Estimación de la probabilidad de cancelación de una póliza individual de automóviles, usando modelos de Machine Learning.

Estimation of the probability of cancellation of an individual car policy, using Machine Learning models.

Juan Carlos Vega Rueda[1]

RESUMEN

El propósito de este artículo es identificar las características de mayor impacto, que aumentan la probabilidad de que un usuario cancele una póliza de seguro de automóvil. Para esto, también es necesario entender el impacto que existe por la cancelación de pólizas (riesgo de caída) y comprender el proceso de tarificación de del costo de pérdida de un seguro. El estudio se enfoca en aquellos riesgos individuales, con un histórico de pólizas emitidas desde enero de 2016 hasta diciembre 2020.

La información de la base de datos esta segmentada de acuerdo con A) la información de la póliza: prima, suma asegurada para algunas coberturas, tiempos de vigencia. B) Información acerca del vehículo: Marca, año modelo, tipo de vehículo, color, servicio. y por último C) información acerca del asegurado: tipo de persona, genero, edad y niveles de ingresos o egresos anuales. Con las variables anteriores, se pretende modelar la probabilidad de cancelación usando algunos modelos de *Machine Learning* como *Random Forest* y *XG Boost*. y se usan las técnicas de los valores SHAP para visualizar y explicar el impacto de las características más importantes en los modelos utilizados.

Los resultados obtenidos muestran que el modelo *Random Forest* tiene mejores resultados, al clasificar clientes que realizan cancelaciones con una precisión del 79%, y se identifican las cinco características más importantes que explican el modelo.

Palabras clave: Riesgo de Caída, Solvencia II, Tarifación, Prima Pura, Machine Learning, Inteligencia Artificial, Random Forest, XG Boost, Valores SHAP.

ABSTRACT

The purpose of this paper is to identify the characteristics of the greatest impact, which used the probability that a user will cancel an auto insurance policy. For this, it is also necessary the impact that exists due to cancellation of policies (risk of falling) and understanding the process of pricing the cost of loss of insurance. The study focuses on those individual risks, with a history of policies issued from January 2016 to December 2020.

[1] Fundación Universitaria Los Libertadores, jcvegar@libertadores.edu.co

The information in the database is segmented according to A) the information of the policy: premium, sum insured for some coverages, validity periods. B) Information about the vehicle: Make, model year, type of vehicle, color, service. and finally, C) information about the insured: type of person, gender, age and levels of income or annual expenses. With the previous variables, it is intended to model the probability of cancellation using some Machine Learning models such as Random Forest and XG Boost. and the techniques of SHAP values are used to visualize and explain the impact of the most important characteristics in the models used.

The results obtained show that the Random Forest model has better results, when classifying clients who make cancellations with an accuracy of 79%, and the five most important characteristics that explain the model are identified.

Key words: Lapse Risk, Solvency II, Pricing, Pure Premium, Machine Learning, Artificial Intelligence, Random Forest, XG Boost, SHAP-Values.

INTRODUCCIÓN

El presente artículo describe de forma general la importancia del riesgo de caída de cartera en una compañía de seguros, especialmente para el ramo de automóviles, y del como este riesgo podría ser considerado en la toma de decisiones y en el cálculo de las estimaciones de las provisiones técnicas y su posible afectación en el proceso de tarificación de la prima pura de riesgo. Para esto, se pretende con este trabajo lograr identificar las variables que ayudan a explicar que un asegurado cancele o no una póliza de seguro individual de automóviles, con el fin de aplicar un modelo de machine learning en específico el modelo de bosques aleatorios o *Random Forest,* y por último clasificar las probabilidades de cancelación de una póliza mediante algunas características del asegurado, del seguro y del riesgo, implicados en la suscripción de la póliza.

La base de datos utilizada registra las emisiones o expediciones de pólizas desde enero 2016 a diciembre 2020, las cuales suman cinco años de información relacionada con seguros individuales de vehículos y de uso particular.

Esta base contiene 121.090 registros, de los cuales 55.476 de ellos han presentado cancelaciones del seguro. La base previamente ha sido tratada con el fin de obtener la información necesaria y detallada para el respectivo análisis, muchos de los nombres de los campos han sido modificados y los campos relacionados con información personal del asegurado han sido anonimizados y no se utilizaron para este estudio, con el fin de proteger la confidencialidad del asegurado.

A continuación, se resumen aspectos generales de solvencia II, riesgo de caída de cartera y del proceso de tarificación, con el fin de dar contexto al problema del presente documento.

SOLVENCIA II Y RIESGO DE CAIDA DE CARTERA

El riesgo de caída de cartera de seguros está relacionado con el impacto que existe o esperado por la cancelación de pólizas o la no renovación de estas, de acuerdo con Millán y Colomina (2001), citado por De Lourdes Gutiérrez Cordero, Segovia-Vargas, & Escamilla (2017), menciona que este riesgo de caída es el conjunto de pólizas que no optan por la renovación a su vencimiento por parte de los asegurados.

