *Diego Basso*[(2)], *María Florencia Pollo-Cattaneo*[(3)]

[(1)] Grupo de Estudio en Metodologías de Ingeniería de Software. Facultad Regional Buenos Aires.
Universidad Tecnológica Nacional, Argentina.
arielderoche@gmail.com

[(2)] Departamento de Ingeniería e Investigaciones Tecnológicas.
Universidad Nacional de La Matanza, Argentina.
diebasso@yahoo.com.ar

[(3)] Grupo de Estudio en Metodologías de Ingeniería de Software. Facultad Regional Buenos Aires.
Universidad Tecnológica Nacional, Argentina.
Flo.pollo@gmail.com

ISSN: 2525-1333
Vol.:4-Nro.1 (Julio-2019)

Resumen:

Introduciéndonos en el competitivo mercado automotor argentino donde las estrategias para captar nuevos clientes varían, nos encontramos con una gran cantidad de bonificaciones y descuentos para que nuevos clientes se adhieran a un plan.

Sin embargo, la captación de nuevos clientes muchas veces desvía el foco de un aspecto importante, la satisfacción de los clientes actuales, ocasionando la rescisión de planes de ahorro ante la falta de estrategias específicas de retención en clientes de este segmento comercial. La predicción de clientes que pueden llegar a rescindir su plan se ha convertido en una arista de interés de estudio por parte de los concesionarios de automotores. Esto se centra en que la cartera de clientes representa uno sus mayores activos, obligando, en consecuencia, no sólo a tener que generar incentivos para aumentar el número de clientes, sino también para mantener los actuales.

En este contexto, este trabajo tiene como objetivo presentar los resultados de la aplicación de distintos algoritmos de aprendizaje automático y técnicas de explotación de datos sobre la información que un concesionario automotor de Argentina tiene de sus clientes, a fin de identificar patrones de comportamiento en aquellos con mayor probabilidad de rescindir el plan de ahorro.

Abstract:

Entering a competitive Argentine automotive market where the strategies to attract new customers vary, we find a large number of bonuses and discounts for new customers to adhere to a plan.

However, attracting new clients often diverts the focus from an important aspect, the satisfaction of current clients, causing the rescission of savings plans in the absence of specific retention strategies in clients of this commercial segment. The prediction of clients that may end up resigning their plan has become an edge of study interest on the part of automotive dealers. This is based on the fact that the client portfolio represents one of its greatest assets, obliging, therefore, not only to generate incentives to increase the number of clients, but also to maintain current ones.

In this context, this paper aims to present the results of the application of different automatic learning algorithms and techniques of data exploitation on the information that a car dealer in Argentina has of its customers, in order to identify patterns of behavior in those more likely to terminate the savings plan.

Palabras Clave: *Churn Analysis, plan de ahorro, minería de datos, árboles de decisión, Naive bayes.*

Key Words: *Churn Analysis, savings plan, data mining, decision trees, Naive bayes.*

Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza
ISSN: 2525-1333
Vol.:4-Nro.1 (Julio-2019)

I. INTRODUCCIÓN

En los últimos años, y a medida que las grandes cantidades de datos de empresas crecen exponencialmente, el concepto de aprendizaje automático o minería de datos (data mining) resulta ser esencial para comprender y proporcionar el conocimiento necesario para la toma de decisiones y el cumplimiento de los objetivos. Todo esto es posible gracias a la aplicación de diferentes algoritmos de aprendizaje automático [1, 2].

El aprendizaje automático [3] es un subcampo de las ciencias de la computación y una rama de la Inteligencia Artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender, es decir, generalizar comportamientos y conocimientos a partir de la información suministrada en forma de ejemplos. En general, un problema de aprendizaje automático utiliza un conjunto de *n* muestras de datos para intentar predecir propiedades de los datos desconocidos.

La retención de clientes, especialmente en tiempos de crisis, resulta ser la principal estrategia de negocio elegida por distintas compañías, teniendo en cuenta que el costo de obtener un nuevo interesado es 5 veces más alto que el de mantener uno existente [4, 7].

El "attrition analysis" o "Churn analysis" son técnicas de minería de datos aplicadas a predecir la fuga de clientes o, explicar las causas de por qué los compradores se van de la compañía. Retener al cliente resulta ser el objetivo principal de este tipo de análisis, para que éste no desista de los servicios o elija la competencia [6, 8, 9].

Concesionarios de planes de ahorro enfrentan una intensa competencia, deseando mantener a tantos clientes como sea posible y por otro lado captar nuevos. En este contexto, como objetivo principal del análisis tenemos la identificación de un grupo de clientes con alta probabilidad de fuga, para que la empresa tome las decisiones y acciones necesarias para retener a aquellos que considere importantes.

En Argentina, la compra de un 0 km puede realizarse de maneras diferentes: por venta convencional, plan de ahorro, créditos prendarios y/o leasing [9].

Los planes de ahorro, en su mayoría, son de 84 cuotas, teniendo en cuenta que al fraccionar el valor del vehículo en mayor cantidad de cuotas, la inversión mensual resulta menor para el consumidor y, a su vez, más accesible. Si bien los planes no cuentan con interés, su valor crece al ritmo que aumenta el precio de la unidad 0 km.

