ISSN 1870-4069

Evaluación del abandono de clientes de una compañía de telecomunicaciones por medio de cuatro modelos de aprendizaje máquina

Jorge Octavio Castro Rodríguez, Ernesto Pérez Vázquez

Universidad Panamericana,
Facultad de Ingeniería,
México

{0234588,0234686}@up.edu.mx

Resumen. Debido a que los clientes son el activo más preciado de las compañías, la retención de clientes es un objetivo esencial dentro de la estrategia de desarrollo de cualquier organización. En un entorno donde los servicios, productos y valor añadido están en constante evolución se vuelve cuestión de supervivencia contar con una estrategia de retención de clientes. Basados en información histórica existente en combinación con herramientas tecnológicas de procesamiento de información el aprendizaje de maquina (ML) nos permite realizar la extracción de conocimiento de una manera poderosa. Este proyecto implementó las siguientes técnicas de ML: Regresión logística, Random forest, SVM y XGBoost. Los resultados representan las características de los clientes que abandonaron. El negocio puede identificar los abandonos que sucederán usando los resultados de las técnicas empleadas. Con base a dicha información la organización debe implementar estrategias de retención para aquellos clientes más propensos a abandonar al negocio y cuyas características estén apegadas a los objetivos de desarrollo organizacional.

Palabras clave: Aprendizaje máquina, aprendizaje supervisado, ajuste de hiperparámetros, principio de parsimonia, abandono de clientes.

Evaluation of Customer Churn of a Telecom Company Using Four Machine Learning Models

Abstract. Clients are the most valuable asset of a company. Therefore, retaining them is a vital part of the strategy inside any organization. Nowadays, services, products and added value are constantly evolving, which makes it critical to possess a strategy for retaining customers. Based on existing historic data in combination with tech tools of processing data for machine learning (ML), we show how to make knowledge extraction in a powerful manner. In this project, we implement the

following techniques of ML: Logistic Regression, Random Forest, SVM and XGBoost. The results represent features of CHURN customers. The company can identify where CHURN occurs by using the results of those techniques. Based on this, a business may implement strategies for the more likely CHURN customers in which features are attached to the objectives of an organizational development.

Keywords: Machine learning, supervised learning, hyperparameters optimization, principle of parsimony, customer CHURN.

1. Introducción

Haciendo énfasis en la base de que es mucho más costoso captar un nuevo cliente que retener a los clientes actuales y el hecho que a periodos más largos de permanencia los clientes generan mayores ganancias, [1] afirma que la retención de clientes incrementa la rentabilidad.

Hoy día las compañías invierten grandes cantidades de recursos económicos, tecnológicos, humanos y estratégicos para aumentar el número de clientes que se tienen registrados en la base de datos histórica del negocio, sin embargo, no se tiene una estrategia clara y enfocada para retener a los clientes actuales pero sobre todo retener a los clientes ideales (clientes que encajen en el tipo y estrategia de nuestro negocio).

Con base a lo anterior es sumamente importante tener conocimiento de los puntos que se darán a continuación, ya que en cualquier negocio donde los costos de captación son elevados y los márgenes de ganancia del producto o servicio son pequeños, la manera más eficiente para generar rentabilidad es con base al consumo recurrente de los productos y/o servicios de la compañía. Por tanto, minimizar la tasa de abandono de los clientes adecuados se vuelve un pilar de supervivencia de las compañías.

- **Conocimiento de clientes.** Conocer cómo se segmentan nuestros clientes para tener claridad de sus gustos, comportamiento he intereses y de esta manera enfocar la creación de nuevos productos y/o servicios de manera dirigida a captar dichas necesidades.
- **Acciones históricas.** Análisis de las acciones históricas que se han realizado para fidelizar a los clientes y a partir de ello identificar la efectividad de dichas estrategias y de esta manera enriquecer la base de información que alimente tanto al conocimiento de nuestros clientes (segmentación) y de esta manera maximizar las opciones y niveles de retención de nuestros clientes.
- **Optimización.** Mejora continua de nuestros servicios, productos, promociones, ofertas comunicados, etc., que genere un continuo y genuino interés en nuestra compañía y en lo que ofrece (optimizar valor agregado de la compañía).

