ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Sociales y Humanísticas

Deserción de clientes en el sector asegurador. Evidencia mediante el análisis de modelos de Machine Learning.

PROYECTO INTEGRADOR

Previo la obtención del Título de:

Economista con mención en gestión empresarial

Presentado por:

José Eduardo Centeno Arízaga

Adriana Graciela Vera Burgos

GUAYAQUIL - ECUADOR

Año: 2020

DEDICATORIA

Este proyecto se lo dedico a mi madre y mi abuela por la eterna protección y apoyo incondicional que me han dado.

José

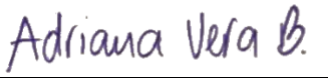
El presente proyecto lo dedico a mi familia en especial a mis abuelitos y padres que siempre me han motivado a superarme.

Adriana

DECLARACIÓN EXPRESA

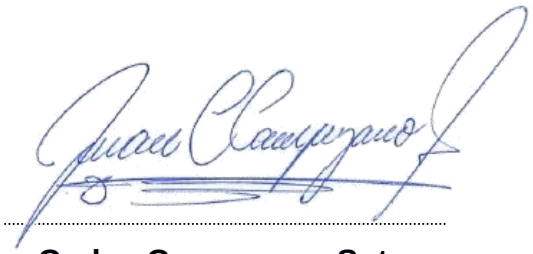
“Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; José Eduardo Centeno Arízaga y Adriana Graciela Vera Burgos damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

José Eduardo Centeno Arízaga



Adriana Graciela Vera Burgos

EVALUADORES



Juan Carlos Campuzano Sotomayor

PROFESOR /TUTOR DE LA MATERIA

RESUMEN

La retención de clientes es parte clave del manejo de la relación con los consumidores. Actualmente y debido a los efectos negativos de la COVID-19 en la economía ecuatoriana y la alta competencia en el sector asegurador, la tarea de estudiar y predecir el comportamiento de deserción de clientes ayuda a las empresas a asegurar renovaciones y mantener niveles de liquidez y solvencia óptimos . En este estudio se utilizó machine learning en cuatro modelos distintos como metodología para predecir la probabilidad de deserción en clientes de una aseguradora ecuatoriana en el ramo vehículos. Los modelos usados fueron árboles de decisión, GLM binomial, random forest y xtreme gradient boosting. Como resultados se obtuvo que el modelo xtreme gradient boosting fue el que mejor rendimiento presentó en las métricas usadas para comparar los modelos que fueron precisión y área bajo la curva, con una precisión del 70,77% y además a través del GLM binomial se descubrieron las variables más influyentes en la renovación de los clientes. Finalmente se concluye que mediante la predicción obtenida se puede segmentar clientes e implementar estrategias de ventas a los clientes que el modelo indique que son más propensos a desistir de la póliza y así mejorar la tasa de retención de la compañía y la relación con los clientes.

Palabras clave: deserción de clientes, machine learning, sector seguros, estadística.

ABSTRACT

Client retention is key in customer relationship management, nowadays due to the negative effects of COVID-19 in the Ecuadorian economy and the high competition in the insurance sector, the task of studying and predicting the customer churn helps companies to assure renovations and keeping optimum levels of liquidity and solvency. In this study machine learning was used in four different models as methodology to predict the probability of customer churn of an insurance company in the motor insurance policies. The models that were used are decision trees, GLM binomial, random forest and xtreme gradient boosting. The results obtained are that the xtreme gradient boosting had the best performance in the metrics used for comparing the models, which were precision and area under the curve, with a performance of 70,77% and also through GLM Binomial the most influential variables in the renovation of policies were discovered. Finally, the conclusion was that through the prediction obtained clients can be segmented into different groups to carry out sales strategies with the clients that the model indicates as most likely to churn and so improve the retention rate of the company and the relationship with customers.

Keywords: *customer churn, machine learning, insurance sector, statistics.*

ÍNDICE GENERAL

ÍNDICE DE GRÁFICOS

ÍNDICE DE TABLAS

CAPITULO 1

1. Introducción

La crisis económica provocada por la COVID-19 ha afectado las operaciones comerciales de los distintos sectores de la economía, siendo uno de estos el sector de seguros. Debido a la situación actual del país conseguir nuevos clientes es una tarea complicada, por lo que asegurar las renovaciones de los clientes de años anteriores se ha convertido una estrategia imprescindible para mantener la liquidez de una compañía y que pueda continuar con sus operaciones.

Al existir confinamiento y restricciones de movilidad, podría creerse que la siniestralidad del mercado asegurador debería bajar considerablemente, pero siempre es importante mantener producción para no perjudicar el funcionamiento del sector. Por esta razón, aprovechar al máximo las renovaciones constituye una prioridad y de cierta forma, aprovechar la crisis para obtener ventaja sobre la competencia.

Actualmente existen herramientas computacionales y estadísticas que pueden ayudar a crear métodos y modelos que permiten aplicar un análisis de datos y así conseguir información valiosa para mejorar el manejo del negocio y los clientes. Con esto, es posible anticiparse a decisiones de los clientes y poder crear campañas u ofrecer beneficios a individuos en específico (obteniendo eficiencia en los procesos), evitando así la deserción de clientes.

La estrategia más utilizada por las compañías de seguros en el país, para mejorar la efectividad de las renovaciones es ofrecer descuentos en el precio de la póliza. Esta práctica ha dado resultados positivos a lo largo de los años, pero resulta importante para mantener una situación adecuada de la empresa en tiempos de la COVID-19, conocer si en la actualidad es posible cambiar el sistema de descuentos, creando procesos más eficientes, en los cuales se realiza un análisis previo, antes de llevar a cabo una acción. Esto se puede lograr por medio de análisis de clientes a través de las herramientas computacionales disponibles.

1.1. Descripción del problema

El sector asegurador constituye una parte importante en la economía de un país. Debido a que impulsa la adquisición de bienes, disminuye incertidumbre por medio de la colectivización de riesgos, así mismo invierte las reservas técnicas en la economía nacional inyectando liquidez. Sin embargo, en Ecuador este sector apenas representa el 1,67% en 2019 como porcentaje del Producto Interno Bruto y en comparación a países desarrollados donde este porcentaje alcanza niveles de 30%.

En el Ecuador el mercado de seguros es competitivo como lo muestra en índice de Herfindahl e Hirschman alcanzando un valor de 738 en el 2019 (Federación Ecuatoriana de Empresas de Seguros, 2020) demostrando que es un mercado poco concentrado y de alta competitividad. La disputa por conseguir aumentar la participación en el mercado ha forzado a las empresas del sector a buscar distintas estrategias para aumentar la captación de clientes.

Debido a la creciente competitividad en los últimos años y la limitada innovación del sector es necesario recurrir a una estrategia de precios, la cual puede resultar peligrosa para compañías que no tengan los fondos suficientes para sobrevivir a esta situación en la que se encuentra el sector.

La situación actual que se vive en el mundo, los niveles de producción de los sectores productivos y por ende el mercado asegurador también se ha visto afectado. Las medidas de bioseguridad ante la pandemia traen consigo restricciones de movilidad vehicular y distanciamiento social, lo cual disminuye la exposición al riesgo de los automóviles, depreciando la valoración que tienen los clientes sobre el seguro, por esta razón la captación de nuevos clientes y la retención de la cartera son una prioridad y un desafío que exige una intervención inmediata a fin de lograr la sostenibilidad del sector.

Finalmente, las aseguradoras deben orientar sus esfuerzos a evitar la deserción de clientes con mecanismos de fidelización y trabajar en la captación de nuevos seguros en el marco de la depresión económica causada por la COVID-19.

1.2. Justificación y delimitación del problema

Los problemas a los que se debe hacer frente son el manejo de los flujos de caja, mantener una idónea atención al cliente con el mínimo riesgo para sus colaboradores, conseguir nuevos clientes en los distintos ramos y mantener los clientes de la compañía (Deloitte, 2020). Por estos desafíos imprevistos para la economía ecuatoriana y sus distintos sectores, mejorar la relación del cliente, la tasa de retención y por ende los ingresos de las empresas resulta importante.

La situación a la que se enfrentan las empresas aseguradoras las induce a analizar sus productos y las estrategias de ventas necesarias para mantener una operación del negocio sana. Tener una predicción sobre la deserción en el servicio permite crear mejores estrategias comerciales y asegurar la retención de mayor proporción de clientes.