Dependiendo de la naturaleza de la compañía, el término de "customer Churn" variará, por lo que el análisis está íntimamente relacionado con la definición que la compañía le brinde al "Churn" [8, 11]. La empresa debe definir qué significa que un cliente deje sus servicios.

Por ejemplo, para el caso de los concesionarios de autos y el concepto de planes de ahorro, definirá "Churn" a la rescisión del contrato, y por ende, el fin de la prestación del plan con la misma [13].

Existen diferentes algoritmos de aprendizaje automático que se pueden aplicar al análisis del Churn [16-18, 23]. Sin embargo, la elección de uno sobre otro y su aplicación en la construcción de un modelo adecuado de predicción del Churn puede dificultarse, generando cada uno de ellos resultados diferentes.

En el contexto detallado se plantea entonces la definición del problema (sección 2), los elementos y metodología utilizada (sección 3), la solución propuesta (sección 4) y los resultados obtenidos (sección 5). Finalmente se

ISSN: 2525-1333
Vol.:4-Nro.1 (Julio-2019)

exponen las conclusiones y futuras líneas de trabajo (sección 6).

II. DEFINICIÓN DEL PROBLEMA

Considerando la información que el concesionario de autos mantiene sobre cada cliente, es posible realizar un análisis que permita relacionar distintas variables y detectar anticipadamente a aquellos con alto potencial de abandono o rescisión, estableciendo así estrategias de fidelización y/o de retención. En la actualidad, existen varias técnicas y algoritmos de aprendizaje automático que permiten abordar el problema de la fuga de clientes, o "Churn Analysis". Sin embargo, la elección de un algoritmo específico resulta ser un trabajo arduo, generando cada uno de ellos resultados diferentes. Esto dificulta además su proceso de aplicación por la no existencia de herramientas que unifiquen esta actividad [8, 13, 15].

A partir de lo enunciado y, acompañando las actuales investigaciones sobre cada uno de los algoritmos de manera individual, surge el interrogante de identificar primeramente los algoritmos que se podrían utilizar, aportando información conceptual de cada uno, seleccionando luego un subconjunto de ellos, y aplicándolos con distintas técnicas de explotación de datos al problema de la rescisión de clientes de planes de ahorro del concesionario.

Asimismo, se pretende identificar las características que describen el comportamiento de los clientes con mayor probabilidad a la rescisión del plan, los factores significativos que inciden en este comportamiento y finalmente realizar una comparación de la performance de los algoritmos utilizados en este trabajo.

III. ELEMENTOS Y METODOLOGÍA DE TRABAJO

Dentro de los algoritmos de aprendizaje automático que a menudo se utilizan para construir modelos de predicción de Churn [16-18, 23] se tienen: Árboles de Decisión, Naive Bayes, Redes Neuronales Artificiales, Regresión Logística, Reglas de Asociación, SVM o Máquinas de vectores de soporte (Support Vector Machines), Análisis de Redes Sociales (Social Network Analysis, SNA), Random Forest, Bagging y Boosting, entre otros.

Teniendo en cuenta lo descripto, se considera oportuna la aplicación de los primeros tres algoritmos mencionados a los datos de una cartera de clientes de planes de ahorro. El resultado será un modelo predictivo que permita identificar clientes o comportamientos que puedan implicar un riesgo de rescisión del plan de ahorro. La elección de estos algoritmos está estrictamente ligada a los objetivos que se plantean en este trabajo. En el caso de los árboles de decisión son una representación simple y una de las técnicas más eficaces para este tipo de problemas de clasificación y predicción. El algoritmo Naive Bayes, además de ofrecer un análisis cualitativo de los atributos y valores que pueden intervenir en el problema, permite dar cuenta de la importancia cuantitativa de esos atributos en el resultado del Churn. Y la Red Neuronal Artificial, porque además de ser uno de los algoritmos más utilizados en el campo de la Inteligencia Artificial, combinado con otras técnicas permitiría también extraer reglas de comportamiento de los clientes a partir de los valores de peso de cada neurona y comparar sus resultados con las reglas obtenidas de los árboles de decisión. La extracción de

reglas desde una red neuronal no entra dentro del alcance del presente trabajo.

Para la recolección de datos se ha utilizado la información de 17 meses de clientes de planes de ahorro de una concesionaria de autos de la República Argentina, situada en la Ciudad Autónoma de Buenos Aires (CABA). La selección de datos considerados no tuvo en cuenta todas las provincias del país, por el extenso volumen de información que implicaba su análisis y procesamiento. En la sección 4.1 se describen con mayor detalle los datos utilizados en este trabajo.

Para este trabajo se ha utilizado el software Weka [10] y Tanagra [12] a fin de comparar y evaluar las soluciones dadas por cada uno. Para su uso se requiere un archivo de formato filas-columnas que contengan en las columnas las variables a analizar y, en cada fila los valores asociados. Este archivo es obtenido del sistema oficial de la marca del concesionario, y luego de realizar el preprocesamiento de datos necesario, es ingresado en el software como *dataset* (conjunto de datos). Entre las variables definidas como columnas del archivo *dataset*, algunas de ellas serán consideradas como atributos explicativos (denominados input), y otro atributo se seleccionará como la clase a predecir.

