DOI: 10.47460/uct.v24i106.399

PREDICCIÓN DE FUGA DE CLIENTES EN UNA EMPRESA DE DISTRIBUCIÓN DE GAS NATURAL MEDIANTE EL USO DE MINERÍA DE DATOS

Troncoso Espinosa, Fredy Humberto[1], Ruiz Tapia, Javiera Valentina[2]
ftroncos@ubiobio.cl[1], jruiz@idgterragis.cl[2]
Departamento de Ingeniería Industrial, Facultad de Ingeniería, Universidad del Bío-Bío, Concepción-Chile[1]
Empresa Gas Sur S.A[2].

Concepción-Chile[1]
Talcahuano-Chile[2]

Recibido (07/10/20), Aceptado (22/10/20)

Resumen: La fuga de clientes es un problema relevante al que enfrentan las empresas de servicios y que les puede generar pérdidas económicas significativas. Identificar los elementos que llevan a un cliente a dejar de consumir un servicio es una tarea compleja, sin embargo, mediante su comportamiento es posible estimar una probabilidad de fuga asociada a cada uno de ellos. Esta investigación aplica minería de datos para la predicción de la fuga de clientes en una empresa de distribución de gas natural, mediante dos técnicas de machine learning: redes neuronales y support vector machine. Los resultados muestran que mediante la aplicación de estas técnicas es posible identificar los clientes con mayor probabilidad de fuga para tomar sobre estas acciones de retención oportunas y focalizadas, minimizando los costos asociados al error en la identificación de estos clientes.

Palabras Clave: Fuga de clientes, Minería de datos, Machine learning, Distribución de gas natural.

PREDICTING CUSTOMER CHURN IN A DISTRIBUTION COMPANY OF NATURAL GAS USING DATA MINING

Abstract: Customer churn is a relevant problem faced by service companies and that can generate significant economic losses. Identifying the elements that lead a customer to stop consuming a service is a complex task. However, through their behavior, it is possible to estimate a churn probability associated with each one of them. This research applies data mining to predict customer churn in a natural gas distribution company, using two machine learning techniques: neural networks and support vector machine. The results show that by applying these techniques it is possible to identify customers with the highest probability of churn to take retention actions timely and focused, minimizing the costs associated with the error in the identification of these customers.

Keywords: Customer churn, Data mining, Machine learning, Natural gas distribution.

I.INTRODUCCIÓN

El área comercial y de servicio al cliente tienen el rol fundamental de poder establecer relaciones comerciales para atraer y retener clientes mediante campañas de marketing. Sin embargo, retener un cliente resulta más rentable para una empresa que atraer uno nuevo [1]. Esta rentabilidad se debe a que un nuevo cliente implica altos gastos operacionales y gastos asociados a acuerdos comerciales, y se trasforma en una fuente de beneficios para la empresa cuando comienza a consumir el servicio contratado y otros servicios derivados.

Según [2], la fuga de un cliente es la acción de cancelar el servicio prestado por la compañía. La fuga puede ser por voluntad propia o bien por que la empresa cancela el contrato. Lo esperado es que el cliente deje de consumir una vez que renuncia al servicio contratado, sin embargo, existen algunos que dejan de consumir el servicio durante un periodo determinado antes de renunciar.

Determinar la propensión de un cliente a la fuga resulta de gran ayuda para enfocar los recursos destinados a la retención de aquellos con mayor propensión. La retención requiere implementar medidas mitigación para que el cliente no renuncie y no deje de consumir los servicios.

Gas Sur S.A. es una empresa que se dedica a la distribución de Gas Natural por tuberías para el sector residencial y comercial en las ciudades de Concepción y Los Ángeles en Chile [3]. Un cliente se fuga para optar por otro servicio de distribución de gas por tuberías, por gas en otros formatos o por otro tipo de combustible. Actualmente las acciones de retención se aplican cuando un cliente deja de consumir en un periodo de 6 meses o menos, luego de esto se considera fugado.

Gas Sur S.A no cuenta con un modelo que permita predecir la de fuga de clientes, por lo que las acciones de retención son tardías, se realiza en base a la experiencia y sin estar fundamentado en un patrón de comportamiento del cliente. Contar con un modelo que permita predecir la fuga de clientes es un elemento esencial de los actuales sistemas de gestión de relación con los clientes, Customer Relationship Management (CRM) [4].

Una herramienta muy utilizada en la predicción de fuga de clientes es la minería de datos mediante técnicas de machine learning. Minería de datos ha sido ampliamente utilizada para esto en la industria de telecomunicaciones [5] [6] e instituciones financieras [7]. También ha sido utilizada para analizar y predecir la fuga de clientes en otros ámbitos como marcas [8] y servicios web [9]. Dentro de las técnica de machine learning más utilizada para la predicción de fuga están las redes neuronales, support vector machine, naive bayes, árbol de decisión and logistic regression analysis [10] y random forest [11].