De manera análoga el estado de empresas de otros sectores repercute al manejo del sector de seguros ya que no podrán renovar sus pólizas o les será más difícil cumplir con sus obligaciones con la aseguradora. Desde otro punto de vista las personas naturales se enfrentan a cambios en sus ingresos a raíz de la emergencia sanitaria, lo que generará también dificultades con los pagos de sus pólizas y la disminución en renovaciones de pólizas.

Cabe mencionar que la realización del proyecto es conveniente ya que en varios estudios anteriores se menciona que implementar estrategias de retención de clientes es 5 veces menos costoso que la puesta en marcha de estrategias comerciales para conseguir nuevos clientes. Además es necesario indicar que a través del proyecto planteado se podrá implementar modelos predictivos de deserción de clientes en empresas aseguradoras permitiendo así el manejo adecuado de su cartera y adicionalmente representa un nuevo instrumento mediante el cual las empresas usan sus bases de datos de clientes para su beneficio y toma de decisiones menos riesgosas ya que la estadística permite tener mayor precisión sobre los datos.

El siguiente gráfico representa la evolución tasa de renovación de la empresa aseguradora en la que se aplicará la metodología propuesta por este proyecto. Como se observa la tasa de renovación desde 2017 tiene una tendencia decreciente

sin embargo en el segundo trimestre del 2020 registra una tasa de renovación del 36% que puede estar vinculado con la situación actual por la COVID-19. Aunque el gráfico no incluye información de todas las empresas que pertenecen al sector permite percibir el comportamiento de la variable importante para el estudio.

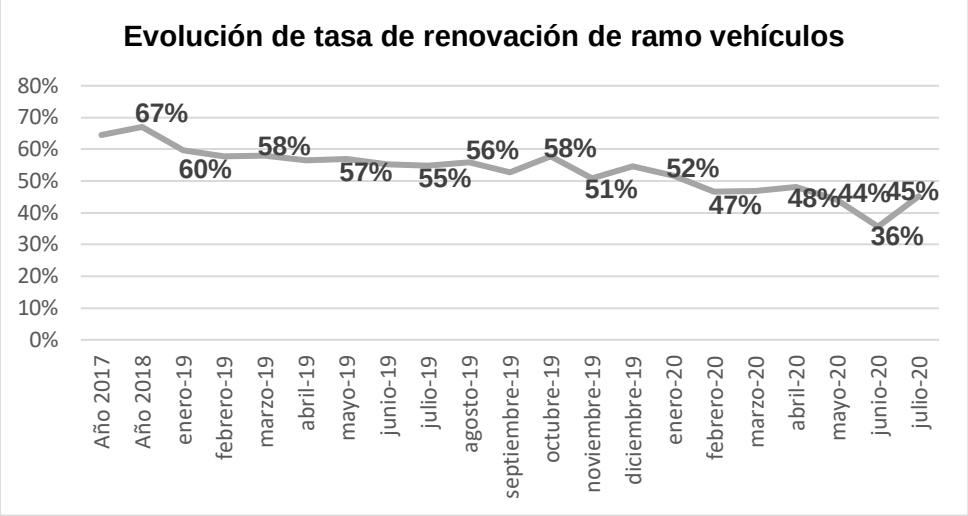


Ilustración 1.1 Evolución de la tasa de renovación de ramo vehículos

1.3. Preguntas de investigación

- ¿En qué proporción se podría mejorar la retención de clientes al implementar modelos de predicción de deserción de cliente?
- ¿Cuáles son las variables con mayor incidencia en el comportamiento de deserción de un cliente?

1.4. Objetivos

1.4.1. Objetivo General

- Estimar la probabilidad de renovación de las pólizas para el diseño de estrategias de ventas efectivas que aumenten los ingresos de las compañías de seguros mediante el uso de modelos de machine learning.

1.4.2. Objetivos específicos

- Diagnosticar el comportamiento del sector asegurador para su posterior tratamiento en base a técnicas de machine learning.
- Estimar varios modelos para un análisis de resultados y selección del más adecuado mediante el uso de machine learning.

- Identificar variables de interés que expliquen el comportamiento de los clientes para una toma de decisión anticipada.

1.5. Marco Teórico

1.5.1. Sector de Seguros en Ecuador antes de pandemia de la COVID-19

En el año 2019 la situación del sector asegurador había mejorado con respecto a años anteriores debido a un aumento de la demanda de seguros en su mayoría en el sector público y además de las múltiples estrategias que se implementaron por atraer individuos hacia el sector asegurador mediante creación de seguros más fáciles de aplicar. Los frutos de estas estrategias de las compañías del sector permitieron un crecimiento del 6,4% en 2019.

Como se observa en el gráfico desde el 2001 hasta la actualidad la evolución de la prima neta total emitida ha tenido un crecimiento sostenido con cierta disminución por el terremoto del 2016 que afectó a la economía en general del país.

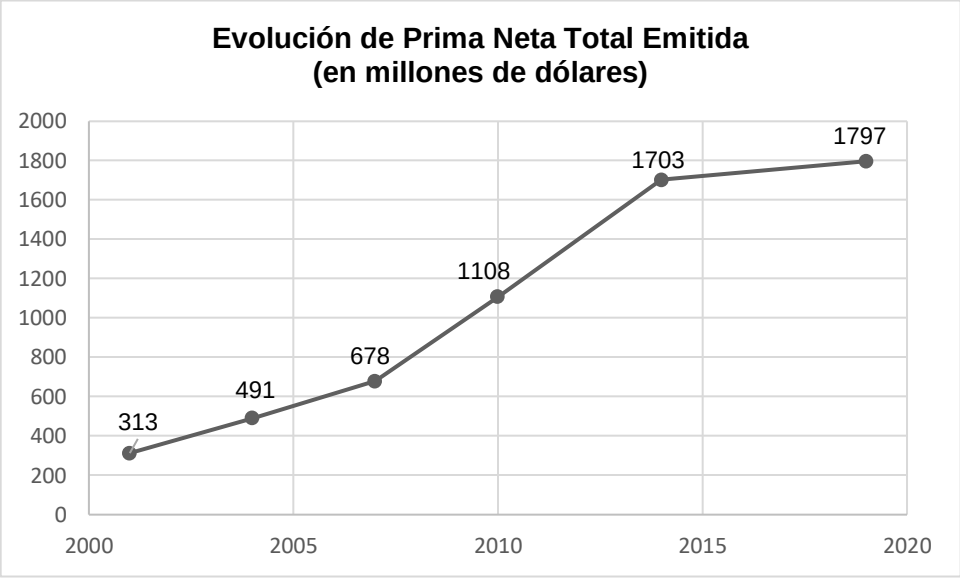


Ilustración 1.1 Evolución de la prima neta total emitida en Ecuador

Fuente: FEDESEG

Se ha encontrado que las razones principales por las cuales hay una baja cultura aseguradora en el mercado nacional son que existe una baja credibilidad de este servicio, una percepción de costos altos ya que lo perciben como un gasto mas no una inversión y finalmente un desconocimiento del funcionamiento de este servicio. Además, una percepción de la mayoría de los ecuatorianos sobre los seguros es que, no son necesarios sino una obligación por un tercero como son los bancos, al realizar un préstamo vehicular o hipotecario obligan a asegurar el bien (El Universo, 2018).

Más aún se tiene que el mercado asegurador no incrementa de tamaño en el país, resultado de esto es la participación del 1,67% en 2019 en el Producto Interno Bruto Nominal de la Prima Neta Emitida Nominal mismo nivel alcanzo en el 2014 (Federación EcuatorIana de Empresas de Seguros, 2020). En un análisis realizado por (Navarro & Wahren, 2017) se muestra que el índice de penetración del sector seguros en el Ecuador está por debado del promedio de latinoamérica el cuál es de 2,17% para el año 2015 cuando para ese año el promedio del país era del 1,65%. Los países latinoamericanos con mayor índice de penetración son Chile y Argentina con 4,71% y 3,11% respectivamente.

1.5.2. Efectos de la COVID-19 en Ecuador y el mercado asegurador

La COVID-19 es una enfermedad viral originada en China en la provincia de Hubei, la cual fue declarada pandemia a inicios del año 2020 por la OMS. Se ha convertido en un problema mundial, el cual ha acabado con miles de vidas y ha dejado al sistema de salud público y privado colapsado. Junto con el sistema de salud, la economía se dirige hacía una gran recesión, que pone en peligro a todos los sectores sin distinción.

La pandemia ha resultado muy costosa para el sector privado. El distanciamiento social, las restricciones de movilidad vehicular y la imposibilidad de trabajar presencialmente han generado grandes pérdidas cercanas a los 500 millones de dólares y un incremento del 4.4% al 13.3% en el desempleo nacional entre junio 2019 y 2020 según datos de (INEC, 2020).