Se han utilizado los mismos datos para probar con todos los algoritmos. Para realizar las pruebas se ha aplicado validación cruzada de 10 iteraciones. La validación cruzada o cross-validation es una técnica que se aplica para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre las diferentes particiones, con el objetivo de estimar la exactitud (o precisión) del modelo al llevarlo a la práctica.

Como estimador de la bondad de los modelos obtenidos se han utilizado las métricas de [27]: *Exactitud*, que muestra mejores resultados cuando más se acerca a 1; *Tasa de Errores*, que garantiza que el modelo tiene la capacidad de predecir mejor nuevos casos cuando su valor se acerca a 0; *Cobertura de Regla*, que expresa mejor calidad y utilidad de la regla cuando su valor tiende a 1 y *Grado de Incidencia del valor de un Atributo* que aporta más significancia cuando su valor tiende a 1.

IV. SOLUCIÓN PROPUESTA

A continuación, se presenta la descripción y preparación de los datos utilizados, se realiza una explicación conceptual de los algoritmos propuestos y se aplican los mismos utilizando algunas técnicas de explotación de datos, con el objetivo de responder a los interrogantes planteados en la Definición del Problema.

V. DESCRIPCIÓN DE LOS DATOS INICIALES

Para el desarrollo del trabajo se han utilizado 15.004 registros asociados a clientes que poseen planes de ahorro, correspondiente al período Diciembre 2016 - Abril 2018. Se consideraron clientes de todo el territorio nacional, exceptuando la provincia de Buenos Aires y CABA por el gran volumen de datos que representaban. Por motivos de confidencialidad, los datos personales de los clientes de planes de ahorro no pudieron ser incluidos ya que requerían autorización especial de la gerencia del concesionario, los cuales no se obtuvieron hasta avanzado el trabajo de investigación. En base a la experiencia y al conocimiento del analista del negocio, es posible identificar los atributos relevantes para construir

un modelo predictivo. En primera instancia se logró identificar un conjunto de 9 atributos iniciales que pueden explicar el comportamiento del cliente: Provincia, Fecha Inicio, Vehículo, Medio de Pago, Deuda Moroso, Cuotas Pagas, Licita-Cancela, Cuotas a Vencer y Estado del Plan.

VI. PREPARACIÓN DE LOS DATOS

Previo al procesamiento de los datos, los mismos se integraron en una única tabla y se corrigieron las inconsistencias que presentaban. Se identificaron valores faltantes, los cuales se completaron en base al análisis de los datos mensuales anteriores y posteriores al estudiado. Se aplicó transformación de datos, estandarización de valores sobre un rango y se incluyeron nuevos atributos, indicados en la TABLA 1.

Atributo	**Descripción**	**Tipo de Atributo**
Región	Es la región geográfica donde se encuentra la provincia	Categórico ó Nominal
Mes Inicio	Mes de inicio del plan	Numérico
Año Inicio	Año de inicio del plan	Numérico
Modelo	Clasificación que permite agrupar los vehículos de un plan de ahorro	Categórico ó Nominal

TABLA 1. Descripción de nuevos atributos

Luego de la limpieza de los datos, se obtuvo una distribución balanceada entre clientes que rescindieron el plan (53,7%) de los que no lo hicieron (46,3%). Posteriormente y utilizando una función aleatoria se selecciona un 70% de los registros para datos de entrenamiento y los restantes como datos de prueba del modelo de predicción del Churn obtenido. En la Tabla 3 se muestra una lista parcial de los datos usados.

VII. ALGORITMOS A UTILIZAR

La elección de los algoritmos a utilizar, además de lo ya fundamentado en la sección 3, se basa también en diferentes investigaciones que señalan su utilización más frecuente en problemas de predicción del Churn [14]. Se ha relevado la base de datos científica Science Direct [16] enfocándose en los trabajos de los últimos diez años. Asimismo, se han recolectado investigaciones utilizando el buscador académico Google Scholar y el buscador académico de la Universidad Tecnológica Nacional (UTN) [17, 18].

VIII. APLICACIÓN DE ALGORITMOS Y TÉCNICAS

Para identificar las características que describen el comportamiento de los clientes que rescinden el plan de ahorro se utilizaron los árboles de decisión C4.5 [24] y J48 [10] sobre el dataset que contiene los datos de entrenamiento descriptos en la sección 4.1 y 4.2. Este dataset se cargó en las herramientas Tanagra y Weka, configurando como atributo objetivo del Churn al "Estado del Plan", que toma los valores Rescindido o No Rescindido, y los restantes atributos como entrada de los algoritmos. Asimismo, se estableció un tamaño mínimo de 5 hojas y nivel de confianza del 25%.

De esta configuración y luego de la etapa de entrenamiento y prueba de los modelos, se obtuvieron 2 reglas generales que describen el comportamiento general de los clientes que rescinden el plan. Estas reglas se describen en la Tabla 6 de la siguiente sección.