Dado los buenos resultados obtenidos se propone la utilización de minería de datos para la predicción de la fuga de clientes en el consumo de gas natural. Gas Sur S.A. posee un sistema que registra la información demográfica, de consumo y reclamos de cada cliente la cual será utilizada para este fin. Se utilizarán diversas técnicas de machine learning y se determinará aquella con el mejor desempeño para ser implementada. La elección de la mejor técnica permitirá identificar de mejor manera los clientes con mayor probabilidad de fuga, tomar medidas de retención en forma oportuna y minimizar los costos asociados al error en la clasificación.

II.METODOLOGÍA

Para identificar las variables más influyentes en la fuga de clientes y obtener un modelo predictivo, se utilizó la metodología Knowledge Discovery in Databases KDD [12]. Este es un proceso que en forma iterativa explora grandes volúmenes de datos para determinar patrones, como se muestra en la Figura 1. Esta metodología está compuesta por cinco etapas [13]:

•Selección de datos, donde se determinan las fuentes de datos y el tipo de información a utilizar.

•Pre procesamiento de la base de datos, con el fin de tener información más confiable, la que aporte mayor valor a la predicción. Esta etapa incorpora el análisis de datos faltantes, de datos inconsistentes, y el análisis de datos fuera de rango.

•Transformación y selección de variables, que engloba cualquier proceso que modifique la forma de los datos para generar nuevas variables, que enriquezcan la información y obtener un mejor patrón. Luego se procede a seleccionar aquellas variables más importantes.

•Minería de datos, en la cual se aplican las técnicas que permitirán extraer el patrón relevante desde los datos, como las técnicas de machine learning o algoritmos de clasificación. Las técnicas de machine learning aprenden el patrón general oculto en los datos y luego lo utilizan para generar una predicción. La predicción consiste en asignar un registro u observación a una clase o grupo previamente definido.

• Interpretación y evaluación, donde se interpretan los patrones de datos que se descubrieron y se evalúa el impacto del modelo en su futura implementación.

ISSN 2542-3401/ 1316-4821 UNIVERSIDAD, CIENCIA y TECNOLOGÍA *Vol. 24, Nº 106 Noviembre 2020 (pp. 79-87)*

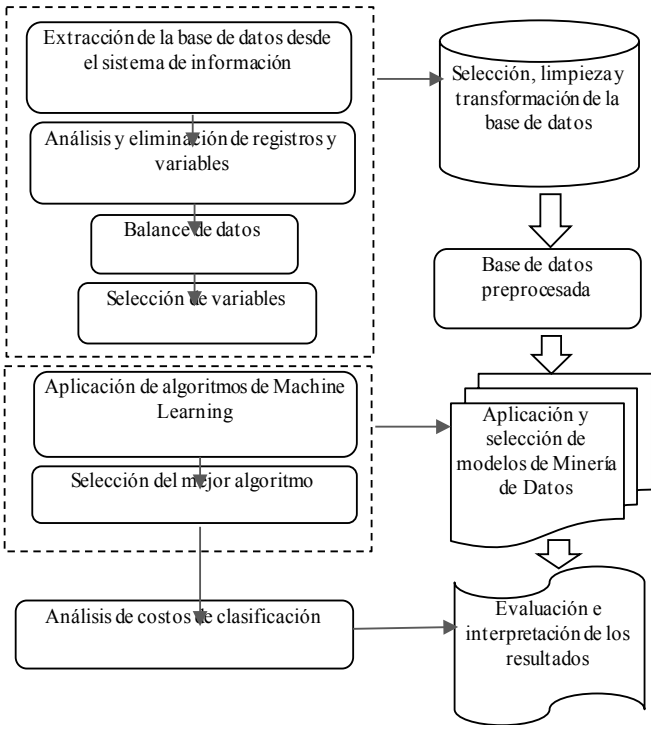

Fig. 1. Procesos de datos dentro de la metodología Knowledge Discovery in Databases KDD

III.DESARROLLO

A.Selección de datos

Los datos proporcionados por Gas Sur S.A comprenden el registro demográfico, de contrato, de reclamos y servicio al cliente, y datos asociados al consumo para un total de 36.484 clientes. Los registros de consumo considerados van desde enero de 2018 hasta febrero de 2020. La información asociada a cada cliente es la siguiente:

Datos demográficos y de contrato

•Tipo cliente: Se determina que se analizarán los clientes Residencial y Comercial

•Comuna: Comuna en la que reside el cliente que contrata el servicio.

•Última tarifa: último plan de tarifas o bolsas de consumo contratadas.

•Último descuento: Último descuento que les se asignado.

Datos de reclamos y servicio al cliente: Información sobre las llamadas realizadas por el cliente a esta área de la empresa por cada cliente.

•Cantidad de llamados de emergencia durante los últimos 6 meses.

•Cantidad de llamados de emergencia durante los últimos 12 meses.

Red de distribución: Puede ser por anulación de planes por parte del cliente, garantías de artefactos, modificación de datos del cliente, informes de lecturas, solicitud de conexiones, etc.