Es necesario identificar y poder hacer el control adecuado de este riesgo, pues de acuerdo con las normativas de Solvencia II es fundamental que las compañías aseguradoras prevean los impactos financieros que pueden ocasionar las cancelaciones de las pólizas y así constituir las provisiones o reservas necesarias. Las normas dadas por Solvencia II, están compuestas por una estructura formada por tres pilares, de acuerdo con UNESPA (2015), estos pilares contemplan lo siguiente: Pilar I, basado en el cálculo técnico de la carga de capital, hace referencia a los requerimientos cuantitativos. El Pilar II, representa los requerimientos cualitativos necesarios para el desarrollo de una adecuada supervisión de los riesgos. Y por último el Pilar III que considera la transparencia en la divulgación de la información tanto al sector público y los reportes necesarios al supervisor.

Solvencia II da como herramienta importante la composición del SCR, que consiste en el cálculo de capital de solvencia o Solvency Capital Requirement, que se define como el capital necesario para hacer frente a las posibles pérdidas económicas teniendo en cuenta todos los riesgos cuantificables a los que está expuesta, en un horizonte temporal de un año y con un nivel de confianza del 99.5% (VaR al 99.5%) De Lourdes Gutiérrez Cordero et al (2017).

Por lo anterior, es necesario identificar las posibles pérdidas económicas y esto es los riesgos a los que la compañía se expone, algunos de estos riesgos son los riesgos de mercado (de interés, de acciones, inmobiliario, de cambio, spread y concentración), riesgos de contraparte, riesgo de activos y riesgos de suscripción, en este último se encuentra el riesgo de caída de cartera.

Para el cálculo del SCR, solvencia II propone dos caminos para hacerlo, uno es el modelo estándar y el otro es la utilización de modelos internos, los cuales se basan en la experiencia propia de cada compañía, utilizando sus supuestos y el comportamiento histórico de cada uno de estos riesgos Ayuso (2012). La cuantificación del riesgo se lleva a cabo con la utilización

de métodos estadísticos debidamente validados, proporcionando el rigor técnico-actuarial sobre el que se fundamentan dichos modelos.

TARIFICACIÓN EN SEGUROS

La identificación de la caída de una póliza, tanto en cancelación o no renovación de esta es una medida que permitiría ajustar el sistema de tarificación y el cálculo de la respectiva tasa pura de riesgo, de acuerdo con Calzón, C. B. et al. (2008) un sistema de tarificación es el conjunto de principios técnicos que ayudan a elaborar una tarifa, el objetivo es determinar el valor de la prima y que esta sea equitativa de acuerdo con el nivel de riesgo expuesto. Uno de estos principios técnicos es la exposición del riesgo, en este caso del vehículo asegurado, que se caracteriza por ser la unidad básica de riesgo que subyace al seguro.

Ahora bien, la medida de exposición utilizada para fines de elaboración de tarifas varía considerablemente según la línea de negocio o el tipo de riesgo. Por ejemplo, un auto deportivo rojo conducido por una persona joven durante un año representa una exposición distinta a la que tendría una persona mayor que conduce un auto familiar. Con el análisis apropiado de la exposición se obtiene el valor de la prima que es la suma que el asegurado paga por el seguro mientras dure la cobertura de este.

El riesgo para las compañías de seguros en aquellas pólizas canceladas consiste en que parte de esta prima se devuelva al no devengarse y que a su vez se haya manifestado una reclamación. Esto ocasionaría desviaciones en las reservas hechas, ya que algunas de estas se estiman partiendo del supuesto de que la cartera original de pólizas al inicio de una cobertura se mantenga constante al final de esta, pero cuando no es así se generan algunas desviaciones no programadas en la siniestralidad.

Para calcular la prima pura (costo de pérdida), la cual se define por Hadidi (2015) como la medida de la pérdida esperada por exposición, se deben considerar dos ratios importantes la frecuencia y la severidad.

Sin embargo, estas medidas de frecuencia y severidad para la estimación de la prima pura de algunos productos se complementan con la utilización de métodos probabilísticos (muchos basados en modelos lineales generalizados GLM) o técnicas de inteligencia artificial que permitan identificar y comprender todos aquellos factores que impactan en la materialización de un riesgo. Por esto, al fijar el precio de un nuevo producto de seguro, el actuario debe buscar información interna que pueda tener alguna relación con el nuevo producto o adquirir datos externos relevantes o riesgos adicionales (caída de cartera) que permitan la estimación más apropiada de la prima.

METODOLOGIA

Para el desarrollo de este proyecto se abordarán las técnicas utilizadas en la inteligencia artificial, específicamente las utilizadas por el *Machine Learning*, ya que esta proveerá el aprendizaje necesario de los parámetros que se utilizarán para la estimación de la probabilidad de cancelación, y así determinar la respectiva clasificación del perfil del asegurado, que podría de acuerdo con su probabilidad realizar la cancelación de la póliza de seguros.

Los modelos que se utilizan para resolver el objetivo del presente proyecto son los determinados por las técnicas de Bosques Aleatorios (*Random Forest*) y el algoritmo *Xg-Boost*, el cual consiste en un ensamblado secuencial de árboles de decisión.