Luego, para identificar si los clientes que rescinden el plan tienen características homogéneas,

independientemente de la región geográfica en la que viven, se seleccionó en primer lugar a los clientes por su región de pertenencia (Centro, Cuyo, Litoral, Norte y Patagonia), sobre los que aplicaron los mismos algoritmos C4.5 y J48. En este análisis, se obtuvieron 2 reglas que describen el comportamiento de los clientes que rescinden el plan en la región Centro y Cuyo, respectivamente, y 3 reglas para las regiones Litoral, Norte y Patagonia. Las mismas se describen en las Tablas 7 a 11 de la siguiente sección.

Además de identificar las características generales y regionales que describen a los clientes que rescinden el plan de ahorro, resulta de interés poder ponderar aquellas características más significativas que inciden sobre este resultado del Churn. Para ello, se aplicó una técnica que aplica redes bayesianas con el algoritmo Naive Bayes [22]. En este caso, se configuró como atributo objetivo del Churn al "Estado del Plan", y los restantes atributos como entrada del algoritmo previa discretización de aquellos de tipo continuo.

El último de los modelos construidos para la predicción de la rescisión de planes de ahorro está basado en la aplicación de redes neuronales artificiales. La topología de la red adoptada es un Perceptrón de tres capas: la de entrada, oculta y capa de salida. Las variables para la capa de entrada corresponden a los atributos indicados en las Tablas 1 y 2, mientras que la capa de salida corresponde al "Estado del Plan", objetivo del Churn. Debido a que una sola capa oculta es suficiente para la mayoría de los problemas [25, 26], la topología del Perceptrón se configuró de esta manera, comenzando con dos a seis neuronas para generar el menor error y evitar el sobreentrenamiento (overfitting).

IX. RESULTADOS OBTENIDOS

A continuación se muestran los resultados obtenidos de la aplicación de cada uno de los algoritmos y técnicas propuestas, dando respuesta a los interrogantes planteados en la sección 2.

X. ÁRBOLES DE DECISIÓN

Para el estudio del mejor árbol de decisión se consideraron los algoritmos C4.5 y J48. La selección del mejor modelo de predicción para la rescisión de los planes de ahorro se efectuó basándose en los datos de prueba. Las Tablas 4 y 5 muestran la información y exactitud predictiva de cada algoritmo aplicado.

Predicción del Estado de Rescisión			
	Rescindido	No Rescin.	Tasa de Error
Para casos de entrenamiento			
Rescindido	5563	46	0,82%
No Rescin.	14	4880	0,28%
Para casos de prueba			
Rescindido	2417	30	1,22%
No Rescin.	7	2047	0,34%

TABLA 4. Información Predictiva con C4.5

Predicción del Estado de Rescisión			
	Rescindido	No Rescin.	Tasa de Error
Para casos de entrenamiento			
Rescindido	5528	81	1,44%
No Rescin.	25	4869	0,51%
Para casos de prueba			
Rescindido	2410	37	1,51%
No Rescin.	4	2050	0,19%

TABLA 5. Información Predictiva con J48

Como puede deducirse de las Tablas 4 y 5, el modelo predijo los datos de entrenamiento con el 99,42% (C4.5)

Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza
ISSN: 2525-1333
Vol.:4-Nro.1 (Julio-2019)

y 98,99% (J48) de exactitud y los datos de prueba con un 99,17% (C4.5) y 99,08% (J48). Asimismo, el modelo predice los planes en estado "Rescindido" con una tasa de aciertos del 99,18% (C4.5) y 98,56% (J48) en la etapa de entrenamiento y de 98,78% (C4.5) y 98,49% (J48) durante la de prueba, mientras que los planes "No Rescindidos" los predice con un 99,72% (C4.5) y 99,49% (J48) de aciertos en la etapa de entrenamiento y con 99,66% (C4.5) y 99,81% (J48) en la de prueba. Por consiguiente, si bien ambos modelos presentan muy buena capacidad de predicción y generalización, se considera que el algoritmo J48 resultó ser levemente superior al momento de predecir los planes de ahorro que pueden resultar rescindidos.

Por otra parte, de la aplicación del algoritmo J48, se obtuvieron 2 reglas que describen el comportamiento general de los clientes que pueden rescindir sus planes, las cuales se detallan en la Tabla 6:

R_1	Si Licita_Cancela = 0 y Medio de Pago = 'DESCONOCIDO' y Deuda Moroso = 0 y Cuotas Pagas <= 10 y Cuotas a Vencer <= 73
R_2	Si Licita_Cancela = 0 y Medio de Pago = 'DESCONOCIDO' y Deuda Moroso = 0 y 10 < Cuotas Pagas <= 18 y Cuotas a Vencer <= 65

TABLA 6. Clientes que rescinden el plan de ahorro

A partir de las reglas identificadas se describe el comportamiento general de estos clientes:

Regla 1: Los planes que se rescinden son de clientes que no licitaron ni cancelaron su plan y que tampoco presentan mora en el pago, abonaron a lo sumo 10 cuotas y le restan pagar menos de 73. En este caso, se desconoce el medio de pago (pudiendo ser tarjeta o débito automático). Esta regla da una cobertura del 82% de los casos analizados.

Regla 2: Los planes que se rescinden son de clientes que no licitaron ni cancelaron su plan, tampoco presentan mora en el pago, abonaron entre 10 y 18 cuotas y le restan pagar menos de 65. En este caso, también se desconoce el medio de pago elegido por el cliente. Esta regla da una cobertura del 7% de los casos analizados.