•Cantidad de llamados por red de distribución realizados los últimos 6 meses.

•Cantidad de llamados por red de distribución realizados los últimos 12 meses.

Gestión de reclamos: Pueden ser por facturas mal emitidas, demoras en la atención, entre otros.

•Cantidad de llamados por gestión de reclamos realizados los últimos 6 meses.

•Cantidad de llamados por gestión de reclamos realizados los últimos 12 meses.

•Cantidad total de llamadas realizado el cliente los últimos 6 meses.

•Cantidad de llamadas que ha realizado el cliente los últimos 12 meses.

Información asociada al consumo del cliente: Se obtuvieron los consumos históricos de la cartera de clientes desde enero de 2017 hasta febrero de 2020 de la empresa en estudio. En general se observa un patrón estacional en el consumo. En general el consumo de gas disminuye en los meses en época de verano y aumenta en invierno.

B.Pre procesamiento del conjunto de datos

La empresa abastece de gas a 4 tipos de clientes: residenciales, comerciales, gran comercio y centrales térmicas. Se decide extraer de la base de datos al Gran Comercio y a las centrales térmicas. Se decide que estos se excluyen de la base, ya que sus fugas son previamente acordadas con la empresa y no dependen de los atributos del modelo.

Dado el rango de datos (enero de 2018 – febrero de 2020) se eliminó de la base de datos a todos aquellos clientes que no representan un consumo regular de gas, es decir a todos los clientes ingresados como clientes nuevos durante el año 2019. También se eliminó a aquellos clientes sobre los cuales no se puede realizar una acción de retención, es decir, aquellos que tuvieron más de seis meses continuos sin consumo durante el año 2019.

La identificación de datos atípicos se realizó para cada variable mediante la regla de tres sigmas [14] . Los datos faltantes y fuera de rango fueron reemplazados mediante el valor de la variable de un registro similar [15].

C.Creación y transformación de variables

La creación de nuevas variables permite mejorar la calidad de la información de manera que el patrón a obtener mediante machine learning tenga un mayor grado de asertividad. Para esto se consideró elementos representativos del comportamiento del consumo de cada cliente similares a los presentados en [16]. Las variables creadas fueron las siguientes:

•Antigüedad cliente: Cantidad de meses en que el cliente ha tenido un consumo.

•Desviación estándar 3: Desviación estándar de los últimos 3 meses de consumo

•Desviación estándar 6: Desviación estándar de los últimos 6 meses de consumo.

•Desviación estándar 12: Desviación estándar de los últimos 12 meses de consumo.

•Promedio 3: Promedio de los últimos 3 meses de consumo del cliente.

•Promedio 6: Promedio de los últimos 6 meses de consumo del cliente.

•Promedio 12: Promedio de los últimos 12 meses de consumo del cliente.

•Último consumo: Consumo en metros cúbicos del cliente el último mes que consumió.

•Consumo hace un año: Consumo en metros cúbicos 12 meses antes del último consumo.

•Diferencia año anterior: Proporción de la diferencia de consumo del cliente de su último mes de consumo con respecto al consumo que tuvo hace 12 meses.

Considerando todas las variables, se identificó a aquellas con alta correlación entre sí mediante una matriz de correlación. Una alta correlación entre variables implica que cada una de ellas explica un fenómeno en forma similar. Por esta razón se decide dejar sólo una variable entre aquellas que presentan una correlación mayor a un 0.9. Las variables que fueron eliminadas del estudio fueron Promedio 3, Promedio 6, Promedio 12 y Desviación estándar 12.

D.Selección de variables

La selección de variables es necesaria para generar modelos más sencillos y mejorar el desempeño de las técnicas de machine learning [17]. Para medir la importancia de las variables se utilizó la ganancia de información. La ganancia de información mide la cantidad de información contenida en un atributo y que explica la variable a predecir. A mayor ganancia de información más importante es la variable para la predicción. Para medir la ganancia de información de cada variable fue necesario categorizar las variables numéricas. Luego de categorizar las variables numéricas se seleccionó entre todas las variables aquellas con un valor de ganancia de información mayor o igual al promedio [18]. La variable a predecir es aquella que define si un cliente se fugó o no se fugó durante enero de 2020. La Tabla I muestra las variables seleccionadas y su respectivo valor de ganancia de información. Es posible observar que las variables más importantes son aquellas relacionadas al comportamiento del consumo del cliente.

Tabla I: Ganancia de información de atributos seleccionado

Atributo	Ganancia de Información
Antigüedad Cliente	0,697
Diferencia año anterior	0,235
Último consumo	0,123
Desviación estándar 3	0,108
Consumo hace un año	0,076

IVRESULTADOS

A.Minería de datos

Para obtener el patrón que caracteriza a los clientes que se fugan y luego obtener una predicción respecto a la fuga, se entrenó y probo dos técnicas de machine learning: redes neuronales artificiales y support vector machine.