Breiman (2001) define el *Random Forest* como un clasificador que consta de una colección de árboles de clasificación estructurados $\{h\,(x, \Theta_k), k\, =\, 1, \dots\}$ donde las $\{\Theta_k\}$ son vectores aleatorios independientes distribuidos de forma idéntica y cada árbol emite un voto unitario para la clase más popular para la entrada x. Es decir, el modelo toma muestras aleatorias del conjunto de datos de entrenamiento, realiza para cada una de las muestras un árbol y luego pondera los resultados.

Por otro lado, Espinoza (2020) menciona que el algoritmo XG Boost (Extreme Gradiente Boosting) es una técnica de aprendizaje supervisado, que se basa en arboles de decisión. Algunas de las características mencionadas por Chen y Guesterin (2016) en su trabajo *Xgboost: A scalable tree boosting system* son las siguientes:

- Los árboles de decisión utilizados en el algoritmo XG Boost, se agregan secuencialmente con el objetivo de aprender de los resultados de los arboles previos corrigiendo el error producido por cada uno de ellos, hasta que el error no pueda ser corregido, a este proceso se le conoce como gradiente descendente.
- El algoritmo XG Boost puede manejar bases de datos grandes y con múltiples variables, maneja valores perdidos y tiene buena velocidad en su ejecución.
- Sin embargo, el algoritmo posee algunas desventajas solo trabaja con vectores numéricos, los parámetros deben ser ajustados previamente y esto puede ocasionar el consumo de muchos recursos computacionales, Espinoza (2020).

ANÁLISIS EXPLORATORIO Y DESCRIPTIVO.

La base se compone de 16 variables predictoras y una variable objetivo que indica si la póliza fue o no cancelada. En la **Tabla 1** se describen las variables. Además, para la efectiva protección de datos, se han creado variables por niveles o categorías para los campos de ingresos y egresos, en la tabla 2 se muestra la distribución de los niveles de ingresos de acuerdo con los registros y número de cancelaciones.

Tabla 1: Estructura de la base de cancelaciones.

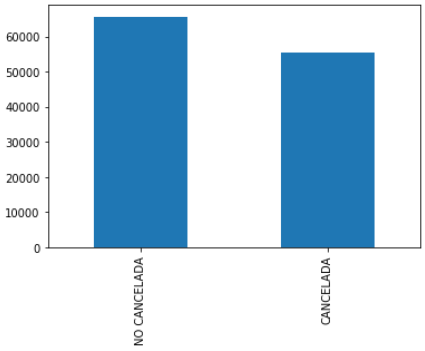
No.	Variable	Descripción
1	MES_EMISION	Mes en el que se realiza la expedición de la póliza.
2	AAAA_MODELO	Año del modelo del vehículo.
3	MARCA	Marca del vehículo, agrupada en 18 marcas principales.
4	COLOR	Color del vehículo, agrupada en 11 colores principales.
5	CILINDRAJE	Cilindraje del vehículo.
6	GENERO	Género del asegurado (masculino, femenino).
7	ESTADO_CIVIL	Estado civil del asegurado (Casado, Union libre, Divorciado, Soltero, Viudo, No especificado).
8	EDAD	Edad del asegurado.
9	CLASIFICACION_EDAD	Grupo etario al que pertenece el asegurado de acuerdo con la edad.
10	SINIESTRADA	Indica si la póliza fue siniestrada o no.
11	#SINIESTROS	Indica el número de siniestros, en el caso de ser siniestrada.
12	SA_RCE	Suma asegurada por la cobertura de responsabilidad civil.
13	SA_PMDH	Suma asegurada por perdidas menores relacionadas con daños o hurto.
14	SA_PSDH	Suma asegurada por perdidas severas relacionadas con daños o hurto.
15	PRIMA	Prima pagada por el seguro del vehículo.
16	NIVEL_INGRESOS	Nivel de ingresos del asegurado.
17	NIVEL_EGRESOS	Nivel de egresos del asegurado.
18	COD_CANCELADA	Identificador de cancelación de la póliza (1=cancelada, 0=No cancelada)

Tabla 2: Distribución de registros y cancelaciones por nivel de ingresos.

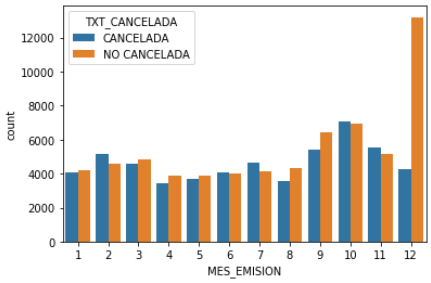
Nivel de ingresos	Registros	Cancelaciones
1	10.371	5.171
2	479	150
3	3.461	1.220
4	1.826	673
5	8.098	2.702
6	11.918	3.972
7	9.765	3.102
8	11.109	3.623
9	4.845	1.573
10	2.076	658
11	2.198	685
12	631	224
13	16.986	6.169
14	37.327	25.554
	121.090	55.476

A continuación, se mostrarán algunas descripciones de las variables categóricas y su impacto para el modelo a trabajar.