Del análisis realizado se observa que cuando el cliente pagó menos de 10 cuotas tiene mayor probabilidad de rescindir el plan, disminuyendo esta situación a medida que abona más cuotas. Asimismo, el 99% de los planes que se abonan con medios de pago "tarjeta" o "débito automático" son de clientes que no rescinden el plan, siendo este un indicador importante a tener en cuenta. Con el objetivo de identificar alguna otra regla relevante, se aplicó el algoritmo C4.5, obteniéndose exactamente las mismas reglas generales que con el algoritmo J48.

Respecto al análisis realizado por región, luego de aplicar el algoritmo J48 se obtuvieron las siguientes reglas que describen el comportamiento de los clientes que rescinden el plan, detalladas en las Tablas 7 a 11:

R_1	Si Licita_Cancela = 0 y Medio de Pago = 'DESCONOCIDO' y Deuda Moroso = 0 y Cuotas Pagas <= 9
R_2	Si Licita_Cancela = 0 y Medio de Pago = 'DESCONOCIDO' y Deuda Moroso = 0 y 9 < Cuotas Pagas <= 26

Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza
ISSN: 2525-1333
Vol.:4-Nro.1 (Julio-2019)

TABLA 7. Reglas de comportamiento - Región Centro

A partir de las reglas descubiertas en la Tabla 7 se identifica que los clientes de la región Centro con mayor predominancia a rescindir el plan de ahorro, son aquellos que abonaron a lo sumo 9 cuotas, no licitaron ni cancelaron el plan y no tienen mora en el pago. Esta regla da una cobertura del 87,6% sobre los clientes que rescinden los planes en esta región. Otra posibilidad, aunque con una cobertura bastante menor del 9,6%, incluye a los clientes que abonaron entre 9 y 26 cuotas.

En cuanto a los clientes de la región de Cuyo, a partir de las reglas descubiertas en la Tabla 8 se identifica que los de mayor tendencia a rescindir el plan son aquellos que abonaron a lo sumo 10 cuotas, no licitaron ni cancelaron el plan y tampoco tienen mora en el pago. Esta regla da una cobertura del 90,8% sobre los clientes que rescinden los planes en esta región. Otra regla, aunque con una cobertura menor del 9,6%, incluye a los clientes que comenzaron el plan antes del año 2015 y abonaron entre 10 y 28 cuotas.

Respecto a las reglas descubiertas en la Tabla 9 se identifica que los clientes de la región Litoral con mayor tendencia a rescindir el plan, son aquellos que abonaron a lo sumo 11 cuotas, no licitaron ni cancelaron el plan y no tienen mora en el pago. Esta regla da una cobertura del 82,7% sobre los clientes que rescinden los planes en esta región. Otras reglas, aunque con una cobertura del 7,4% y 5,4% respectivamente, incluyen a los clientes que tienen pagas a lo sumo 4 cuotas y le restan pagar más de 72 o bien que habiendo pagado entre 11 y 28 cuotas comenzaron el plan antes del año 2016.

R_1	Si Licita_Cancela = 0 y Medio de Pago = 'DESCONOCIDO' y Deuda Moroso = 0 y Cuotas Pagas <= 10
R_2	Si Licita_Cancela = 0 y Medio de Pago = 'DESCONOCIDO' y Deuda Moroso = 0 y 10 < Cuotas Pagas <= 28 y Año Inicio <= 2015

TABLA 8. Reglas de comportamiento - Región Cuyo

R_1	Si Licita_Cancela = 0 y Medio de Pago = 'DESCONOCIDO' y Deuda Moroso = 0 y Cuotas Pagas <= 11 y Cuotas a Vencer <= 72
R_2	Si Licita_Cancela = 0 y Medio de Pago = 'DESCONOCIDO' y Deuda Moroso = 0 y Cuotas Pagas <= 4 y Cuotas a Vencer > 72
R_3	Si Licita_Cancela = 0 y Medio de Pago = 'DESCONOCIDO' y Deuda Moroso = 0 y 11 < Cuotas Pagas <= 28 y Año Inicio <= 2015

TABLA 9. Reglas de comportamiento - Región Litoral.

De la Tabla 10, correspondiente a la región Norte se obtienen 3 reglas que describen el comportamiento de los clientes con mayor predominancia a rescindir el plan. Por un lado, se identifica que quienes pagaron a lo sumo una cuota son los más propensos a rescindir el plan, dando una cobertura del 41,4% sobre el total de clientes de esta región. Por otra parte, y con una cobertura similar a la regla anterior, se presenta la posibilidad de rescisión en clientes que, habiendo abonado entre 2 y 6 cuotas, no licitaron ni cancelaron el plan, ni tampoco tienen mora

Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza
ISSN: 2525-1333
Vol.:4-Nro.1 (Julio-2019)

en el pago. Otra posibilidad, aunque menor, incluye también a los clientes que comenzaron el plan de ahorro entre los años 2014 y 2016 y abonaron entre 6 y 15 cuotas. Esta última situación presenta una cobertura del 12% sobre los clientes de esta región.