Las redes neuronales artificiales (NN), son una representación de las neuronas biológicas las cuales, al ser estimuladas y alcanzar cierto umbral, reaccionan. En una neuronal artificial el estímulo lo generan las variables y la reacción representa un valor de salida el cual puede ser la predicción [19]. Las redes neuronales artificiales están formadas por varias capas de neuronas interconectadas. La capa de neuronas que reciben las variables es llamada capa de entrada. Luego el procesamiento de la información de la capa de entrada se realiza en una siguiente capa llamada capa oculta. La capa oculta puede poseer una o varias capas. El procesamiento de la información en la capa oculta se realiza mediante la interconexión ponderada de sus neuronas. El resultado de este procesamiento lo recibe un conjunto de neuronas llamado capa de salida, las cuales entregan finalmente la predicción o clasificación [20]. El entrenamiento de la red neuronal consiste principalmente en determinar el valor de los ponderadores de las interconexiones entre las neuronas. El valor final de estos ponderadores representa el patrón para la predicción. Dentro de los parámetros a ajustar durante el en-

ISSN 2542-3401/ 1316-4821 UNIVERSIDAD, CIENCIA y TECNOLOGÍA *Vol. 24, Nº 106 Noviembre 2020 (pp. 79-87)*

trenamiento de una red neuronal artificial se encuentra el número de ciclos de entrenamiento, que es el número de veces que pasará la base de datos completa por la red. También está la tasa de aprendizaje que determina la sensibilidad al cambio del valor de los ponderadores en cada ciclo y el factor de momento que acelera la convergencia del cambio de los ponderadores en cada ciclo. El algoritmo de red neuronal artificial entrenado en esta investigación es el incorporado en el software RapidMiner llamado Neural Net. Este algoritmo de red neuronal artificial es del tipo perceptrón multicapa que ajusta sus ponderadores mediante un algoritmo de retropropagación [21].

Support vector machine (SVM) es un modelo de programación no lineal que busca el hiperplano que mejor separa el conjunto de datos en dos regiones o clases dadas, como Fuga o No Fuga de un cliente [22]. El hiperplano encontrado representa el patrón que permitirá la predicción o clasificación [1]. Es común que las observaciones de dos clases no sean linealmente separables, por lo que agrega un parámetro de regularización C el cual representa la tolerancia admitida del incumplimiento de las observaciones que admite el hiperplano. También se incorporar diferentes funciones de kernel que devuelven el producto interno entre dos puntos en un espacio de características adecuado para la separación de las clases y un parámetro de optimización ε. El algoritmo support vector machine entrenado en esta investigación es el incorporado en el software RapidMiner llamado Support Vector Machine el cual es una implementación realizada en lenguaje Java por Stefan Rueping llamado mySVM. El algoritmo mySVM trabaja con funciones de pérdida lineales o cuadráticas e incluso asimétricas [21].

Para el proceso de entrenamiento, validación y obtención de parámetros de ambos algoritmos se utilizó la técnica de hold-out cross-validation para series de tiempo [23] [24] [25] . Esta técnica considera la división temporal de los datos en un conjunto de entrenamiento y prueba como se muestra en la Figura 2.

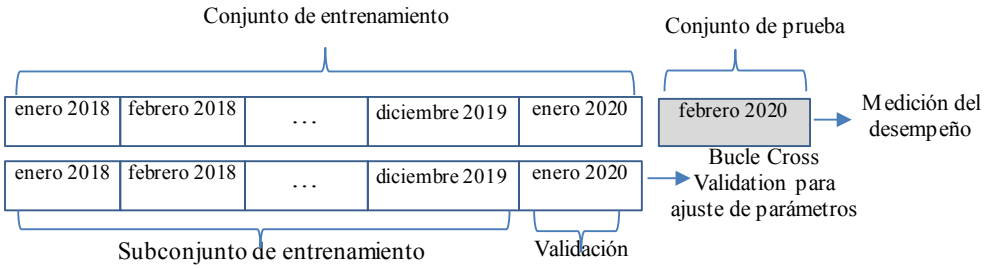

Fig. 2. Proceso de entrenamiento y validación de los algoritmos de machine learning

En la Figura 2 se muestra que el conjunto de entrenamiento formado por los registros desde enero 2018 a enero 2020 se divide en un subconjunto de entrenamiento (enero 2018 a diciembre 2019) y validación (enero 2020) para el ajuste de parámetros. Este ajuste de parámetros se realiza mediante un proceso iterativo hasta encontrar los parámetros óptimos de cada algoritmo. Con los parámetros ajustados, se entrena el modelo mediante el conjunto de entrenamiento y se predice la Fuga o No Fuga de un cliente para febrero de 2020. Mediante esta predicción se obtiene el desempeño predictivo de cada modelo.

La Tabla II muestra el resultado de la cantidad de clientes fugados versus la cantidad de clientes que no se fugan en enero 2020. Se observa un alto desbalance entre las clases No fuga y Fuga.