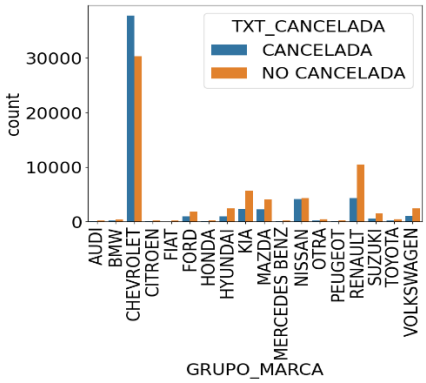
Gráfica 1: Distribución variable objetivo



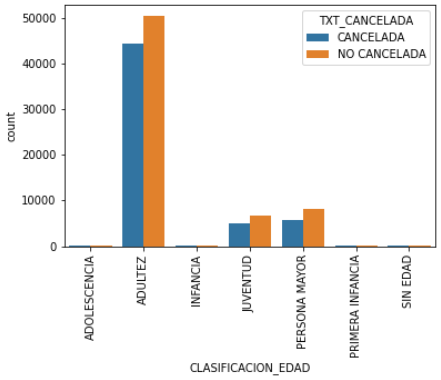
Gráfica 2: Distribución de cancelaciones por mes de emisión.



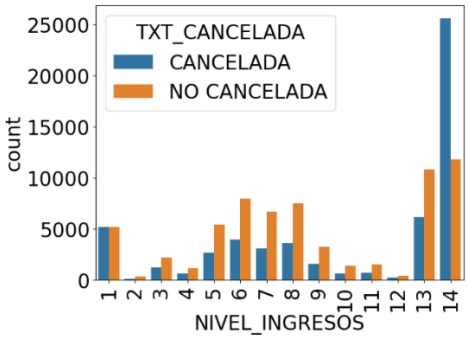
Gráfica 3: Distribución de cancelaciones por categoría de marca.



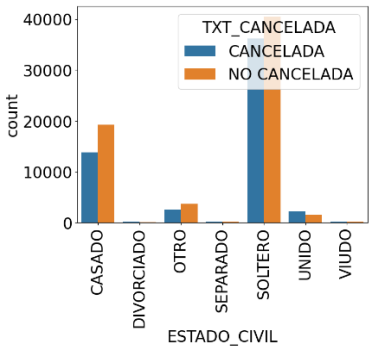
Gráfica 4: Distribución cancelaciones por grupo etario.



Gráfica 5: Distribución de las cancelaciones por nivel de ingresos.



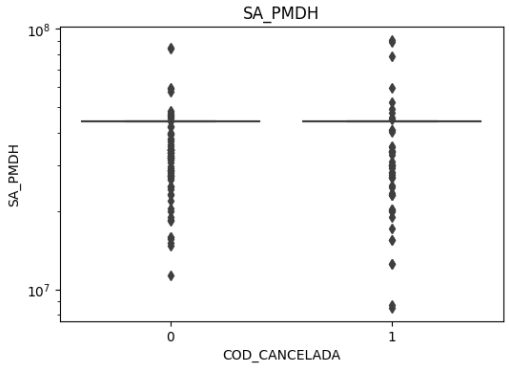
Gráfica 6: Distribución de cancelaciones por estado civil.



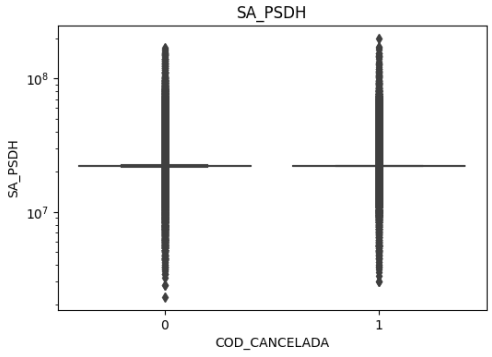
La **Gráfica 1**. Muestra la distribución de la variable objetivo con respecto al espacio muestral, se puede evidenciar que no existe un desbalanceo grande de variable la respuesta, ya que las cancelaciones representan el 45.81% del total de los registros.

En la **Gráfica 2**. Se observa que, dependiendo del mes de emisión, puede ocurrir mayores cancelaciones, por ejemplo, se aprecia que al parecer se emiten más pólizas en el mes de diciembre con respecto a los demás meses y esto podría ocasionar un mayor número de cancelaciones. Por otro lado, **la Gráfica 3**. Evidencia que las marcas con más cancelaciones son Chevrolet y Renault. Las demás gráficas muestran el comportamiento de las cancelaciones en función de las variables categóricas: Estado civil, nivel de ingresos y grupo etario.

Gráfica 7: Comportamiento Suma Asegurada por pérdida menor daños y hurtos.



Gráfica 8: Comportamiento Suma Asegurada por pérdida severa daños y hurtos.