Dado a lo anterior, el objetivo establecido para este artículo es ejecutar una metodología de ciencia de datos que nos permita llevar a cabo la implementación de cuatro modelos distintos de aprendizaje automático supervisado, a partir de un set de datos históricos en común, y con base a los resultados de cada uno de los modelos realizar un análisis comparativo que nos ayude finalmente a determinar el modelo óptimo.

Para llevar a cabo la implementación de los modelos elaboraremos para cada uno de ellos los siguientes pasos generales (en la sección número cuatro se describe a mayor detalle las etapas de la metodología implementada):

- Recolección de datos (set de datos compartidos).
- Pre-procesamiento.
- Extracción de características.
- Selección de características.
- Creación de modelos.
- Evaluación de modelos.

Este documento se encuentra estructurado de la siguiente manera:

- **Trabajos relacionados.** En esta sección se observan los trabajos históricos de autores distintos que se tomaron como referencia base para el desarrollo de este artículo, dichos trabajos y autores fueron seleccionados debido a las similitudes con respecto a nuestro objetivo inicial (financiero-banca) como con nuestro objetivo final (Telecomunicaciones - considerando las limitaciones)
- **Limitaciones.** En este apartado se describen los elementos restrictivos para llevar el estudio presentado en este documento
- **Material y método.** En este bloque narra la propuesta del trabajo, los materiales utilizados y la metodología llevada a cabo
- **Experimentación.** Se describen los experimentos realizados y las métricas de evaluación
- **Resultados y discusión.** En este módulo se plasman los resultados obtenidos por cada uno de los modelos con la finalidad de abrir el debate dirigido a determinar el modelo más adecuado para el set de datos utilizados
- **Conclusiones.** Sección que se utiliza para detallar las conclusiones del proyecto y consideraciones futuras del estudio.

"La infidelidad siempre tiene como constantes la insatisfacción, desapego y desconfianza".

2. Trabajos relacionados

En Keramati et al. [2] se enfocan en la predicción de abandono de clientes en los servicios de banca electrónica utilizando el método de árbol de decisiones de aprendizaje máquina. Raj [3] utiliza aprendizaje máquina para predecir el abandono de clientes de telecomunicaciones haciendo uso de KNN, random forest,

Gaussian Naive Bayes, Support Vector Machine, LightGBMClassifier, XGBoost. Leon Palafox muestra el proceso para el entendimiento de negocio perfilado al abandono de cliente, entendimiento de datos y su exploración, Guo-en and Wei-dong [4] se enfocan en construir un modelo de predicción de abandono de clientes usando SVM en la industria de telecomunicaciones. De manera tabular los podemos resumir en la figura 1:

Investigador y/o año	Industria	Metodo(s)
AbbasKeramati, HajarGhaneei and SeyedMohammadMirmohammadi (2016)	Banca	Árbol de decisión
Pavan Raj (2018)	Telecomunicaciones	KNN, random forest, Gaussian Naive Bayes, Support Vector Machine, LightGBMClassifier, XGBoost
Leon Palafox	Telecomunicaciones	Análisis exploratorio

Fig. 1. Trabajos relacionados.

3. Limitaciones

Inicialmente se tenía planeado realizar el ejercicio con una base de datos de Nacional Monte de Piedad, sin embargo, no fue posible debido a las restricciones de la institución para compartir información productiva lo que represento un limitante respecto al desarrollo de este estudio, por lo que se optó por trabajar con información pública del sitio Kaggle. La investigación futura tendrá como objetivo realzar este ejercicio con información productiva esperando lograr superar los problemas de privacidad asociados, de igual manera se planea programar un despliegue el cual no se ejecutara debido al corto tiempo con el cual se contó para obtener las autorizaciones correspondientes.