Por la región Patagonia, tal como se observa en la Tabla 11, también se obtienen 3 reglas que describen el comportamiento de los clientes con tendencia a rescindir el plan. Por un lado, se identifica a aquellos clientes que a lo sumo pagaron una cuota, otorgando una cobertura del 38,9% sobre el total de clientes de esta región. Además, y con una cobertura del 44%, se tiene a los clientes que, habiendo abonado entre 2 y 9 cuotas, no licitaron ni cancelaron el plan, tampoco tienen mora en el pago y le restan abonar 74 cuotas. Como última posibilidad, aunque en menor proporción de cobertura (10,5%), se identifica a los clientes que pagaron entre 9 y 24 cuotas, adeudan hasta 74 e iniciaron el plan antes de 2016.

R_1	Si Cuotas Pagas <= 1
R_2	Si Licita_Cancela = 0 y Medio de Pago = 'DESCONOCIDO' y Deuda Moroso = 0 y 1 < Cuotas Pagas <= 6
R_3	Si Licita_Cancela = 0 y Medio de Pago = 'DESCONOCIDO' y Deuda Moroso = 0 y 6 < Cuotas Pagas <= 15 y Año Inicio <= 2016

TABLA 10. Reglas de comportamiento - Región Norte

R_1	Si Cuotas Pagas <= 1
R_2	Si Licita_Cancela = 0 y Deuda Moroso = 0 y 1 < Cuotas Pagas <= 9 y Cuotas a Vencer <= 74
R_3	Si Medio de Pago = 'DESCONOCIDO' y 9 < Cuotas Pagas <= 24 y Cuota a Vencer <= 74 y Año Inicio <= 2015

TABLA 11. Reglas de comportamiento - Región Patagonia

XI. NAIVE BAYES

Para el estudio de la red bayesiana se aplicó el algoritmo Naive Bayes utilizado el software Weka y Tanagra. Ambas herramientas presentaron los mismos resultados sobre los datos de entrenamiento y prueba. En la Tabla 12 se muestra la información y exactitud predictiva del algoritmo aplicado.

Predicción del Estado de Rescisión			
	Rescindido	No Rescin.	Tasa de Error
Para casos de entrenamiento			
Rescindido	5504	105	1,87%
No Rescin.	*919*	3975	18,77%
Para casos de prueba			
Rescindido	2388	59	2,41%
No Rescin.	*388*	1666	18,88%

TABLA 12. Información Predictiva con Naive Bayes

Como se observa en la Tabla anterior, el modelo predijo los datos de entrenamiento con el 90,25% de exactitud y los de prueba con un 90,07%. También, que el modelo predice los planes en estado "Rescindido" con una tasa de aciertos del 98,13% en la etapa de entrenamiento y del 97,59% durante la de prueba, mientras que los planes "No Rescindidos" los predice con una tasa del 81,23% en la etapa de entrenamiento y con 81,12% en la de prueba. Si bien el modelo tiene buena capacidad de predicción y generalización, tiende a confundir los planes que no fueron rescindidos como que sí van a

serlo. No obstante, desde el punto de vista del impacto para el concesionario no resultará alto, ya que el cliente en definitiva no rescindirá el plan de ahorro, siendo simplemente un indicador de "atención".

Por otra parte, para identificar el o los factores de mayor incidencia sobre los planes rescindidos, se tiene en cuenta el árbol de ponderación obtenido de aplicar una red bayesiana con el algoritmo Naive Bayes. Este árbol muestra en qué medida la variación de los valores de cada atributo incide en el estado de "Rescisión" de un plan.

Del análisis de atributos significativos se observa que los de mayor incidencia sobre el estado de un plan "Rescindido" son: la REGION, donde el valor "Litoral" da una incidencia del 42,3%; el MODELO de autos "Compacto" con una significancia del 57,3%; el atributo DEUDA MOROSO con una incidencia del 99,8% sobre los clientes que tienen hasta 5 cuotas en mora; el atributo CUOTAS PAGAS con una significancia del 86,5% en aquellos planes que pagaron hasta 9 cuotas y el atributo LICITA_CANCELA con una incidencia del 99,8% cuando no se superan las 8 cuotas licitadas/canceladas. En el caso del atributo CUOTAS AVENCER, si bien tiene incidencia en el resultado de la rescisión de un plan, el comportamiento queda extrapolado entre los que adeudan menos de 16 cuotas y más de 50. Los restantes atributos del algoritmo no son significantes. No obstante, se mencionan las conclusiones sobre la incidencia que los mismos aportan al modelo de predicción del Churn construido.

Las 22 provincias dan una incidencia entre 1,21% y 9%. Las más relevantes son Chaco (6,99%), Neuquén (6%), Corrientes (7,56%), Entre Ríos (7,76%), Mendoza (7,18%) y Santa Fe (9,04%). Ninguna tiene más significancia sobre otra, por lo que no se consideró un atributo significativo en el modelo.

Todos los meses tienen una incidencia entre 6,35% y 10,8%, por lo que tampoco resultan significativos.