Tabla II: Clientes fugados a enero 2020

	N° clientes	Porcentaje
No fuga	36.334	99,59%
Fuga	150	0,41%

En estas circunstancias, los algoritmos de machine learning presentarán una tendencia de clasificación hacia la clase mayoritaria. Para mitigar este problema se utiliza la técnica de balance SMOTE (Synthetic Minority Oversampling Method) que genera nuevas instancias de la clase minoritaria para equilibrar la base de datos en base a la técnica del vecino más cercano [26]. En su implementación se buscó los 5 vecinos más cercanos y se generó instancias artificiales, de tal forma que el total de filas de la clase Fuga y No Fuga sean 20.000.

Para medir el desempeño de las técnicas de machine learning, se utilizó la Matriz de confusión mostrado en la Tabla III y las medidas de desempeño Accuracy, Precision y Recall [27].

Tabla III: Matriz de confusión

		Clasificación real	
		Fuga	No Fuga
Clasificación Predicha	Fuga	Verdadero Positivo	Falso Positivo
	No Fuga	Falso Negativo	Verdadero Negativo

La medida Accuracy que mide el desempeño general del modelo y se obtiene la suma de los Verdadero Positivo y Verdadero Negativo dividido por el total de datos en la matriz. Recall representa el porcentaje de clientes que se fugan que fueron clasificadas correctamente y se obtiene al dividir el número de Verdaderos Positivos entre la suma de los Verdaderos Positivos y Falsos Negativos. Precision representa el porcentaje de clientes que se fugan entre el total de clientes predichos como fuga y se obtiene al dividir Verdadero Positivo entre la suma de los Verdadero Positivo y Falso Positivo.

La Tabla IV muestra los valores de la Matriz de Confusión, y las medidas de desempeño para los mejores desempeños de los algoritmos luego de ajustar sus respectivos parámetros. El mejor desempeño lo muestra el algoritmo Support Vector Machine ajustado con kernel radial y parámetro $\gamma = 0.75$ (el parámetro γ ajusta rendimiento del kernel al problema en cuestión), parámetro c=0 y parámetro $\varepsilon = 0.001$ (SVM 3). El segundo mejor desempeño (similar al anterior) lo muestra el algoritmo Support Vector Machine ajustado con kernel radial y parámetro $\gamma = 0.5$ parámetro c=0 y parámetro $\varepsilon = 0.001$ (SVM 2). Finalmente, el tercer mejor desempeño lo muestra el algoritmo Neural Net ajustado con 200 ciclos de entrenamiento, tasa de aprendizaje 0.01 y factor de momento 0.9 (NN 1).

Tabla IV: Desempeño de los algoritmos seleccionados

	Verdadero Positivo	Falso Positivo	Falso Negativo	Verdadero Negativo	Accuracy	Precision	Recall
NN1	19641	3592	359	16408	90.12%	84.54%	98.21%
SVM2	18500	893	1500	19107	94.02%	95.40%	92.50%
SVM3	18665	942	1335	19058	94.31%	95.20%	93.33%

También se obtuvo para cada algoritmo la curva ROC y el valor del área bajo esta curva llamada AUC. La curva ROC es un gráfico que representa la relación entre las proporciones de la tasa Verdadero Positivos y Falso Positivo para distintos valores de corte o umbral [28]. La tasa Verdadero Positivo es el Recall o también llamado Sensitivity. La tasa Falso Positivo se obtiene mediante la división de los Falsos Positivos entre la suma de los Falsos Positivos y Verdaderos Negativos y representa la cantidad de individuos que no se fugaron y que fueron predichos como fugados por el modelo. A este valor se le conoce también como 1- Specificity. El valor de umbral o valor de corte es el valor que define si la predicción entregada por un algoritmo será de la clase Fuga o No Fuga. Es necesario definir este valor pues la predicción de los algoritmos es un valor entre cero y uno llamado confidence. En la curva ROC la exactitud en el desempeño de un algoritmo de machine learning aumenta a medida que la curva se desplaza desde la diagonal hacia el vértice superior izquierdo. Si la clasificación fuera perfecta (Sensitivity =1, 1- Specificity=1) la curva llegaría a dicho punto. La Figura 3 muestra el desempeño de los algoritmos según el área bajo la curva AUC.

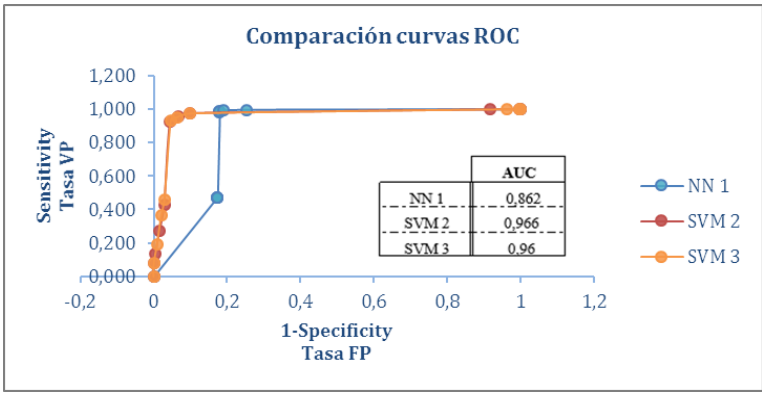

Fig. 3. Curva ROC y AUC de los algoritmos de machine learning seleccionados

ISSN 2542-3401/ 1316-4821

Según el AUC el algoritmo de mejor desempeño fue SVM 2, el cual será considerado como el mejor modelo. Se ajustará el desempeño de esta configuración del algoritmo Support Vector Machine mediante la búsqueda del valor de umbral que minimiza costo de error de clasificación [29]. Para esto se incorpora la función de costos de clasificación dada por la siguiente ecuación:

$$osto = (1 - Tasa\ VP) * Costo(Error\ Tipo\ I) + Tasa\ FP * Costo(Error\ Tipo\ II) \quad (1)$$

El error tipo I ocurre cuando se clasifica como No Fuga a un cliente que si se fugará. Este costo de clasificación es el más caro y corresponde a la pérdida de un cliente. El error tipo II ocurre cuando se clasifica a un cliente como Fuga cuando realmente no se fuga. El costo del error tipo II es la aplicación de una acción de retención a un cliente que no se fugará. Según la información entregada por Gas Sur S.A. la acción de retención es un descuento de un 30% en promedio sobre el consumo regular. El costo de error tipo I es el incurrido al perder a un cliente sobre el cual no se tomó ninguna acción de retención. Este costo en Gas Sur S.A es de un 70% sobre su consumo regular, es decir el consumo total (100%) menos el 30% de descuento por retención si el cliente se hubiese clasificado correctamente.

El valor de umbral que minimiza el costo de error de clasificación se encuentra al intersectar la curva de costos con la curva ROC como se muestra en la Figura 4. El valor de umbral obtenido de esta intersección es 0,707.

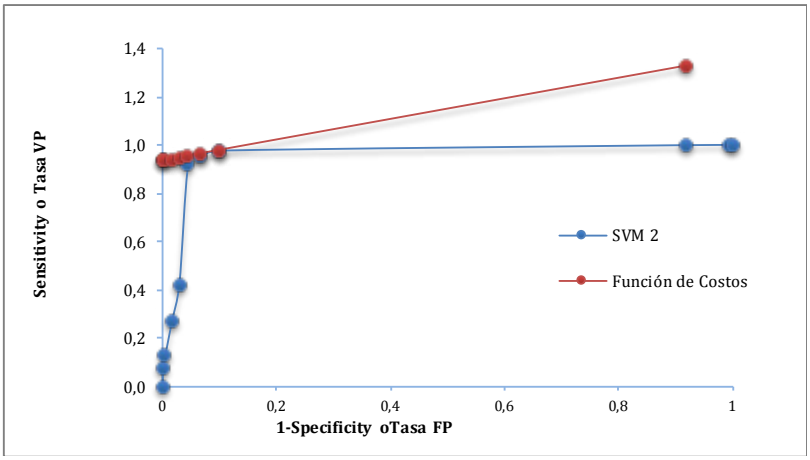

Fig. 4. Intersección de la curva de costos de clasificación y curva ROC.

B.Discusión

El valor de umbral de 0.707 e implica que utilizando SVM2 todos los clientes con un valor de predicción mayor o igual a este valor deben ser considerados como Fuga. El valor de la pendiente de la función de costos resulta muy relevante para la determinación del umbral óptimo, es por eso que la determinación de los costos es decisiva para determinar la forma de interpretar los resultados de la técnica de machine learning seleccionada.

La Figura 5 muestra un histograma con los valores de la predicción de cada cliente para el mes de febrero de 2020. Por ejemplo, de acuerdo este, existen 24.142 clientes que tienen probabilidad de fuga entre un 22% a un 32%, rango que tiene la mayor cantidad de clientes.

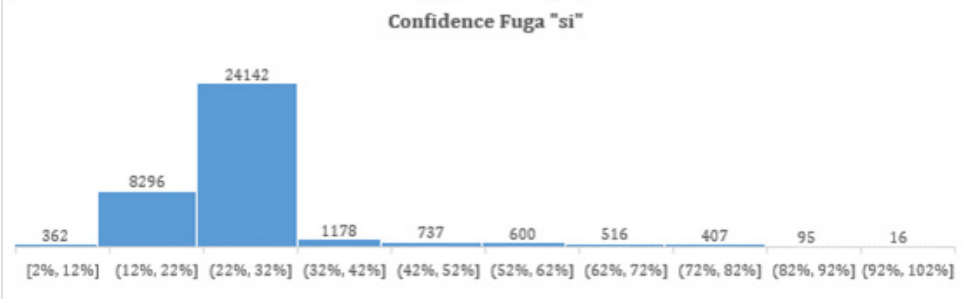

Fig. 5: Histograma de los valores de predicción del algoritmo SVM seleccionado.