4. Materiales y método

El estudio se llevó a cabo conforme al esquema denominado CRISPDM, (en la figura 2 podemos observar el flujo general utilizado):

- **Entendimiento del negocio** El entendimiento del negocio debe partir del conocimiento de los objetivos estratégicos de las organizaciones ya que a partir de dicho conocimiento se pueden establecer líneas base de observación y/o acción. Debido a las limitantes del set de datos se entiende que este apartado queda pendiente para futuros ejercicios, sin embargo, cabe recalcar

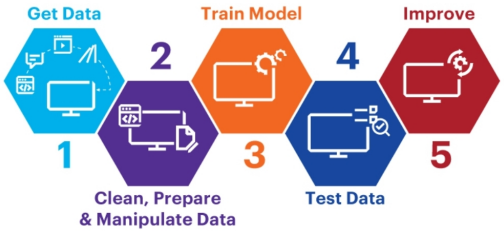
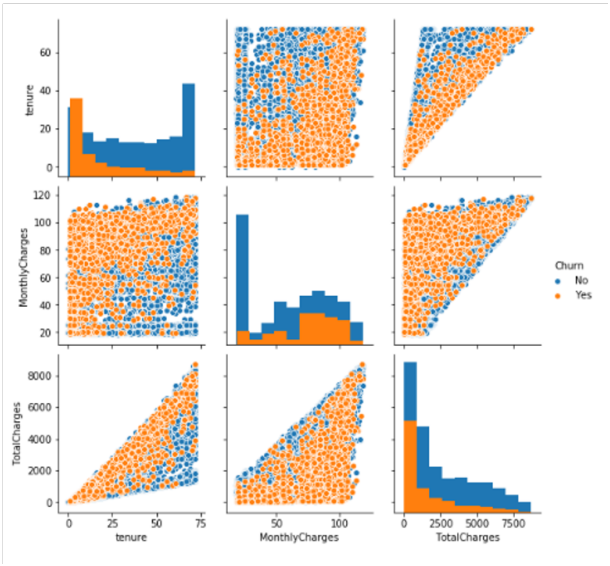


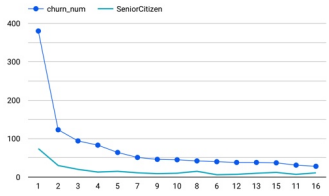
Fig. 2. Flujo de trabajo.

que con base al entendimiento de la metodología empleada este estudio podrá ser ejecutado con mayor exactitud y experiencia en cuanto se logre obtener información productiva.

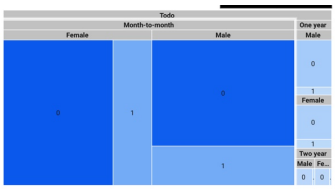
- **Entendimiento de los datos** Como parte de la exploración de datos descubrimos comportamientos valiosos para la toma de decisiones, por ejemplo, figura 3:
  - El abandono temprano de clientes jóvenes.
  - Las personas que pagan menos permanecen por mayor tiempo.
  - Las personas que cuentan con el servicio de telefonía tienden a tener una taza alta de abandono.
  - Las personas que utilizan fibra óptica tienden a presentar una taza alta de abandono.
  - El abandono no se relaciona con el género del cliente.
  - Al finalizar el primer mes de servicio es donde se nota la mayor tasa de abandono.
  - La correlación de las variables pertenecientes al set de datos.
  - Con forme los clientes permanecen mayor tiempo con el servicio la tasa de abandono disminuye.
- **Preparación de los datos** La información se preparó mediante Python. Dado que encontramos casos en los cuales tuvimos que eliminar espacios (TotalCharges) o eliminar valores vacíos (TotalCharges), se opto por excluir 11 observaciones por contener en una columna valores vacíos, normalizamos los valores, excluimos las variables gender y customerID ya que con base al análisis exploratorio observamos que no tienen valor representativo, como se muestra en la figura 3. Para efectos de este estudio el set de datos fue dividido en dos subsets de datos, el primero de ellos contiene el 70 porciento de las observaciones y fue reservado para la etapa de entrenamiento, el 30 porciento de la información fue utilizada para la etapa de evaluación.
- **Modelado** Se detalla en el apartado 5 (Experimentación) de este documento.
- **Evaluación** Se detalla en el apartado 5 (Experimentación) de este documento.
- **Despliegue** Se observa en la sección de limitantes.