Según la Federación Ecuatoriana de Empresas de Seguros la venta de pólizas en general ha caído desde el inicio de la pandemia y la declaratoria de emergencia en el país. Los datos recopilados por FEDESEG muestran que existe en el mes de junio de 2019 al junio de 2020 una variación en las primas netas emitidas por aseguradoras del Ecuador -7,7% alcanzando una diferencia de 69,3 millones de dólares.

Sin embargo, la pandemia no solo ha generado pérdidas económicas, sino que también se han perdido muchas vidas. En particular se registra que en los meses de marzo, abril y mayo la mortalidad se incrementó en un 400%. Esta sobre mortalidad afecta directamente al sector de seguros, específicamente al ramo de vida, aumentando su siniestralidad. A pesar de que en esta investigación, no se aplicará el análisis en este ramo, es importante mencionar, que la situación actual está generando pérdidas en el sector, las cuales, con el fin de sobrepasar la crisis, deben ser solventadas con otros ramos que no se vean afectados directamente por la pandemia, estrategias de reducción de costos y métodos para conservar clientes y evitar cualquier fuga adicional de capital.

(Revista Vistazo, 2020) señala que 19 de las empresas aseguradoras del país han registrado una caída en el volúmen de las primas emitidas netas de enero a mayo del presente año comparandolas con los niveles de emisión del año 2019. Esto se traduce a una afectación en el 63% del sector mientras que la proporción restante tiene incrementos no significativos.

1.5.3. Retención de clientes como estrategia en el mercado de seguros

La retención de clientes es una práctica muy usada en diferentes sectores incluido en el sector seguros, donde es un tema de alta importancia dada el elevado grado de competencia. La incorporación de nuevos clientes en la cartera involucra costos de emisión, inspección, corretaje, etc. Kotler (como se citó en (Guadarrama y Rosales 2015)), comenta sobre cómo el consumidor final se ve beneficiado cuando las empresas focalizan sus prioridades hacia ellos. Esta es una forma de

aportar valor a los procesos de una compañía, aumentando la fidelización de los clientes.

Como se menciona en (Guadarrama y Rosales 2015) “Una organización pierde el 50% de sus clientes cada cinco años y por regla general captar un cliente nuevo requiere un esfuerzo cinco veces mayor que conservar a uno ya existente (Burnett, 2002)”.

Debido a esto,(Peppers & Rogers, 2006), sugieren realizar segmentación en la cartera de clientes y determinar indicadores clave sobre ellos: Los clientes que generan un alto beneficio y los que generan bajos beneficios, el tipo de comunicación que se debe tener con ambos y los productos que resulta más conveniente ofrecerles. Así mismo en (Weinstein, 2001) la segmentación de clientes permite implementar estrategias de marketing de retención distintas, enfocando las estrategias más efectivas y rentables en los de clase alta y con alta incidencia en sus ventas y para los clientes eventuales o esporádicos se deberían implementar programas costo eficientes y así no desperdiciar recursos en bajas probabilidades de recompra.

Para (Weinstein, 2001) para maximizar su valor las empresas deben desarrollar la relación con el cliente teniendo en cuenta los siguientes elementos: fidelidad, retención, valor, satisfacción y rentabilidad.

En un estudio realizado por el European Central Bank (citado en Miranda, Rey y Weber 2005) se logró evidenciar como se ve afectado el beneficio acumulado en los flujos futuros de esta institución bancaria. En el gráfico se puede apreciar que mientras mayor sea la retención de clientes, mayor será la acumulación de beneficios futuros. El gráfico muestra una proyección, para distintos índices de retención, dentro de un horizonte temporal de 25 años. También es posible evidenciar en el gráfico que pequeñas variaciones el este índice, pueden aportar grandes diferencias en los beneficios, lo cual muestra la importancia de la retención de cartera.

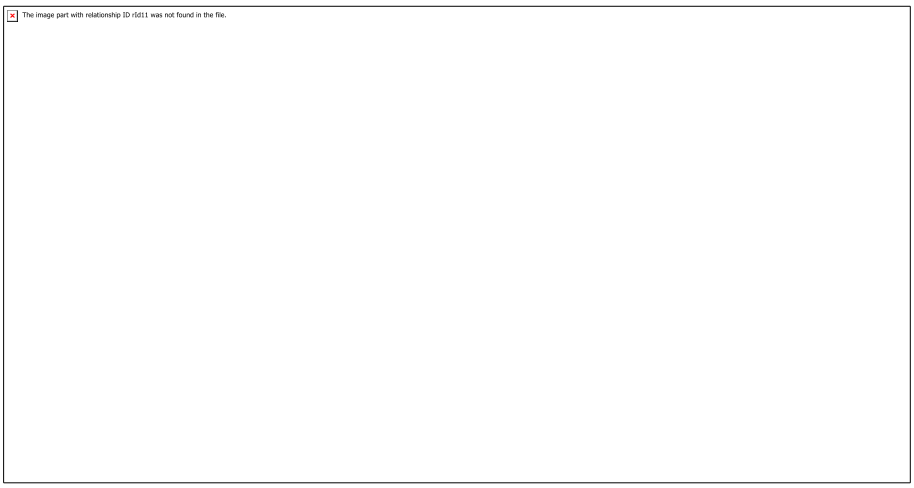


Ilustración 1.2 Beneficio acumulado respecto a la tasa de retención

Fuente: (Miranda, Rey, & Weber, 2005)

1.5.4. Metodologías utilizadas en estudios previos sobre fuga de clientes

En (Miranda, Rey, & Weber, 2005) se usa la metodología de Support Vector Machines, que consiste en encontrar un hiperplano de separación que divida el espacio de entrada en dos regiones maximizando el margen de separación y minimizar el error de clasificación. En este trabajo se encuentra que SVM es más efectiva que otro tipo de modelo para reconocer los clientes más propensos a la fuga en una empresa que se dedica a la actividad bancaria.

El estudio de (Spiteri & Azzopardi, 2018) se enfoca en una empresa aseguradora de vehículos y busca identificar los clientes con riesgo de irse y el tiempo que tiene la compañía hasta antes de la cancelación para tomar medidas para retener a esos clientes. Como conclusiones del estudio se tiene que la metodología de Random Forest fue la mejor para predecir la fuga de clientes con una exactitud del 89,9%. Además, pudieron concluir que las pólizas que tenían altas probabilidades de ser canceladas eran de clientes con altos índices de reclamos.

En (Gunther, Tvete, Aas, Sandnes, & Borgan, 2014) utilizan datos mensuales de pólizas de clientes de una empresa aseguradora para construir un modelo de predicción de fuga de clientes cada mes mediante una regresión logística con GAM. Como parte de sus resultados obtuvieron que a menor antigüedad del cliente hay

un mayor riesgo de fuga, y los clientes que han regresado a la compañía luego de una cancelación son más probables a una cancelación posterior. Como punto importante consideran a las variables de descuentos y los cambios de la variable en el tiempo como más incidente en el modelo desarrollado en el estudio.

Los autores (He, Xiong, & Tsai, 2020) realizan el planteamiento de métodos de Machine Learning para obtener el mejor modelo que permita predecir la fuga de clientes de una empresa aseguradora para que se tomen las decisiones acertadas antes de la fecha de vencimiento de los clientes en riesgo de fuga. Como métrica para escoger el método que mejor se acopla a la base de datos usan el Área bajo la curva AUC por sus siglas en inglés la cual analiza aleatoriamente valores positivos y negativos obtenidos con el modelo y calcula los falsos positivos y verdaderos positivos y sus ratios, analizando así el desempeño del modelo.

Al estudiar la fuga de clientes se necesita obtener características o definir perfiles de clientes para identificar que aspectos influyen en mayor grado a que un cliente sea propenso a irse con la competencia. En el análisis realizado por (Huigevoort, 2015) sobre la fuga de clientes en una empresa aseguradora holandesa de salud, se menciona que una de las metodologías que más ayuda a poder identificar e interpretar las características son las regresiones logísticas, ya que mediante el planteamiento de las regresiones se puede conocer cuáles son las variables que tienen más peso en la decisión de fuga del cliente. Un beneficio adicional mencionado de esta metodología es que permite obtener información que puede ser fácilmente usada por otros departamentos

La elección de variables debe de ser rigurosa y estar basada en un análisis de la situación de la empresa que se usa para el estudio y el contexto en el que se aplican los modelos, ya que regulaciones del país y formatos del seguro pueden afectar a como se analiza una variable o la incidencia de esta en la fuga de clientes de una empresa en específico. Para el caso de seguros en la literatura revisada las variables comúnmente usadas pueden clasificarse en sociodemográficas, variables de la compañía, del cliente y variables relacionadas al producto. La primera agrupación contiene variables como edad, género, estadio civil, ciudad. La segunda agrupa información sobre cantidad de pólizas del cliente, número de reclamos, años

de permanencia del cliente en la compañía, y por último las que se relacionan con los productos como el valor de la prima, descuentos, tipo de producto, siniestros y sus montos, entre otros.