Las **gráficas 7 y 8**, muestran el comportamiento de la suma asegurada por pérdidas menores relacionadas con daños o hurtos y la suma asegurada por pérdida severa daños o hurtos, se evidencia que las distribuciones son similares para aquellas pólizas canceladas, como para las que no han sido canceladas, sin embargo, para el tratamiento del modelo, estas variables numéricas incluyendo la variable PRIMA serán normalizadas.

SELECCIÓN Y AJUSTE DE VARIABLES.

Para realizar el modelamiento se deberá realizar un preprocesamiento de las variables, pero previamente es necesario seleccionar las variables que más impacten a la variable objetivo. Para esto se utilizará una medida de asociación entre dos variables (la respuesta con cada variable explicativa), esta medida estará dada por el coeficiente de Cramer. De acuerdo con Aguilar (2017) el coeficiente de Cramer es una medida simétrica que se utiliza para medir la intensidad que puede haber entre dos o más variables según la escala nominal. También Aguilar (2017) afirma que el coeficiente de V de Cramer se lo escoge cuando los valores son medidas independientes, según el tamaño de la muestra de los cuales da valores simétricos, y la interrelacionada entre dos variables de los que se pueden tomar 2 valores posibles.

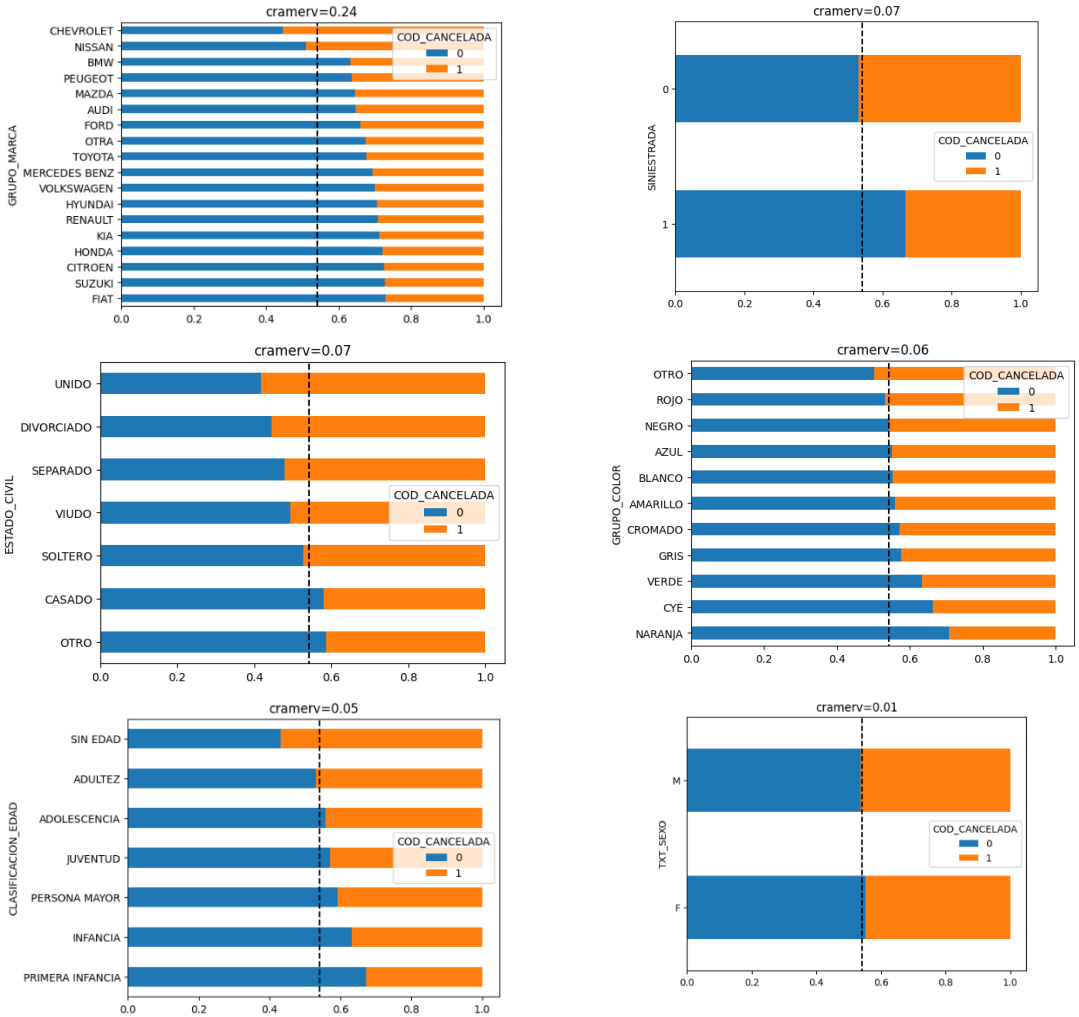
Al aplicar la medida del coeficiente Cramer sobre las variables categóricas (Marca, Siniestrada, Estado civil, Grupo color, grupo etario y género) se obtiene que las principales variables con más contribución son:

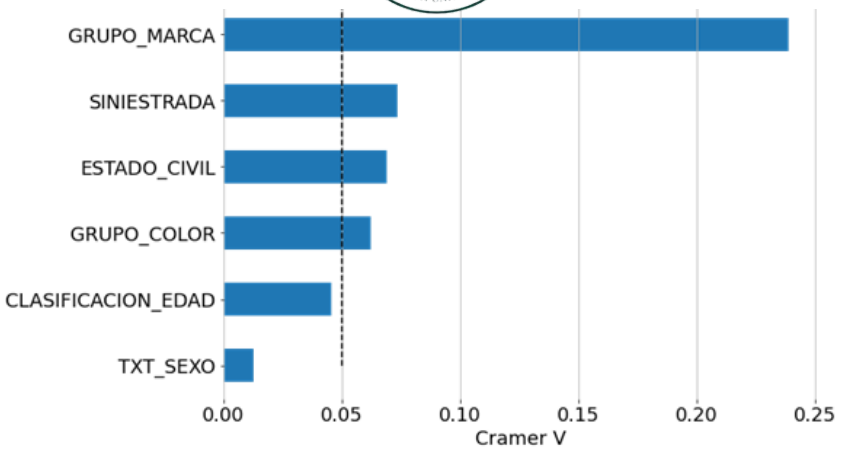
Tabla 3: Top variables categóricas según el coeficiente de Cramer.

Variable	Coeficiente
Grupo_Marca	0,2389
Siniestrada	0,0735
Estado_Civil	0,0689
Grupo_Color	0,0622

La **gráfica 9** muestra el resumen gráfico de las asociaciones de las variables categóricas con respecto a las cancelaciones.

Gráfica 9: Intensidad o asociación de las variables categóricas con respecto a las cancelaciones.



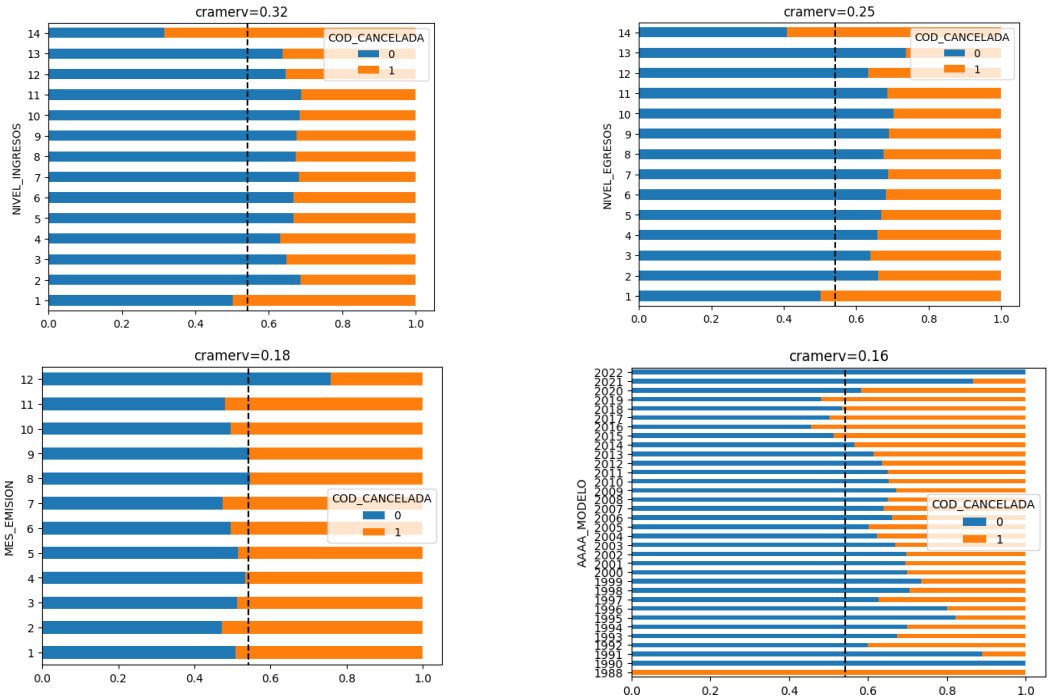


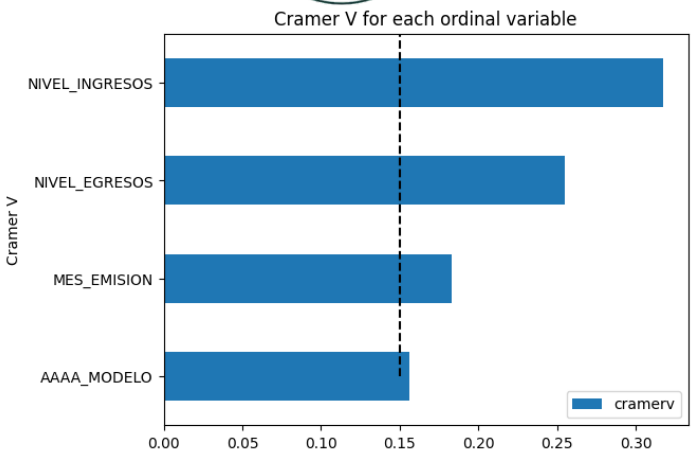
Para las variables ordinales se tiene las siguientes asociaciones:

Tabla 4: Top variables ordinales según el coeficiente de Cramer.