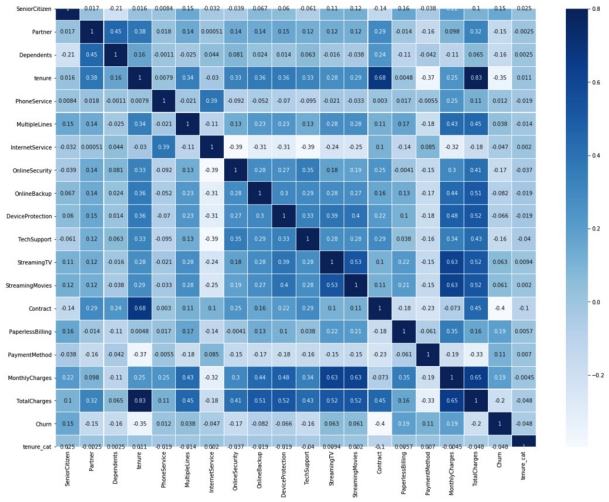
(a) Análisis exploratorio.



(b) Abandono por mes por Seniority.



(c) Relación de abandono por genero.



(d) Matriz de correlación.

Fig. 3. Análisis exploratorio.

4.1. Descripción general

Se programará la ejecución de cuatro modelos de aprendizaje maquina con el objetivo de seleccionar el modelo que nos proporcione la alternativa óptima para apoyar a la toma de decisiones, en la figura 4 se puede observar las variables utilizadas para evaluar los cuatro modelos para el caso de estudio de abandono de clientes.

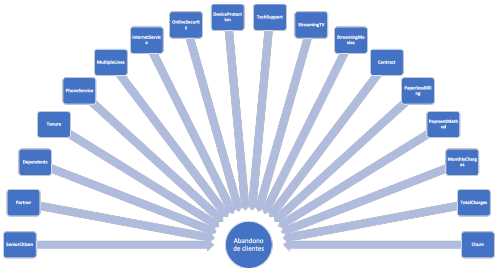


Fig. 4. Mapa conceptual para un modelo de abandono de clientes.

4.2. Materiales

Para llevar a cabo el caso de estudio utilizamos las siguientes herramientas: Python v3, Google Colaboratory, Google data studio, Anaconda.

El resumen del set de datos utilizado se describe en la figura 5.

Artefacto	Descripción
Total de observaciones analizadas	7043
Formato del archivo	csv
Columnas totales	21
Duplicados customerID	0
Columnas homologadas (descripciones)	7
Edición de tipo de dato de columna existente (TotalCharges, String to int)	1
Variables nulas	11 (TotalCharges)
Entendimiento de columna SeniorCitizen	1=Sr (mayor), 2=Jr (joven)
Entendimiento de columna tenure	Meses de permanencia
Entendimiento de columna PhoneService	Telefonía móvil
Total de variables	19
Variables excluidas (customerID, gender)	2
customerID (string)	7043
Gender (category)	2 tipos
SeniorCitizen (Integer)	2 tipos
Partner (category)	2 tipos
Dependents (category)	2 tipos
Tenure (Integer)	73 tipos
PhoneService (category)	2 tipos
MultipleLines (category)	3 tipos
InternetService (category)	3 tipos
OnlineSecurity (category)	3 tipos
OnlineBackup (category)	3 tipos
DeviceProtection (category)	3 tipos
TechSupport (category)	3 tipos
StreamingTV (category)	3 tipos
StreamingMovies (category)	3 tipos
Contract (category)	3 tipos
PaperlessBilling (category)	2 tipos
PaymentMethod (category)	4 tipos
MonthlyCharges (Integer)	1585 tipos
TotalCharges (Integer)	6531 tipos
Churn (category)	2 tipos

Fig. 5. Resumen del set de datos.

4.3. Métodos de aprendizaje

Los métodos de aprendizaje máquina utilizados en este trabajo se describen a continuación.

Regresión logistica. Chitarroni [5] define la regresión logística como un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia hemos puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas.

- **Ventajas:** Al ser un método de clasificación fácil, rápido y simple, tendremos un menor costo computacional que los otros modelos que usaremos para ajustar este set de datos, de esta manera se tiene la opción de usar un rango más amplio de hiperparámetros lo que posiblemente nos dará entrada a tener un mejor desempeño del modelo ajustado.
- **Desventajas:** En este tipo de modelo se requiere de una adecuada selección de características para que este tenga un mayor rendimiento. Dado esto se necesita pensar que al hacer una selección de características mayormente enfocada a la visión de negocio podríamos tener alguna desventaja en este punto.