CAPITULO 2

2. METODOLOGIA

Recolección de datos

Los datos pertenecen a una empresa aseguradora del Ecuador que ofrece seguros de varios ramos, como lo son vehículos, hogar, accidentes personales, lucro cesante, etc. Sin embargo, para esta investigación se usaron datos del ramo de vehículos ya que esta categoría era la que presentaba mayor cantidad de datos disponibles y es el ramo de mayor comercialización de la empresa. La información de la base de datos comprende el periodo de enero de 2016 a junio de 2020.

2.1. Descripción de las variables utilizadas

A continuación, se describen las variables usadas para el planteamiento de los modelos.

Tabla 2.1 Descripción de variables

Nombre de variable	Descripción de la variable
Prima	Valor que paga el cliente por la póliza
Producto	Tipo de producto contratado por el cliente
Tipo de vehículo	Tipo de vehículo asegurado en la póliza
Tipo de Agente	Canal comercial de donde proviene el cliente (financiero, directo o bróker)
Antigüedad Vehículo	Años transcurridos desde la fabricación del vehículo
Pólizas contratadas	Número de pólizas contratadas por el cliente en la aseguradora
Antigüedad del cliente	Número de años del cliente en la empresa aseguradora
Ciudad	Ciudad en la que se generó la póliza
Género	Género de quién contrata la póliza (femenino o masculino)
Edad	Edad de quién contrata la póliza
Color de vehículo	Color del vehículo asegurado en la póliza
Renovación	Indica si el cliente renovó o no renovó la póliza
Siniestralidad	Porcentaje de la prima que es desembolsado en siniestros por la aseguradora

Suma Asegurada	Valor por el cual el vehículo está asegurado en la póliza. Debe ser aproximado al valor comercial del vehículo.
Disminución de suma asegurada	Indica si el cliente decide disminuir la suma asegurada del vehículo, en comparación a la vigencia anterior.

En los anexos del documento se pueden encontrar gráficos referentes al comportamiento de las variables usadas en la investigación.

2.2. Partición y preprocesamiento de datos

Siguiendo la metodología de machine learning se decidió realizar una partición de los datos de 75% training y 25% testing, con el fin de tener una proporción considerable de datos al momento de probar los resultados de la modelación, permitiendo crear las matrices de confusión, las cuales serán usadas para comparar la predicción de los modelos.

Se utilizó también la técnica de K-Fold Cross validation, en la cual se crearon 5 folds. Por medio de la partición y validación cruzada es posible disminuir el sobreajuste en las predicciones, por lo tanto, estas herramientas ayudan a tener estimaciones menos volátiles y sensibles a datos externos al modelo. La K-Fold CV, realiza nuevas particiones de training y testing, distintas en cada iteración, sobre los datos que fueron elegidos para ser modelados. La K indica el número de iteraciones que se usarán.

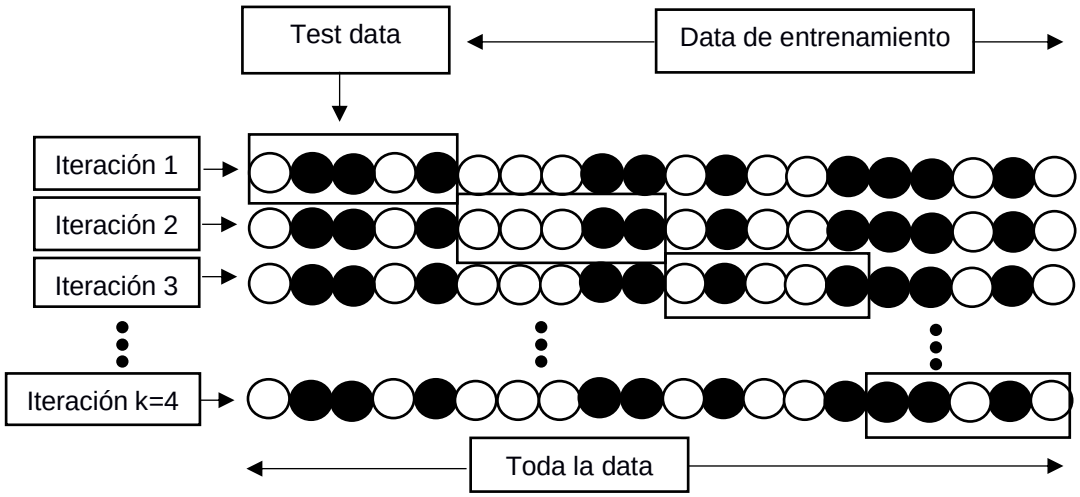


Ilustración 2.1 K-Fold Cross validation

2.3. Desarrollo de los modelos de Machine Learning

Machine learning es una rama de la inteligencia artificial que permite programar a computadoras para optimizar su precisión en el análisis de datos o experiencias pasadas. Permite capturar patrones o características específicas de la data con el objetivo de adquirir conocimientos nuevos de los datos. A través de este aprendizaje el sistema creado adquiere habilidades para aprender y adaptarse a cambios y así proveer nuevas soluciones para estos nuevos escenarios (Alpaydin, 2014).

2.3.1. GLM Binomial

Conocidos también como modelos de regresión logística los modelos de GLM permiten predecir o estimar el comportamiento de una variable binaria o categórica. Debido a que la variable dependiente es binaria el modelo obliga a una formulación no lineal y así obtener valores estimados dentro de 1 y 0 (Stock & Watson, 2012).

Un modelo Logit con varios regresores está dado de la siguiente forma:

$$\Pr(Y = 1|X_1, X_2, \dots X_k) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \quad (2.1)$$

$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

La ventaja de usar este modelo para predecir la deserción de clientes es que produce información interpretable y los coeficientes de las variables del modelo permiten observar la participación estimada de la variable en la deserción de clientes y así poder obtener un perfil de cliente propenso de deserción. Para poder interpretar los resultados de este modelo, es necesario calcular las probabilidades estimadas y las diferencias de estas.

2.3.2. Arboles de decisión

Son estructuras en forma de árbol que representan sets de decisiones capaces de generar reglas de clasificaciones para la información de una base de datos. La precisión de árboles de decisión en la predicción y estudio de deserción de clientes dependiendo de la forma de los datos puede ser alta (Vafeiadis, Diamantaras, Sarigiannidis, & Chatzisavvas, 2015).

Los árboles de decisión son comúnmente utilizados para resolver problemas de clasificación. Su forma de funcionamiento está basada en realizar una serie de preguntas en base a las variables independientes, a medida que cada pregunta es contestada, surge una siguiente hasta llegar a una conclusión. El modelo inicia con una pregunta principal llamada raíz, de donde surgen el resto de las preguntas que se van generando. Todas las preguntas generadas a partir de la raíz son llamadas nodos internos. Luego, a la conclusión de cada nodo interno se le llama hoja o nodo terminal.

Existen varios algoritmos para realizar árboles de decisión, para la siguiente investigación se utilizó el método “Bagged CART”, el cual no posee hiperparámetros de ajuste.