Según el modelo los clientes que potencialmente podrían fugarse (confidence sobre 0.707) son 518. Un elemento a tener en consideración para aplicar acciones de retención sobre estos clientes es la variable Último consumo. A Gas Sur S.A maximiza su utilidad al retener dentro de los clientes con mayor probabilidad de fuga a aquellos que presenten un mayor consumo. De acuerdo a esto, se deben realizar acciones de retención, las que están ligadas principalmente a descuentos en el consumo, a los clientes con un valor de predicción mayor o igual a 0.707 y que presentan un último consumo Medio a Alto.

La Figura 6 muestra un diagrama de los componentes de debería tener el sistema para la identificación de clientes que se fugarán.

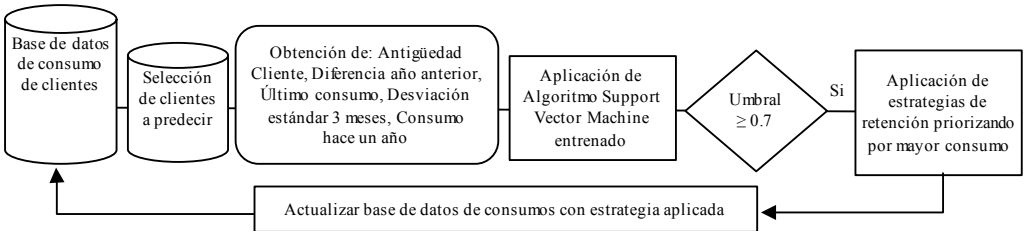

Fig. 6: Componentes del sistema de identificación y retención de clientes que se fugarán.

V.CONCLUSIONES

Las variables más importantes y que permitieron maximizar el desempeño de las técnicas de machine learning entrenadas para la identificación de un patrón de fuga de clientes de la empresa Gas Sur S.A fueron: Antigüedad del cliente, Proporción de la diferencia del consumo del último mes respecto al año anterior, Último consumo registrado, Desviación estándar de los últimos 3 meses de consumo y Consumo en metros cúbicos 12 meses antes del último consumo. Estas variables son de fácil obtención y reflejan el comportamiento de consumo del cliente. El no considerar factores sociodemográficos o de servicio al cliente permite que el modelo pueda tener una aplicabilidad más universal en la empresa.

Considerando la Matriz de Confusión, la curva ROC y el AUC de cada algoritmo de machine learning entrenado, se determinó que el mejor algoritmo fue Support Vector Machine (mySVM) ajustado con kernel radial y parámetro $\gamma = 0.5$ parámetro c=0 y parámetro $\varepsilon = 0.001$.

Respecto a los costos de clasificación, en la fuga de clientes de Gas Sur S.A. el costo de error tipo I está asociado a la pérdida del consumo por parte de un cliente y el costo de error tipo II a aplicar un descuento a un cliente que no se fugará. Estos costos corresponden respectivamente a un 70% y 30% del consumo de un cliente. Mediante la incorporación de estos costos de error de clasificación al modelo, fue posible determinar que el valor de umbral que minimiza el costo de error de clasificación en la predicción. Este valor de umbral es 0.707 e implica que cualquier cliente que tenga un valor de predicción igual o superior a este valor será clasificado como un cliente que se fugará.

La utilización del algoritmo Support Vector Machine seleccionado permitirá a Gas Sur S.A. identificar aquellos clientes con mayor probabilidad de fuga, priorizar las acciones de retención en aquellos con un mayor consumo y minimizar el costo de error de clasificación, de manera de maximizar el beneficio de la acción de retención de los clientes.

REFERENCIAS

[1]J. Miranda, P. Rey y R. Weber, «Predicción de Fugas de Clientes para una Institución Financiera Mediante Support Vector Machines,» Revista Ingeniería de SistemasVolumen XIX,, pp. 49-68, 2005.

[2]P. A. Pérez V., «Modelo de predicción de fuga de clientes de telefonía movil post pago,» Universidad de Chile, Santiago, Chile, 2014.

[3]Gas Sur S.A., «https://www.gassur.cl/Quienes-Somos/,» [En línea].

[4]J. Xiao, X. Jiang, C. He y G. Teng, «Churn prediction in customer relationship management via GMDH-based multiple classifiers ensemble,» IEEE Intelligent Systems, vol. 31, nº 2, pp. 37-44, 2016.

[5]A. M. Almana, M. S. Aksoy y R. Alzahrani, «A survey on data mining techniques in customer churn analysis for telecom industry,» International Journal of Engineering Research and Applications, vol. 4, nº 5, pp. 165-171, 2014.

[6]A. Jelvez, M. Moreno, V. Ovalle, C. Torres y F. Troncoso, «Modelo predictivo de fuga de clientes utilizando mineríaa de datos para una empresa de telecomunicaciones en chile,» Universidad, Ciencia y Tecnología, vol. 18, nº 72, pp. 100-109, 2014.