Random forest. Random Forest es una combinación de árboles de decisión, tal que cada árbol depende de los valores de un vector aleatorio, independiente y con la misma distribución para cada uno de estos [6]. Todos los árboles tienen la misma distribución en el bosque (forest), pero son forzados a ser diferentes. Esto reduce la correlación. Existen diversas técnicas de construcción de conjuntos de clasificadores. En muchas ocasiones, la forma de llevar a cabo estas técnicas es a partir de un conjunto de números aleatorios:

- En Bagging Breiman [7], cada clasificador se construye a partir de una muestra bootstrap, la cual se genera utilizando tantos números aleatorios como elementos tenga el conjunto de entrenamiento.
- En random subspace Ho [8], cada clasificador base utiliza sólo un subconjunto de atributos seleccionado aleatoriamente de entre el total de las variables. De esta forma, cada clasificador se restringe a un subespacio aleatorio de atributos.
- Las técnicas output flipping y class switching, Breiman [9]); Martínez-Muñoz and Suárez [10], se basan en la manipulación aleatoria de las etiquetas de clase, por lo que necesitan de números aleatorios para seleccionar los datos cuyas etiquetas serán cambiadas.
- Randomization Dietterich [11], introduce aleatoriedad en el algoritmo de aprendizaje construyendo conjuntos de clasificadores con árboles de decisión en los que el valor de corte es seleccionado aleatoriamente entre los F mejores cortes posibles.

A continuación, se describen las ventajas y desventajas mas representativas del uso de este modelo:

- **Ventajas:** Este un modelo que funciona bien para métodos de clasificación y parte de sus bondades es que la preparación de los datos que requiere es mínima y al utilizar múltiples árboles se logra reducir el riesgo de sobre ajuste de nuestro modelo.
- **Desventajas:** En este modelo a diferencia del anterior el costo computacional y tiempo de entrenamiento se incrementará con base en los hiperparámetros a usar. Por otro lado, nos enfrentaremos a problemas como el no poder garantizar que el árbol generado sea el óptimo.

Si el clasificador base utilizado es un árbol de decisión, el concepto random forest [12] engloba todas estas técnicas.

XGBoost. Extreme Gradient Boosting (XGBoost), es una técnica de clasificación supervisada que utiliza un conjunto de árboles de decisión. XGBoost es un algoritmo de Machine Learning basado en un árbol de decisiones que utiliza un marco de impulso de gradiente. En los problemas de predicción que involucran datos no estructurados (imágenes, texto, etc.), las redes neuronales artificiales tienden a superar a todos los demás algoritmos o marcos. Sin embargo, cuando se trata de datos estructurados / tabulares de pequeño a mediano, los algoritmos basados en el árbol de decisión se consideran los mejores en su clase en este momento. La figura 6 muestra la evolución de los algoritmos basados en árboles a lo largo de los años.

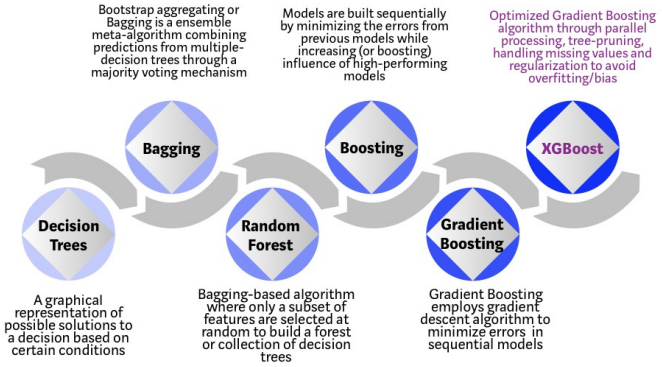


Fig. 6. Evolución del algoritmo XGBoost [13].

- **Ventajas:** Este un modelo que nos funciona bien para temas de clasificación y fundamentalmente nos ayudara para evitar temas de sobre ajuste, permitiendo restringir el crecimiento de los árboles.