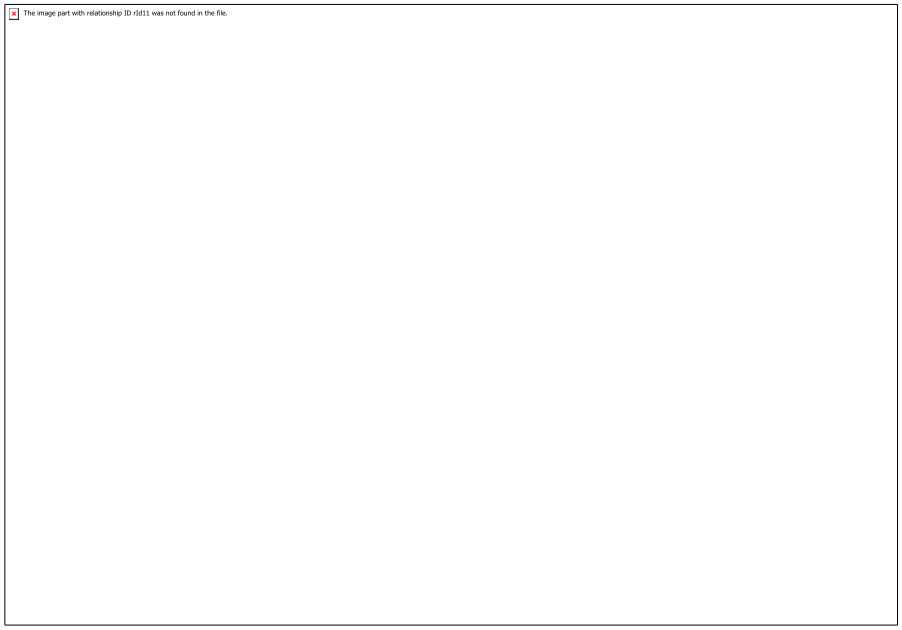


Ilustración 2.2 Problema de clasificación de árboles de decisión

2.3.3. Random Forest

La estructuración del random forest es bastante parecida a la mencionada anteriormente en arboles de decisión, porque está basada en una partición recursiva binaria. Estas particiones van tomando decisiones binarias en las cuales dividen a las variables en dos partes, de esta forma, se va produciendo un mapa de decisiones. El nodo “raíz” o principal del modelo, es el punto de partida para todas

las decisiones. Todos los nodos se parten en dos ramas de decisiones. Cuando un nodo ya no posee partición, es llamado nodo terminal, el cual indica la probabilidad estimada que se quiere conocer (Cutler, Cutler, & John, 2012).

Hiperparámetros a configurar:

- Mtry: # de predictores seleccionados de forma aleatoria. Los valores a escoger son mínimo 1 y máximo el número de variables independientes que tenga el modelo
- Split rule: Regla de separación. Para problemas de clasificación se puede usar las reglas: "gini", "extratrees" o"hellinger".
- Minimum node size: Tamaño de nodo mínimo

2.3.4. Xtreme Gradient Boosting (XGB)

(Malik, 2020) define a este modelo como una versión mejorada del modelo Gradient Boosting machines, la cual está diseñada para tener mayor eficiencia computacional, de manera que puede generar estimaciones de forma más rápida. El algoritmo trabaja de forma secuencial, ensamblando modelos, de esta forma aprende en cada iteración. Para generar las estimaciones, intenta convertir los predictores débiles en predictores fuertes, de esta forma aumenta la precisión del modelo.

Hiperparámetros a configurar:

- Nrounds: # de iteraciones de Boosting
- Max Depth: Profundidad máxima del árbol
- Eta: Ratio de aprendizaje para llegar al punto óptimo entre desviación y varianza.
- Gamma: Porcentaje mínimo de pérdida
- Column sample by tree: Ratio de remuestreo por columna
- Minimum child weight: Suma minima de pesos
- Subsample: Porcentaje de remuestreo

2.4. Métricas para la comparación de modelos

2.4.1. Estadístico Kappa

Permite comparar la precisión del modelo contra la precisión de un sistema aleatorio. Se obtiene mediante la siguiente fórmula.

$$kappa = \frac{Precisión\ total - Precisión\ aleatoria}{1 - Precisión\ aleatoria} \quad (2.2)$$

Donde la precisión total y precisión aleatoria se obtienen

$$PT = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.3) \qquad PA = \frac{VN*NP+VP1*VP2}{Total*Total} \quad (2.4)$$

Donde:

TP= Total positivos

TN=Total negativos

FP=Falsos negativos

FN=Falsos positivos

VN= Verdaderos negativos

NP=Negativos predichos

VP1=Verdaderos predichos

VP2=Verdaderos Positivos

Total=Total de casos

2.4.2. Sensibilidad

Muestra que tan bien predice los clientes que renuevan la póliza ya que mide la proporción de los verdaderos positivos que si fueron correctamente asignados como positivos por cada modelo estudiado (Gaur & Dubey , 2018).

$$S = \frac{TP}{TP+FN} \quad (2.5)$$

2.4.3. Especificidad

Refleja que tan bien se predice los clientes que no renuevan la póliza ya que esta métrica es la proporción de verdaderos negativos que fueron clasificados como negativos por los modelos aplicados (Gaur & Dubey , 2018).

$$SP = \frac{TN}{TN+FP} \quad (2.6)$$

2.4.4. Curva ROC (receive operating characteristics curve)

Es un método estadístico que permite analizar la precisión de tests realizados a los datos, a través de la curva ROC se puede identificar en que punto se obtiene la más alta especificidad y sensibilidad (Cerda & Cifuentes, 2011), es decir la capacidad de los modelos planteados para diferenciar los clientes propensos a fugas de los no propensos. Se forma graficando en el axis y la sensibilidad y en el axis x 1-especificidad de distintos valores de las pruebas realizados en el estudio. Una curva ROC refleja buenos resultados de la predicción cuando se encuentra alejada de la línea donde y=x son iguales ya que ubicarse en esos puntos de la diagonal significa que el modelo no es bueno discriminando los casos ya que obtiene tantos verdaderos positivos como falsos positivos.

2.4.5. Área bajo la curva AUC

Esta métrica se la consigue como su nombre lo explica calculando el área por debajo de la curva y este valor explica la compensación entre tasa de verdaderos positivos y tasa de falsos positivos (He, Xiong, & Tsai, 2020). Esta métrica es considerada mejor para comparar rendimientos de modelos que han usado una base de datos donde las clasificaciones son desequilibradas. Esto implica que cuando se comparan modelos mediante esta métrica el que tenga una AUC mayor o cercana a 1 es el que mejor predice la clasificación de los caso

CAPITULO 3

3. RESULTADOS Y ANALISIS

Al aplicar las metodologías mencionadas en el capítulo anterior, junto con el direccionamiento hacia resolver los problemas planteados, se procedió a modelar los datos. Los resultados obtenidos en los 4 modelos propuestos son los siguientes.

Tabla 3.1 Tabla de Resultados de modelos

Modelo	Precisión	Intervalo de Confianza	Estadístico Kappa	Sensibilidad	Especificidad	AUC
GLM	0.6773	(0.6686, 0.6858)	0.1933	0.8851	0.2856	0.5857
Arboles de decisión	0.6979	(0.6894, 0.7063)	0.2774	0.8558	0.4003	0.6395
Random Forest	0.7053	(0.6969,0.7137)	0.321	0.8219	0.4856	0.6522
Xtreme Gradient Boosting	0.7078	(0.6994,0.7162)	0.3242	0.8280	0.4814	0.6628

Se evidenció que, en todas las métricas consideradas para el análisis, precisión, sensibilidad, especificidad y área bajo la curva, los modelos de Xtreme Gradient Boosting XGB y Random Forest tuvieron los mejores resultados y bastante parecidos. Finalmente se escogió al XGB como el método para la resolución del problema planteado, puesto que tuvo la mayor precisión y área bajo la curva con 0.7078 y 0.6628 respectivamente. La métrica de precisión mide que tan bien el modelo identifica a los clientes de la base como renovados o no renovados en conjunto.

Tabla 3.2 Matriz de confusión de modelo XGB

		Valores reales	
Predicción		**No renovados**	**Renovados**
	No renovados	6196	2059
	Renovados	1287	1911

En la tabla 3.2 se muestra la precisión del modelo XGB al intentar predecir las pólizas que serán y no serán renovadas. La diagonal principal de la matriz indica los casos para los cuales el modelo dio una predicción acertada. Se puede observar en la primera columna, que se pudo predecir 6.196 de 7.483 pólizas que no serían renovadas, lo cual predice un 82,80% de los casos (Sensibilidad). En la segunda columna se pudo predecir 1.911 de 3.970 pólizas que serían renovadas, lo cual es un 48,14% de los casos (Especificidad).

Si se aplicara el modelo para la toma de decisiones de la compañía, la campaña comercial o de marketing se aplicaría en las pólizas de la primera fila las cuales dan un total de 8.255 (6.196 + 2.059 = 8.255). Si arbitrariamente, se supone un 50% de efectividad en la campaña aplicada, se lograría renovar 7.068 pólizas (6.196*50% +2.059 +1.911 =7.068). Las pólizas renovadas pasarían de 3.970 a 7.068, lo cual simbolizaría un incremento del 178% en el índice de renovación.