Los años de mayor incidencia son el 2016 (29,51%) y 2017 (18,95%). Los restantes dan una incidencia entre 0,1% y 12,9% sobre el estado de rescisión del plan. Este atributo tampoco es significativo ya que corresponde a los valores de la muestra utilizada en el trabajo.

XII. REDES NEURONALES ARTIFICIALES

Para el estudio de la mejor Red Neuronal Artificial (RNA) se consideró una topología de dos a seis nodos en la capa oculta, obteniéndose seis modelos. La selección del mejor modelo se efectuó basándose en los datos de prueba. Los resultados de la búsqueda de la mejor RNA se muestran en la Tabla 13.

Capa Oculta	**Error del Modelo**	
	Entrenamiento	**Prueba**
2 nodos	**0,94%**	**1,16%**
3 nodos	1%	1,18%
4 nodos	1,1%	1,58%
5 nodos	1,05%	1,31%
6 nodos	1,23	2,09

TABLA 13. Resultados de la mejor red neuronal

En esta tabla se puede observar que la tasa de clasificaciones correctas del estado de todos los planes de ahorro es del 99,06% con un modelo de 2 nodos en la capa oculta, del 99% con el modelo de 3 nodos, mientras que con un número mayor de nodos la tasa de clasificaciones correctas desciende por debajo de este valor. Además, se resalta en negrita el mejor modelo de RNA seleccionado. La selección se basa en una mejora

por debajo de la primera cifra significativa en el porcentaje de clasificación correcta.

En la Tabla 14 se muestra la información y exactitud predictiva de la Red Neuronal Artificial, la cual utiliza un perceptrón de 2 nodos en la capa oculta. Como se observa, el modelo predijo los datos de entrenamiento con el 99,06% de exactitud y los de prueba con un 98,84%. También, que el modelo predice los planes en estado "Rescindido" con una tasa de aciertos del 98,56% durante la etapa de entrenamiento y del 98,53% durante la de prueba, mientras que los planes "No Rescindidos" los predice con una tasa del 99,64% en la etapa de entrenamiento y del 99,23% en la de prueba.

Predicción del Estado de Rescisión (2 nodos)			
	Rescindido	No Rescin.	Tasa de Error
Para casos de entrenamiento			
Rescindido	5528	81	1,44%
No Rescin.	18	4876	0,36%
Para casos de prueba			
Rescindido	2411	36	1,47%
No Rescin.	16	2038	0,77%

. **TABLA 14. Información Predictiva con Red Neuronal.**

XIII. CONCLUSIONES

Los concesionarios de autos necesitan analizar diferentes indicadores que contribuyan a la toma de decisiones a mediano y largo plazo con el fin de implementar estrategias que mejoren su eficiencia y posicionamiento en un mercado muy competitivo.

La explotación de información, el aprendizaje automático y la minería de datos son fundamentales para proporcionar patrones de conocimiento de poblaciones de datos basándose en el análisis y exploración de los mismos, conjuntamente con la aplicación de procesos, técnicas y algoritmos a utilizar.

El análisis comparativo realizado a los resultados obtenidos muestra que el algoritmo de árboles de decisión resultó ser el mejor modelo de predicción del Churn para los planes de ahorro, con una exactitud de clasificación del 99,08% y una tasa de error del 0,2% en la predicción de planes rescindidos y no rescindidos. Además, permitió la identificación de reglas de comportamiento que pueden presentar los clientes con intenciones de rescindir un plan.

En segundo lugar ubicamos a la Red Neuronal, con una exactitud del modelo obtenido del 98,84% y una tasa aciertos del 98,53% y 99,23% en la predicción de los planes rescindidos y no rescindidos, respectivamente. A pesar del desempeño satisfactorio alcanzado, con frecuencia estos modelos son criticados en la medida que se consideran cajas negras que no permiten hacer inferencias acerca de la manera en que las variables de la capa de entrada afectan a los resultados del modelo, en este caso la predicción del Churn.

Por último se ubica el algoritmo Naive Bayes, que presenta una exactitud del modelo del 90,07% y una tasa de aciertos de planes rescindidos y no rescindidos del 97,59% y 81,12%, respectivamente. No obstante, este último algoritmo facilitó la identificación de los atributos más significativos que influyen en la posibilidad de que un cliente rescinda un plan.

Teniendo en cuenta lo descripto, en el trabajo se pudo identificar que la principal alerta que un cliente manifiesta en la fase previa del abandono del plan es haber pagado menos de 12 cuotas de un plan para la compra de un automóvil del tipo familiar o compacto.

Ante esta situación, los concesionarios pueden estar atentos a la proximidad de este número de cuotas, realizando diferentes alternativas de marketing o similares para los clientes.

Acorde a lo expuesto, como futuras líneas de desarrollo se propone por un lado la entrega al concesionario de toda esta información a fin de poder realizar distintas estrategias de contacto y/o citas para revertir el Churn. Por otro lado, se plantea incorporar los datos personales y socio-económicos de los clientes, que no pudieron incluirse en esta etapa del trabajo, a fin de ajustar los modelos y validar la existencia de otras características a tener en cuenta. Asimismo, se propone actualizar la evolución de los actuales planes proyectados a la fecha, estudiando la injerencia de factores micro o macroeconómicos sobre los planes de ahorro. Por último, se plantea el interés de aplicar las Redes Neuronales con técnicas que permitan identificar reglas de comportamiento sobre clientes con posibilidad de rescindir planes de ahorro y extender la aplicación a modelos que combinen varios algoritmos de clasificación.