– **Desventajas:** Dado que este tipo de modelos es más difícil de afinar que el modelo anterior ("Random Forest"), podemos pensar que entre mejor queramos optimizar temas como; número de árboles, profundidad de árboles y tasa de aprendizaje, tendremos tiempos y costos de entrenamiento más elevados que el modelo anterior.

Máquinas de soporte vectorial. Betancourt [14] explica que una Máquina de Soporte Vectorial (SVM) al ser un clasificador de una sola clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un kernel gaussiano u otro tipo de kernel a un espacio de características en un espacio dimensional más alto, donde se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en todas las clases distintas, cada una formando un agrupamiento.

– **Ventajas:** De todos los modelos mencionados anteriormente este modelo tiene una mejor solución debido a que toma ventaja de los siguientes puntos:
  - Utiliza el truco del kernel para resolver soluciones complejas.
  - Utiliza una función de optimización convexa, debido a que los mínimos globales siempre son alcanzables.
  - El concepto de "Hinge loss" proporciona mayor precisión.
  - Los valores atípicos se pueden manejar bien utilizando suavización sobre la constante C.
– **Desventajas:** Si bien es un modelo que nos brinda grandes ventajas, sobre los modelos anteriores, debemos de tomar en consideración los siguientes aspectos:
  - El concepto de "Hinge loss" conduce a la escasez (matriz poblada principalmente con ceros).
  - Los hiperparámetros y los kernel deben ajustarse cuidadosamente para una precisión suficiente.
  - El tiempo y costo computacional de entrenamiento son elevados.

5. Experimentación

En esta sección contrastaremos el desempeño de los modelos:

– **5.1** Regresión Logistica,
– **5.2** Random Forest,
– **5.3** XGBoost,
– **5.4** Máquinas de Soporte Vectorial.

Para generar los modelos realizamos una división del set de datos en el cual 70 porciento de las observaciones se usaron para entrenar el modelo y el 30 porciento restante para probar el modelo, adicional se hizo uso la técnica de

validación por cruce de k-pliegue para mejorar el desempeño de los modelos, el valor de K se definió en 10 (pliegues).

Utilizamos los indicadores de desempeño descritos en la figura 7 para evaluar y contrastar el desempeño entre los modelos y de esta manera definir nuestra recomendación del modelo más óptimo.

Accuracy: Mide el porcentaje de casos que el modelo ha acertado.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Precision: Mide la calidad del modelo.

$$Precision = \frac{TP}{(TP + FP)}$$

Recall: Informa sobre la cantidad que el modelo es capaz de identificar.

$$Recall = \frac{TP}{(TP + FN)}$$

F1-Score: Compara el rendimiento combinado de la precisión y la exhaustividad.

$$F1Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

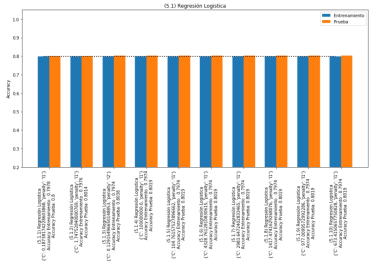
Fig. 7. Interpretación de indicadores.

Se realizó un ranking de desempeño de los hiperparámetros utilizados por modelo con la finalidad de definir los valores óptimos de los hiperparámetros a fijar por modelos, en la figura 8 se muestra un ejemplo de la evaluación de hiperparámetros realizada a la regresión logística de este estudio (las figuras de los modelos restantes se pueden observar en el libro de trabajo que contiene el código que soporta este articulo). Además, en esta misma figura 8 se puede apreciar el listado de hiperparámetros seleccionados a utilizar por cada uno de los modelos.

6. Resultados y discusión

6.1. Resultados

En la figura 9 se observan los resultados obtenidos tras haber evaluado los modelos enunciados en este estudio y el ranking que presentó cada uno de ellos (evaluados con distintos valores en hiperparámetros). Adicional cabe destacar que utilizamos RFE para verificar si el desempeño de los modelos mejoraba, sin embargo no notamos diferencia por lo que la ejecución final no considera RFE (en el libro de trabajo del documento se aprecia el código ejecutado a para esta puntualización).



(a) Ejemplo de comparativa de desempeño de hiperparametro (RL).