Es importante mencionar que los valores que se presentan en la tabla anterior pertenecen a la proporción de la base de testing que ingresa al modelo XGB por lo que se puede observar que el modelo ha aprendido de los datos a través del machine learning.

3.1. Muestra de probabilidades estimadas

Con los resultados del modelo escogido anteriormente, se procedió a realizar estimaciones de la probabilidad de renovación. La siguiente tabla contiene 10 ejemplos de la base de datos escogidos aleatoriamente. Para obtener una mejor presentación de las probabilidades estimadas se decidió mantener un grupo de variables fijas en los siguientes valores como se detalla a continuación:

- Producto: Auto Individual
- Antigüedad de cliente: 1 año
- Tipo de vehículo: Automóvil
- Tipo de Agente: Broker
- Pólizas contratadas en la compañía: 1
- Ciudad: Guayaquil
- Género: Masculino
- Pagos: $0
- Siniestralidad: 0%

Tabla 3.3 Probabilidad de renovación estimada

Prima	Antigüedad vehículo	Color	Edad	Suma asegurada	Dismin. S.A.	Probabilidad renovación estimada	Renovación Observada
$ 732.06	2	Gris	81	$24,500.00	1	**81.19%**	**1**
$ 875.65	3	Gris	67	$30,500.00	1	**75.66%**	**1**
$ 855.00	0	Gris	30	$25,000.00	1	**74.69%**	**0**
$ 952.00	4	Gris	58	$28,000.00	0	**68.95%**	**1**
$ 936.70	3	Gris	39	$34,000.00	1	**68.10%**	**1**
$ 595.08	6	Gris	63	$18,000.00	1	**67.81%**	**1**
$ 570.00	5	Gris	77	$15,000.00	1	**67.55%**	**1**
$ 760.75	5	Gris	65	$17,900.00	0	**67.24%**	**1**
$ 549.10	2	Gris	57	$17,000.00	1	**66.97%**	**1**
$1,193.40	3	Gris	35	$35,100.00	1	**64.03%**	**1**

Con esta muestra realizada en la tabla 3.3, se puede observar las probabilidades estimadas con el modelo para un grupo específico de observaciones, contrastadas con el dato observado de renovación.

El modelo GLM Binomial permitió identificar cuáles eran las variables que en mayor medida aumentan la deserción de clientes en las pólizas de vehículos. De esta manera, no solo se logró conseguir una predicción de casos de deserción, sino también comprender a la cartera de la compañía y sus decisiones.

Tabla 3.4 Resultados GLM

		Renovación	
Predictores	*Odds Ratios*	*CI*	*p*
Intercepto	0.48	0.46 – 0.49	**<0.001**

Producto Auto-Individual	1.58	1.24 – 2.01	**<0.001**
Producto Autos-Individual Abierto	1.40	1.23 – 1.60	**<0.001**
Prima	1.20	1.16 – 1.23	**<0.001**
Antigüedad cliente	1.15	1.12 – 1.19	**<0.001**
Tipo vehículo camioneta	0.97	0.94 – 0.99	**0.004**
Tipo vehículo todo terreno	0.96	0.93 – 0.98	**<0.001**
Tipo vehículo van	0.97	0.94 – 0.99	**0.018**
Tipo agente directo	1.04	1.02 – 1.07	**<0.001**
Tipo agente financiera	0.85	0.68 – 1.06	0.149
Antigüedad vehículo	0.90	0.87 – 0.92	**<0.001**
Pólizas contratadas	0.92	0.89 – 0.95	**<0.001**
Ciudad Guayaquil	1.62	1.51 – 1.73	**<0.001**
Ciudad Manta	1.17	1.13 – 1.21	**<0.001**
Ciudad Quito	1.47	1.37 – 1.58	**<0.001**
Resto Sierra	1.06	1.02 – 1.10	**0.002**
Género masculino	0.95	0.93 – 0.97	**<0.001**
Color vehículo azul	1.53	1.38 – 1.69	**<0.001**
Color vehículo beige	1.32	1.22 – 1.42	**<0.001**
Color vehículo blanco	1.83	1.55 – 2.16	**<0.001**
Color vehículo celeste	1.15	1.10 – 1.20	**<0.001**
Color vehículo dorado	1.40	1.27 – 1.53	**<0.001**
Color vehículo gris	2.25	1.83 – 2.78	**<0.001**
Color vehículo marrón	1.18	1.12 – 1.24	**<0.001**
Color vehículo naranja	1.07	1.04 – 1.10	**<0.001**
Color vehículo negro	1.69	1.46 – 1.95	**<0.001**
Color vehículo rojo	1.61	1.43 – 1.82	**<0.001**
Color vehículo verde	1.19	1.12 – 1.26	**<0.001**
Color vehículo vino	1.42	1.28 – 1.57	**<0.001**
Edad	1.14	1.11 – 1.16	**<0.001**
Suma asegurada	1.03	1.00 – 1.06	0.069
Disminución suma asegurada	1.09	1.05 – 1.12	**<0.001**
Pago	0.97	0.88 – 1.06	0.509
siniestralidad	0.76	0.69 – 0.84	**<0.001**

A través de los resultados obtenidos de la tabla anterior se puede observar que hay un conjunto de variables que tienen una mayor incidencia en la variable dependiente, la variable dicotómica renovación o no renovación. Dicho lo anterior se planteó un perfil de cliente propenso a la deserción con los resultados del análisis del odds ratio detallado a continuación.

- El Producto Auto Individual tiene una mayor probabilidad de deserción.
- A medida que aumenta la antigüedad del cliente en la compañía, aumenta su probabilidad de deserción, esto puede darse debido a que muchas personas consideran el seguro una obligación por instituciones financieras para la aprobación de un crédito y cuando se termina de cancelar el crédito desisten de este servicio.
- Guayaquil y Quito son las ciudades con clientes menos fieles, esto puede ser explicado por la gran cantidad de aseguradoras que operan en estas ciudades permitiendo que la competencia sea mayor que en ciudades más pequeñas.
- Los vehículos de color blanco y gris tienen mayor probabilidad de deserción en comparación al resto.
- El indicador sobre cambios en la suma asegurada entre vigencias de la póliza aumenta la probabilidad de deserción en mayor medida que otras variables, este cambio en la suma asegurada del vehículo podría ser causado por una pérdida de poder adquisitivo del cliente que lo conlleva a desistir del servicio convirtiendo al cliente en un desertor.
- Las variables edad y género no mostraron una incidencia grande como para ser incluidas en el perfil del cliente.

CAPITULO 4

4. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones

En consonancia con lo expuesto en los resultados se debe resaltar que los modelos de predicción de deserción de clientes son una herramienta de gran utilidad para identificar los clientes más propensos a desertar o en el caso concreto de este estudio a no renovar su póliza. Esto permite implementar estrategias focalizadas a una segmentación de clientes de manera que estas sean eficaces y con una mayor posibilidad de ser efectivas para fidelizar al cliente en la compañía aseguradora.

Con los resultados obtenidos de la modelación, se propone armar estrategias de ventas más agresivas, con el fin de maximizar la retención de clientes y por ende los ingresos de la compañía. En la actualidad, una práctica común en el sector, a nivel nacional e internacional, es aplicar un pequeño descuento plano, a cada cliente luego de su primer año en la compañía, dependiendo de su nivel de siniestralidad, tomando sólo esta variable como medida de diferenciación. Lo que se propone, es tomar en cuenta también su probabilidad de deserción al momento de tomar dicha decisión. De esta forma, la compañía podría ahorrar dinero, no ofreciendo un descuento a un cliente cuya probabilidad de renovación es cercana a 1, mientras que podría ofrecer una mayor tasa de descuento a otro cliente cuya probabilidad de renovación es baja. Todo esto tomando también en cuenta la actual diferenciación que se realiza en cuanto a la siniestralidad. Por esta razón lo que se recomienda es juntar ambas metodologías, con el fin de usar toda la información disponible para mejorar la utilidad de la compañía.

Los modelos aplicados en este estudio y su análisis junto con las métricas pueden ser usados en investigaciones de deserción en demás compañías del sector. Además, al aplicar los modelos predictivos meses antes de la fecha de vencimiento de las pólizas se puede tener mayor visibilidad de los perfiles de clientes que tienen mayor probabilidad de deserción y así realizar un seguimiento

sobre la satisfacción del cliente con el servicio contratado y producto de esto obtener un mejor entendimiento y perspectiva sobre algún problema que esté causando esa baja probabilidad de renovación dando paso a una mejora continua del servicio que conlleve a mejorar la tasa de retención de clientes en la compañía.