Variable	Coeficiente
Nivel_Ingresos	0,3173
Nivel_Egresos	0,2549
Mes_emisión	0,1833
AAAA_Modelo	0,1562

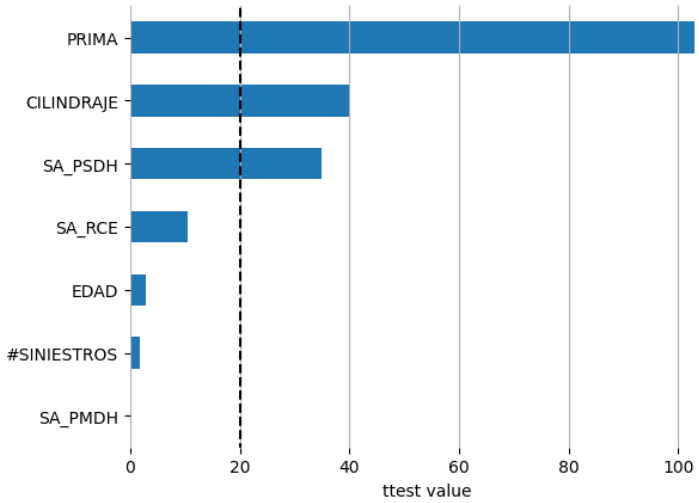
Gráfica 10: Intensidad o asociación de las variables ordinales con respecto a las cancelaciones.





Para el caso de las variables numéricas, al hacer una prueba **t** sobre las asociaciones que tienen sobre la variable objetivo se obtiene lo siguiente:

Gráfica 11: Intensidad o asociación de las variables numéricas con respecto a las cancelaciones.



AJUSTE DE CARACTERÍSITICAS Y ENTRENAMIENTO.

Para el aprendizaje adecuado de los modelos a emplear, se usarán las técnicas *"Feature Engineering"* con el fin de que las variables categorías sean de tipo *dummie* por medio del ajuste OneHot y las variables numéricas se normalizarán para que estas sean equiparables entre sí. De esta manera, el nuevo grupo de variables con los ajustes hechos sería el siguiente:

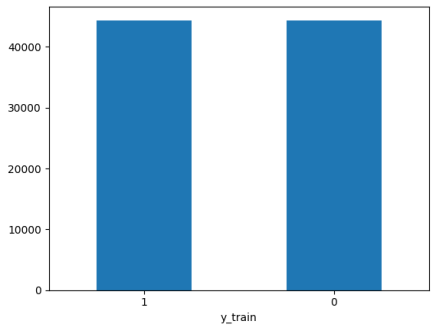
Tabla 5: Grupo de variables ajustadas.

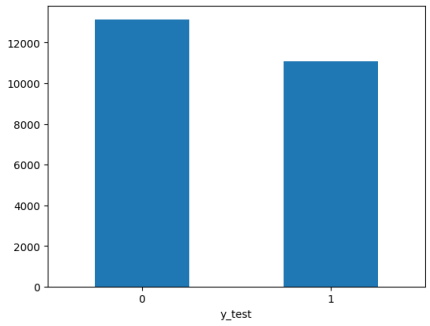
#	Column	Non-Null Count	Dtype
0	AUDI	121090 non-null	float64
1	BMW	121090 non-null	float64
2	CHEVROLET	121090 non-null	float64
3	CITROEN	121090 non-null	float64
4	FIAT	121090 non-null	float64
5	FORD	121090 non-null	float64
6	HONDA	121090 non-null	float64
7	HYUNDAI	121090 non-null	float64
8	KIA	121090 non-null	float64
9	MAZDA	121090 non-null	float64
10	MERCEDES BENZ	121090 non-null	float64
11	NISSAN	121090 non-null	float64
12	OTRA_MARCA	121090 non-null	float64
13	PEUGEOT	121090 non-null	float64
14	RENAULT	121090 non-null	float64
15	SUZUKI	121090 non-null	float64
16	TOYOTA	121090 non-null	float64
17	VOLKSWAGEN	121090 non-null	float64
18	AMARILLO	121090 non-null	float64
19	AZUL	121090 non-null	float64
20	BLANCO	121090 non-null	float64
21	CROMADO	121090 non-null	float64
22	CYÉ	121090 non-null	float64
23	GRIS	121090 non-null	float64
24	NARANJA	121090 non-null	float64
25	NEGRO	121090 non-null	float64
26	OTRO_COLOR	121090 non-null	float64
27	ROJO	121090 non-null	float64
28	VERDE	121090 non-null	float64
29	CASADO	121090 non-null	float64
30	DIVORCIADO	121090 non-null	float64
31	OTRO_EST_CIVIL	121090 non-null	float64
32	SEPARADO	121090 non-null	float64
33	SOLTERO	121090 non-null	float64
34	UNIDO	121090 non-null	float64
35	VIUDO	121090 non-null	float64
36	NO SINIESTRADA	121090 non-null	float64
37	SINIESTRADA	121090 non-null	float64
38	PRIMA	121090 non-null	float64
39	CILINDRAJE	121090 non-null	float64
40	SA_PSDH	121090 non-null	float64
41	EDAD	121090 non-null	int64
42	MES_EMISION	121090 non-null	int64
43	AAAA_MODELO	121090 non-null	int64
44	NIVEL_INGRESOS	121090 non-null	int64
45	NIVEL_EGRESOS	121090 non-null	int64
46	CANCELADA	121090 non-null	int64