Modelo	Hiperparametro
Maquinas de Soporte Vectorial - SVM	{'C': 4, 'gamma': 0.001, 'kernel': 'rbf'}
Regresión Logística	{'C': 0.12915496650148853, 'penalty': 'l2'}
Random Forest	'bootstrap': [True], 'max_depth': [100, 110], 'max_features': ['auto'], 'min_samples_leaf': [3, 4], 'min_samples_split': [8, 10], 'n_estimators': [100, 500, 1000]
XGBoost	'learning_rate': [0.15, 0.20], 'max_depth': [5, 6, 7, 8], 'gamma': [0.4, 0.6], 'colsample_bytree': [0.4], 'n_estimators': [200], 'objective': ['binary:logistic']

(b) Hiperparámetros por modelo.

Fig. 8. Desempeño de hiperparámetros.

Ranking	Modelo	params_cv	Accuracy_Train	Accuracy_test	Recall_test	Precision_test	F1_test
1	(5.4.1) Maquinas de Soporte Vectorial - SVM	{'C': 4, 'gamma': 0.001, 'kernel': 'rbf'}	0.7972	0.8043	0.4957	0.7034	0.5816
2	(5.1.3) Regresión Logistica	{'C': 0.12915496650148853, 'penalty': 'l2'}	0.7974	0.8038	0.5354	0.6813	0.5996
3	(5.4.6) Maquinas de Soporte Vectorial - SVM	{'C': 8, 'gamma': 0.001, 'kernel': 'sigmoid'}	0.7962	0.8033	0.4974	0.6990	0.5812
4	(5.4.3) Maquinas de Soporte Vectorial - SVM	{'C': 7, 'gamma': 0.001, 'kernel': 'sigmoid'}	0.7966	0.8024	0.4853	0.7025	0.5741
5	(5.1.4) Regresión Logistica	{'C': 100000000.0, 'penalty': 'l1'}	0.7974	0.8019	0.5354	0.6754	0.5973
6	(5.1.5) Regresión Logistica	{'C': 104.76157527896662, 'penalty': 'l1'}	0.7974	0.8019	0.5354	0.6754	0.5973
7	(5.2.9) Random Forest	{'bootstrap': True, 'max_depth': 110, 'max_fea...	0.8003	0.8000	0.5095	0.6813	0.5830
8	(5.2.10) Random Forest	{'bootstrap': True, 'max_depth': 100, 'max_fea...	0.8003	0.8000	0.5095	0.6813	0.5830
9	(5.2.1) Random Forest	{'bootstrap': True, 'max_depth': 110, 'max_fea...	0.8009	0.7995	0.5130	0.6781	0.5841
10	(5.3.3) XGBoost	{'colsample_bytree': 0.4, 'gamma': 0.4, 'learn...	0.7926	0.7976	0.5181	0.6696	0.5842
11	(5.3.7) XGBoost	{'colsample_bytree': 0.4, 'gamma': 0.6, 'learn...	0.7899	0.7943	0.5250	0.6566	0.5835
12	(5.3.1) XGBoost	{'colsample_bytree': 0.4, 'gamma': 0.4, 'learn...	0.7936	0.7910	0.5147	0.6507	0.5747

Fig. 9. Ranking de evaluación de performance por modelo.

6.2. Discusión

Con base a la figura 9, podemos argumentar las siguientes observaciones: El mejor indicador de *accuracy* lo tiene el modelo (5.4.1) de Maquinas de Soporte Vectorial - SVM, y en un segundo lugar encontramos que el modelo (5.1.3) de

regresión logística, con base a lo anterior podríamos concluir fría y numéricamente que la opción a utilizar seria la implementación de un modelo de máquina de soporte vectorial SVM, sin embargo si consideramos el uso computacional que conllevó evaluar los modelos, la complejidad de los mismos y la diferencia marginal de resultados (acurracy), nuestra recomendación para este caso de estudio es utilizar el modelo de regresión logística (considerando las limitantes y el uso de hiperparámetros recomendado).

Con base a las condicionantes del set de datos, limitantes del entorno y el tipo de modelos seleccionados al iniciar este estudio es necesario considerar que para determinar un un modelo óptimo no solo se debe contemplar la precisión demostrada por los modelos, dicha decisión debería considerar algunos de los siguientes puntos:

- Tiempo de implementación.
- Infraestructura con la que cuenta la compañía para evaluar un modelo.
- Infraestructura con la que cuenta la compañía para desplegar y operar el modelo.
- Conocimiento del caso de negocio y el objetivo del mismo.
- Conocimiento e interpretación de los datos.
- Entorno.
- Madurez del equipo que desarrolla los modelos.