XVI. REFERENCIAS Y BIBLIOGRAFÍA

[1] *Pollo-Cattaneo. M., Pytel, P., García-Martínez, R., Vegega, C., Amatriain, H., Ramón, H., Mansilla, D., Deroche, A., Cigliuti, P., Saavedra-Martínez, P., Garbarini, R., Rodriguez, D., Britos, P., Tomasello,M. (2013). Prácticas y Aplicaciones de Ingeniería de Requisitos en Proyectos de Explotación de Información. Proceedings del XV Workshop de Investigadores en Ciencias de la Computación, Pág. 171-175. ISBN 978-9-872-81796-1.*

[2] *Barrientos, F., Ríos, S.A. (2013). Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicacione. Universidad de Chile, Santiago, Chile.*

[3] *B. Lantz (2015). "Machine Learing with R Second Edition". Packt Publishing Ltd.*

[4] *He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). Prediction of customer attrition of commercial banks based on SVM model. Procedia Computer Science, 31, 423-430.*

[5] *Chiang, D. A., Wang, Y. F., Lee, S. L., & Lin, C. J. (2003). Goal-oriented sequential pattern for network banking Churn analysis. Expert Systems with Applications, 25(3), 293-302.*

[6] *Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer Churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications policy, 30(10), 552-568.*

[7] *Keramati, A., & Ardabili, S. M. (2011). Churn analysis for an Iranian mobile operator. Telecommunications Policy, 35(4), 344-356.*

[8] *Mutanen, T. (2006). Customer Churn analysis–a case study. Journal of Product and Brand Management, 14(1), 4-13.*

[9] *ACARA (Asociación de Concesionarios de automotores de la República Argentina),* www.acara.org.ar*, Último acceso Julio de 2018.*

[10] *Ian H. Witten, Eibe Frank (1999). Data Mining: Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems. ISBN 978-0-12-374856-0.*

ISSN: 2525-1333
Vol.:4-Nro.1 (Julio-2019)

[11] Kisioglu, P., & Topcu, Y. I. (2011). Applying Bayesian Belief Network approach to customer Churn analysis: A case study on the telecom industry of Turkey. Expert Systems with Applications, 38(6), 7151-7157.

[12] Rakotomalala, R. (2005). Tanagra: data mining software for academic and research purposes. in Actes de EGC'2005, RNTI-E-3, vol. 2, pp. 697-702.

[13] Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer Churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications policy, 30(10), 552-568.

[14] AManso, F. (2015) Análisis de Modelos de Negocios basados en Big Data para Operadores móviles. Universidad de San Andrés. Tesis de Maestría en Gestión de Servicios Tecnológicos y Telecomunicaciones (goo.gl/d88nV2).

[15] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. CRISP-DM 1.0 Step-by-step Data Mining Guide. http://tinyurl.com/crispdm, 2000 (accessed 02.05.17).

[16] Science Direct, http://www.sciencedirect.com/, Último acceso Julio de 2018.

[17] Google Académico, https://scholar.google.com.ar/, Último acceso Julio de 2018.

[18] Bibliotecas Electrónicas UTN, http://portal.bibliotecas.utn.edu.ar/proxy/, Último acceso Julio de 2017.

[19] Britos, P. V., & Britos, P. V. (2005). Minería de datos basada en sistemas inteligentes. Nueva Librería.

[20] Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics, 35(5), 352-359.

[21] Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer Churn prediction. Simulation Modelling Practice and Theory, 55, 1-9.

[22] Britos, P. (2008). Procesos de Explotación de Información basados en Sistemas Inteligentes. Tesis Doctoral. Universidad Nacional de La Plata. Facultad de Informática. Argentina.

[23] Manso, F. (2015). Análisis de Modelos de Negocios Basados en Big Data para Operadores Móviles. Tesis de Magíster en Gestión de Servicios Tecnológicos y de Telecomunicaciones. Universidad de San Andrés.

[24] Quinlan, J. R. (1993). C4.5: Programsfor Machine Learning. Morgan KaufmannPublishers.

[25] Hegazy, T.; Fazio, P. and Moselhi, O. (1994). "Developing practical neural network applications using backpropagation". Microcomputers in Civil Engineering, vol. 9, No. 2 (March), pp. 145-159

[26] Palisade. "Guía para el uso de NeuralTools: Programa auxiliar de redes neuronales para Microsoft® Excel Versión 5.7". Ithaca, NY: Palisade Corporation, 2010. 110 p.

[27] Basso, D. (2014). Propuesta de Métricas para Proyectos de Explotación de Información. Revista Latinoamericana de Ingeniería de Software, 2(4): 157-218, ISSN 2314-2642.

ISSN: 2525-1333
Vol.:4-Nro.1 (Julio-2019)

Recibido: 2019-02-26
Aprobado: 2019-06-16
Datos de edición: Vol. 4 - Nro. 1 - Art. 5
Fecha de edición: 2019-06-28
URL: http://reddi.unlam.edu.ar