Se debe agregar también que el sector asegurador no está lo suficientemente desarrollado en el país comparado con otros países de la región como se ha mencionado anteriormente y esto en parte debido a la falta de cultura aseguradora en el país. Para el desarrollo del sector se necesita mejor educación en temas se seguros y además mejor difusión de los tipos de seguros que se ofrecen actualmente y facilitar la comprensión y manejo de una póliza y con esto aumentar la demanda de este servicio y a su vez disminuir el riesgo para las personas naturales y jurídicas ante cualquier siniestro.

4.2. Recomendaciones

Se recomienda para próximas investigaciones sobre temas similares al de este estudio realizar entrevistas a tomadores de decisiones de empresas del sector para conocer cuál es su perspectiva de la situación actual y también puedan dar su punto de vista sobre cuáles serían las variables que consideran ser importantes para modelos de predicción de deserción.

Los modelos planteados se pueden aplicar en distintos sectores de la economía sin embargo hay que entender el giro de negocio ya que dependiendo del modelo de negocio la información sobre la deserción de un cliente puede estar representada de manera diferente y además las variables a considerar al análisis deben ser estudiadas y estar sustentadas por estudios académicos previos o con entrevistas a expertos y conocedores del funcionamiento del sector.

4.3. Limitaciones

La primera limitación que se encontró fue la cantidad de variables a las que se tuvo acceso. En base a otras investigaciones, resulta de gran utilidad incluir en la modelación variables con respecto a la situación socioeconómica del cliente, por ejemplo: nivel de ingresos, profesión o ocupación, lugar de residencia, etc. También hubiera sido de gran ayuda tener más variables que indiquen la situación personal

de cada cliente, por ejemplo: estado civil, número de hijos, fecha de obtención de licencia de conducir, etc.

Además, el uso información sobre un solo ramo de una empresa para el análisis impide que se pueda conseguir una gran cantidad de datos para que el algoritmo de los modelos trabaje, por lo que se considera que si se poseyera una mayor cantidad de registros de clientes de distintos ramos hubiera permitido realizar un análisis más extensos considerando variables propias de cada ramo.

BIBLIOGRAFIA

Weinstein, A. (2001). Customer retention: A usage segmentation and costumer value approach. *Journal of Targeting, Measurement and Analysis for Marketing*, 259-268.

Miranda, J., Rey, P., & Weber, R. (2005). Predicción de fugas de clientes para una institución financiera mediante Support Vector Machines. *Revista ingeniería de sistemas*, 49-68.

Spiteri, M., & Azzopardi, G. (2018). Customer churn prediction for a motor insurance company. *IEEE*, 173-178.

Gunther, C.-C., Tvete, I., Aas, K., Sandnes, G. I., & Borgan, Ø. (2014). Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*, 58-71.

He, Y., Xiong, Y., & Tsai, Y. (2020). Machine Learning Based Approaches to Predict Customer Churn for an Insurance Company. *The Institute of Electrical and Electronics Engineers, Inc. (IEEE)*, 1-6.

Deloitte. (2020). *Sector Asegurador Riesgos e implicaciones potenciales, derivados del impacto del COVID-19.* Deloitte S-Latam, S.C.

Huigevoort, C. (2015, Abril). Customer churn prediction for an insurance company.

Endara, V. (2020, marzo 09). El sector asegurador creció 6,4% en 2019 Esta noticia ha sido publicada originalmente por Diario EL TELÉGRAFO bajo la siguiente dirección: https://www.eltelegrafo.com.ec/noticias/economia/4/sector-asegurador-crecio-2019 Si va a hacer uso de la misma, por. *El Telégrafo*.

El sector asegurador ecuatoriano en cifras Año 2019. (2020). *Federación Ecuatoriana de Empresas de Seguros.* Retrieved from https://6aab8a7f-de25-4e01-bf7a-2697d046daa5.filesusr.com/ugd/f39f07_8efc328dcab543f3bf55749413963d8e.pdf

Federación Ecuatoriana de Empresas de Seguros. (2020). El sector asegurador ecuatoriano en cifras año 2019. Ecuador.

Vafeiadis, T., Diamantaras, K., Sarigiannidis, G., & Chatzisavvas, K. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 1-9.

El Universo. (2018, abril 3). Cultura sobre seguros, con una baja penetración en Ecuador. *El Universo*.

Federación EcuatorIana de Empresas de Seguros. (2020). *Anuario FEDESEG 2019.* Retrieved from FEDESEG: https://6aab8a7f-de25-4e01-bf7a-2697d046daa5.filesusr.com/ugd/f39f07_5723b917ced642ffb8007b78ce56d6a7.pdf

Alpaydin, E. (2014). *Introduction to Machine Learning.* The MIT Press.

Cerda, J., & Cifuentes, L. (2011). Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos. *Revista chilena de infectología*, 138-141.

Gaur, A., & Dubey , R. (2018). Predicting customer churn prediction in telecom sector using various machine learning techniques. *The Institute of Electrical and Electronics Engineers, Inc.(IEEE)*, 1-5.

Stock, J. H., & Watson, M. W. (2012). *Introducción a la econometría.* Madrid: Pearson Eduación S.A.

Cutler, A., Cutler, R., & John, S. (2012). Random Forests. In *Ensemble Machine Learning: Methods and Applications* (pp. 157-176). Springer, Boston, MA.

Malik, S. (2020). *XGBoost: A Deep Dive into Boosting (Introduction Documentation).*

Revista Vistazo. (2020, julio 12). La solidez del sector asegurador a prueba por el COVID-19.

Navarro, F., & Wahren, P. (2017, septiembre 27). *El sector asegurador en América Latina* . Retrieved from CELAG: https://www.celag.org/el-sector-asegurador-en-america-latina/

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning from theory to algorithms.* New York: Cambridge.