7. Conclusiones

Para este caso de estudio seguimos los pasos de la metodología CRISPDM para evaluar el abandono de clientes de una empresa de telecomunicaciones, teniendo como objetivo realizar un análisis comparativo entre distintos modelos de aprendizaje máquina que nos permita determinar el modelo más óptimo para el set de datos utilizado. Si bien se siguió la metodología CRISPDM es necesario puntualizar nuevamente que no se logró implementar el paso final de despliegue de los modelos, debido a las limitantes ya descritas en este documento, lo que implica que no tenemos parámetros tangibles para observar el desempeño productivo de dichos modelos y por consiguiente dar el seguimiento correspondiente para realizar los ajustes necesarios a los modelos, set de datos, condiciones de negocio, modificación de hiperparámetros, etc. Los apartados anteriores nos apoyan a soportar que una estrategia de retención de clientes debe considerar diversos elementos como lo son:

- La estrategia y objetivos de negocio.
- Conocimiento de nuestros clientes, claridad en las característcas de clientes objetivo.
- Madurez y estrategia de datos.
- Desarrollo y/o robustecimiento de lineamientos de ética.

Con base a lo anterior podríamos implementar la adecuación de servicios, productos y valor agregado actuales basados en las necesidades del cliente (escuchar

al cliente) o bien la creación de campañas perfiladas de retención de clientes (estas pueden ser digitales o físicas) o la creación de un sistema de calificación del cliente (recompensas), etc.

Referencias

1. Verbeke, W., Martens, D., Mues, C., Baesens, B.: Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Syst Appl 38, pp. 2354–2364 (2011)
2. AbbasKeramati, H., SeyedMohammadMirmohammadi: Developing a prediction model for customer churn from electronic banking services using data mining. (2016)
3. Raj, P.: Telecom Customer Churn Prediction https://www.kaggle.com/pavanraj159/telecom-customer-churn-prediction (2018)
4. Guo-en X., Wei-dong J.: Model of customer churn prediction on support vector machine. Syst Eng Theory Pract, 28(1), pp. 71–77 (2008)
5. Chitarroni, H.: La regresión logística. https://racimo.usal.edu.ar/83/1/Chitarroni (2002)
6. Friedman, J., Hastie, T., Tibshirani, R.: Response to Mease and Wyner, Evidence Contrary to the Statistical View of Boosting, JMLR 9, 131–156, 2008. Journal of Machine Learning Research, 9, pp. 175–180 (2008)
7. Breiman, L.: Bagging Predictors. Machine Learning, 24(2), pp. 123–140 (1996)
8. Ho, T. K.: The random subspace method for constructing decision forests. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20(8), pp. 832–844 (1998)
9. Breiman, L.: Randomizing Outputs to Increase Prediction Accuracy. Machine Learning, 40(3), pp. 229–242 (2000)
10. Martínez-Muñoz, G., Suárez, A.: Switching class labels to generate classification ensembles. Pattern Recognition, 38(10), pp.1483–1494 (2005)
11. Dietterich , T. G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, 40(2), pp. 139–157 (2000)
12. Breiman, L.: Random forests. Machine Learning, 45, pp. 5–32 (2001)
13. Vishal Morde: XGBoost Algorithm: Long May She Reign! (2019)
14. Betancourt, G. A.: Las máquinas de soporte vectorial (SVMs). Scientia et technica, 1(27) (2005) 6895/4139
15. Código fuente del estudio realizado[1] (2001)
16. Exploratorio del set de datos [2] (2000)

[1] https://github.com/jorgeo80/UP_MCD/blob/master/MachineLearning/Proyecto/CHURNAnalysis.ipynb,Ingresar->Notebook Python

[2] https://datastudio.google.com/u/0/reporting/1W_R3VelXa8UmWgjxLdlksy7MydMhzD9f/page/YFqGBIngresar->Análisis numero dos realizado con la herramienta google studio