[7]D. Anil Kumar y V. Ravi, «Predicting credit card customer churn in banks using data mining,» Internatio-

nal Journal of Data Analysis Techniques and Strategies, vol. 1, nº 1, pp. 4-28, 2008.
[8]E. Aydoğan, C. Gencer y S. Akbulut, «Churn analysis and customer segmentation of a cosmetics brand using data mining techniques,» Journal of Engineering and Natural Sciences, vol. 26, nº 1, 2008.
[9]G. Dror, D. Pelleg, O. Rokhlenko y I. Szpektor, «Churn prediction in new users of Yahoo! answers,» de Proceedings of the 21st International Conference on World Wide Web, 2012.
[10]T. Vafeiadis, K. Diamantaras, G. Sarigiannidis y K. Chatzisavvas, «A comparison of machine learning techniques for customer churn prediction,» Simulation Modelling Practice and Theory, vol. 55, pp. 1-9, 2015.
[11]Y. Xie, X. Li, E. Ngai y W. Ying, «Customer churn prediction using improved balanced random forests,» Expert Systems with Applications, vol. 36, nº 3, pp. 5445-5449, 2009.
[12]U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, «Knowledge Discovery and Data Mining: Towards a Unifying Framework,» de KDD-96 Proceedings, 1996.
[13]R. Brachman y T. Anand, «The process of knowledge discovery in databases,» de Advances in knowledge discovery and data mining, 1996.
[14]K. Lakshminarayan, S. Harp, R. Goldman y T. Samad, «Imputation of Missing Data Using Machine Learning Techniques,» de KDD, 1996.
[15]B. Nguyen , J. L. Rivero y C. Morell, «Aprendizaje supervisado de funciones de distancia: estado del arte,» Revista Cubana de Ciencias Informáticas, vol. 9, nº 2, pp. 14-28, 2015.
[16]I. Monedero, F. Biscarri, J. Guerrero, M. Peña, M. Roldán y C. León, «Detection of water meter under-registration using statistical algorithms,» Journal of Water Resources Planning and Management, vol. 142, nº 1, p. 04015036, 2016.
[17]I. Guyon y A. Elisseeff, «An introduction to variable and feature selection,» Journal of machine learning research, vol. 3, nº Mar, pp. 1157-1182, 2003.
[18]K. Polat y S. Güneş, «A new feature selection method on classification of medical datasets: Kernel F-score feature selection,» Expert Systems with Applications, vol. 36, nº 7, pp. 10367-10373, 2009.
[19]D. J. Matich, «Redes Neuronales. Conceptos Básicos y Aplicaciones,» de Cátedra: Informática Aplicada ala Ingeniería de Procesos- Orientación I, 2001.
[20]E. Acevedo M., A. Serna A. y E. Serna M., «Principios y Características de las Redes Neuronales Artificiales,» de Desarrollo e Innovación en Ingeniería, Medellín, Editorial Instituto Antioqueño de Investigación, 2017, pp. Capítulo 10, 173-182.
[21]M. Hofmann y R. Klinkenberg, RapidMiner: Data mining use cases and business analytics applications, CRC Press, 2016.
[22]R. Pupale, «Towards Data Science,» 2018. [En línea]. Available: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989.
[23]F. H. Troncoso Espinosa, «Prediction of recidivism in thefts and burglaries using machine learning,» Indian Journal of Science and Technology, vol. 13, nº 6, pp. 696-711, 2020.
[24]L. Tashman, «Out-of-sample tests of forecasting accuracy: an analysis and review,» International journal of forecasting, vol. 16, nº 4, pp. 437-450, 2000.
[25]S. Varma y R. Simon, «Bias in error estimation when using cross-validation for model selection,» BMC bioinformatics, vol. 7, nº 1, p. 91, 2006.
[26]N. V. Chawla, K. W. Bowyer, L. O. Hall y W. Kegelmeyer, «SMOTE: Synthetic Minority Over-sampling Technique,» Journal of Artificial Inteligence Research 16, pp. 321-357, 2002.
[27]M. Sokolova y G. Lapalme, «A systematic analysis of performance measures for classification tasks,» Information processing & management, vol. 45, nº 4, pp. 427-437, 2009.
[28]S. Narkhede, «Understanding AUC-ROC Curve,» Towards Data Science, vol. 26, 2018.
[29]R. Westermann y W. Hager, «Error Probabilities in Educational and Psychological Research,» Journal of Educational Statistics, Vol 11, No 2, pp. 117-146, 1986.

RESUMEN CURRICULAR

Fredy Troncoso Espinosa, Doctor en Sistemas de Ingeniería, Universidad de Chile, Ingeniero Civil Industrial Universidad del Bío-Bío, Chile. Académico e Investigador Departamento de Ingeniería Industrial, Universidad del Bío-Bío. Concepción, Chile

Javiera Ruiz Tapia, Ingeniero Civil Industrial, Universidad del Desarrollo, Magister en Ingeniería Industrial, Universidad del Bío-Bío. Ingeniero Analista de Redes en Empresa Gas Sur, Talcahuano, Chile