ANEXOS

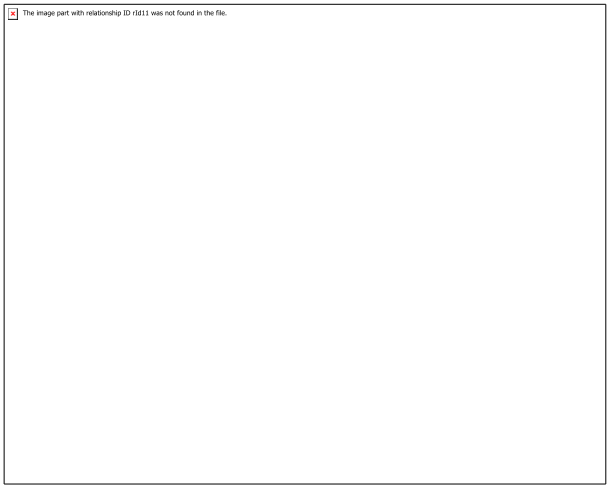


Ilustración 1 Histograma. Antigüedad de clientes

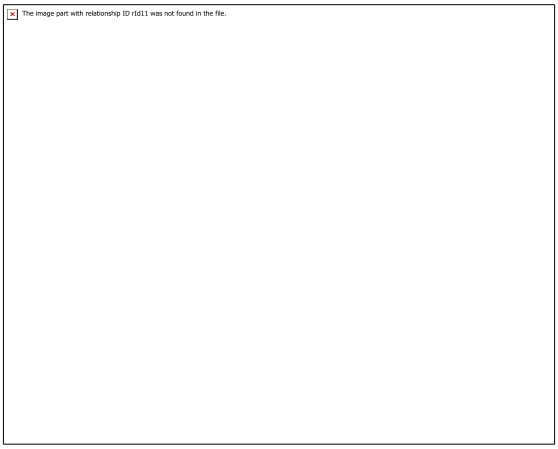


Ilustración 2 Histograma. Edad de Asegurados

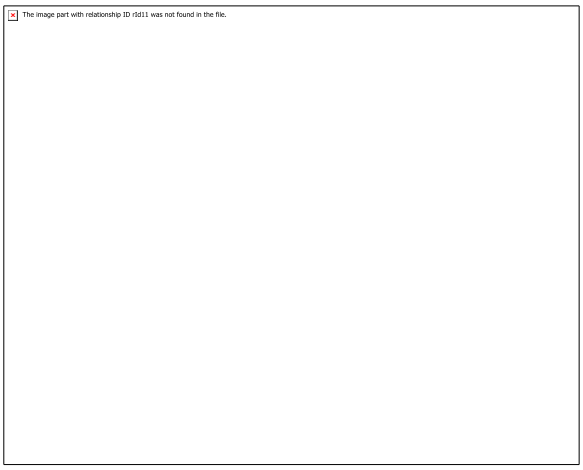


Ilustración 3 Histograma. Número de pólizas contratadas

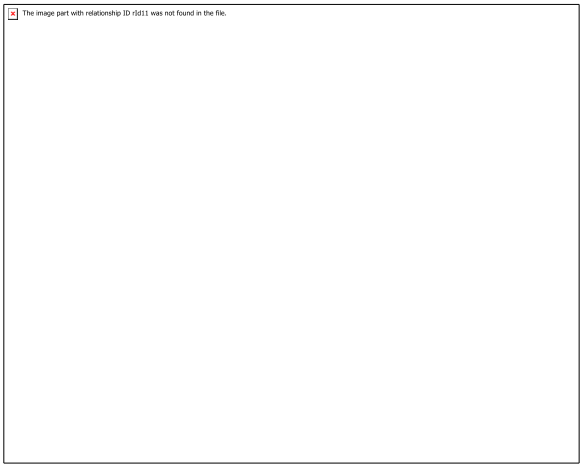


Ilustración 4 Histograma. Antigüedad de Vehículos

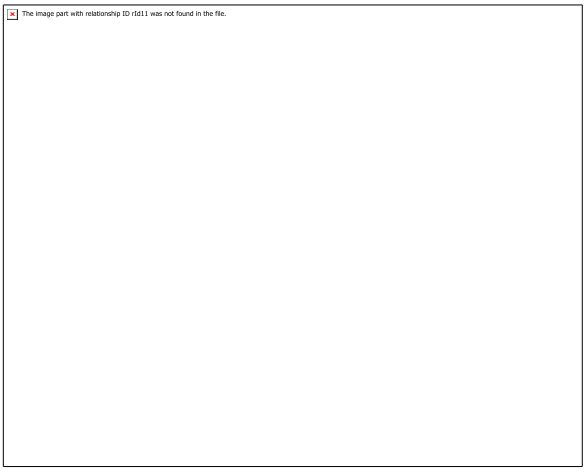


Ilustración 5 Diagrama de Caja. Prima por producto

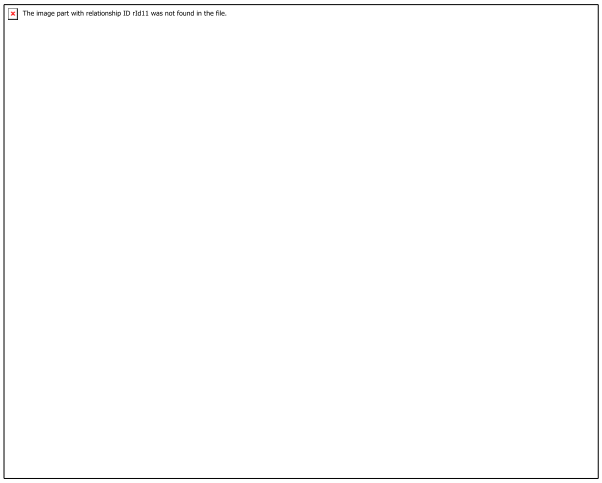


Ilustración 6 Diagrama de Caja. Siniestros por Producto

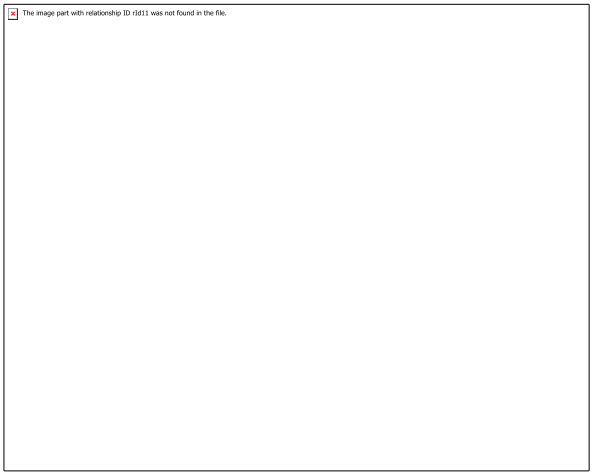


Ilustración 7 Diagrama de Caja. Siniestros por tipo de Vehículo

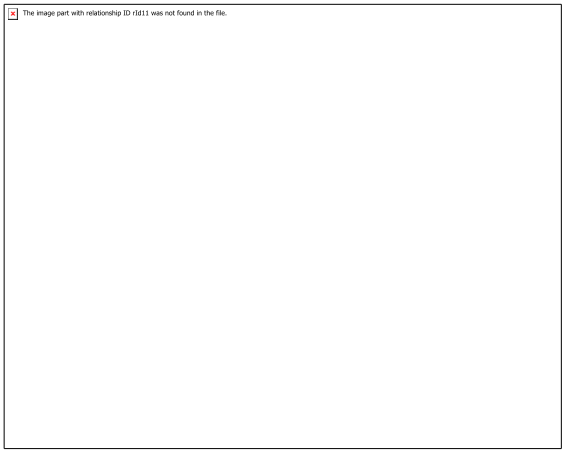


Ilustración 8 Diagrama de Caja. Prima por tipo de vehículo

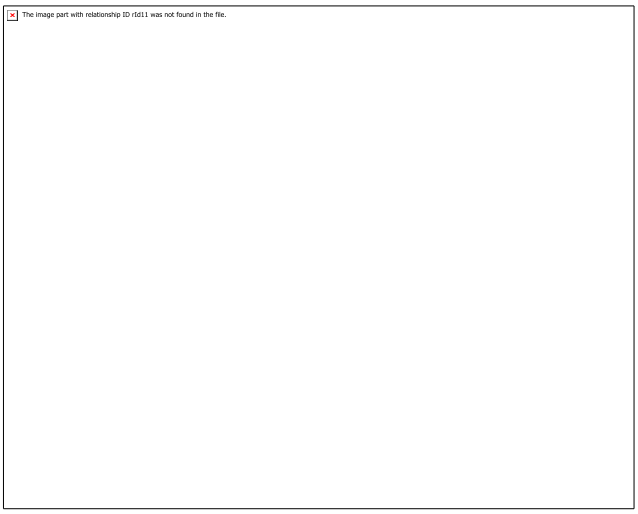


Ilustración 9 Evaluación de hiperparámetros Random Forest

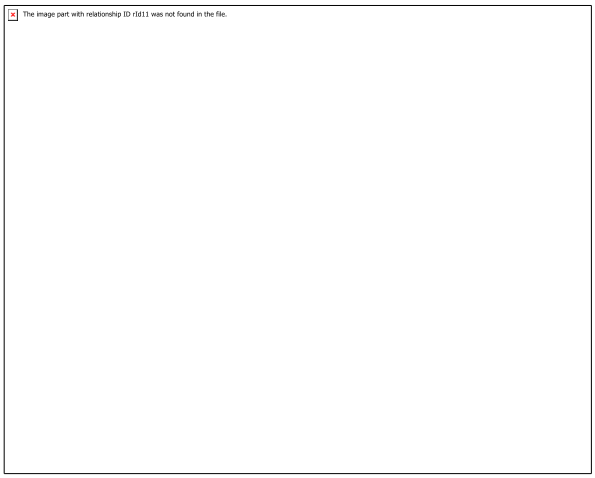


Ilustración 10 Evaluación de hiperparámetros XGB

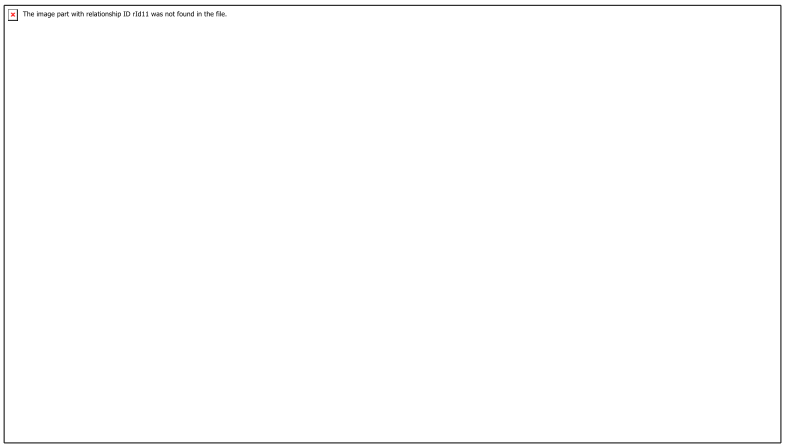


Ilustración 11 Árbol de decisión