A partir del ajuste de variables realizado, se divide la base en dos partes, entrenamiento (80%) y test (20%), la cual se usará para probar y validar los modelos. Esta división de los datos se hace usando la librería *Sklearn* de *Python.*

Se hace el balanceo respectivo de las variables y se valida el respectivo desbalanceo, el cual es de tan solo 15%. Con estos datos se entrenará el modelo Random Forest y el modelo XGBoost.

Gráfica 12: Comparativo (balanceo) entrenamiento y test de la variable objetivo.





ANÁLISIS DE RESULTADOS

Se aplicaron los respectivos modelos *Random Forest* y *XG Boost* a la base de entrenamiento, y se validaron con base a los datos de prueba, de esta manera se obtuvieron los resultados observados en la **Tabla 6**.

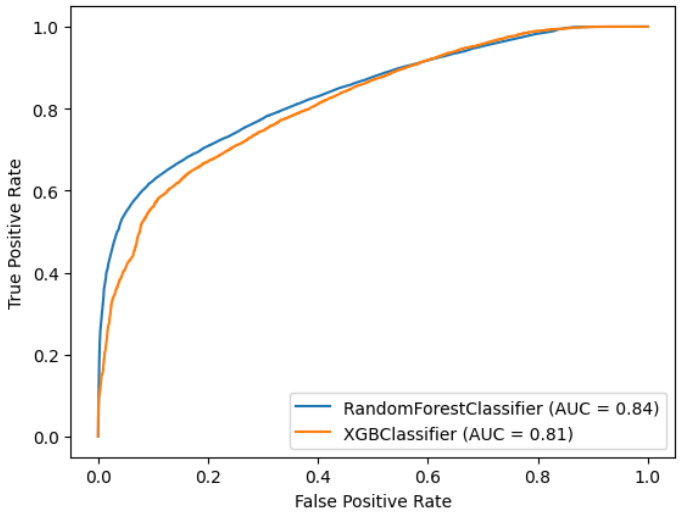
El modelo *Random Forest* presenta una exactitud (*Accuracy)* del 77%, mientras que el *Accuracy* del modelo *XG Boost* es de 75%. Al realizar la comparación entre las distintas métricas se observa que el primer modelo es más optimo que el modelo *XG Boost*.

Tabla 6 Métricas desempeño Random Forest VS XG Boost

	Random Forest		XG Boost	
	No cancelada	Cancelada	No cancelada	Cancelada
Precisión	75%	79%	74%	76%
Sensibilidad	85%	67%	83%	65%
F1 Score	80%	73%	78%	70%
Accuracy	77%		75%	

Cuando se comparan los modelos usando las curvas ROC (*Receiver Operating Characteristic*) y la métrica AUC (Área bajo la curva), métricas que se utilizan para conocer el rendimiento global de las pruebas realizadas, se observa que el modelo de mayor rendimiento con un AUC del 84% ha sido el modelo *Random Forest*.

Gráfica 13 Curvas ROC y métrica AUC

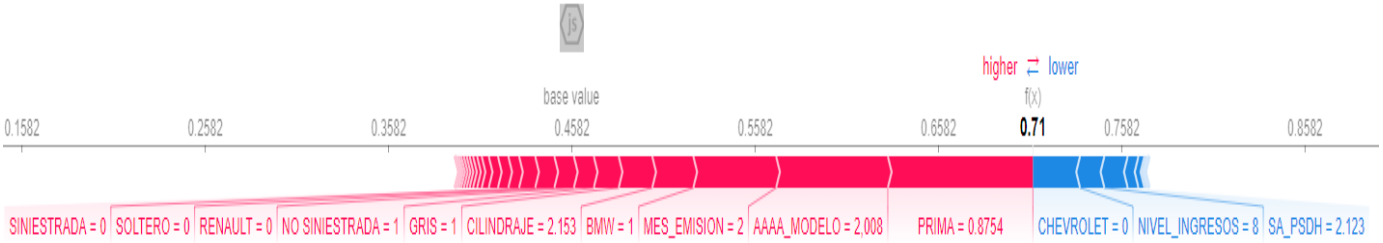


Por esta razón, el análisis de los resultados se enfocará en el modelo *Random Forest.*

Por otro lado, al usar los valores SHAP (*shap values*[2]), se podrá visualizar y analizar las características más influyentes que afectan la decisión de la cancelación de una póliza. Los valores SHAP también pueden ser utilizados para observar el comportamiento individual de cada asegurado, esto podría utilizarse para generar alertas tempranas por parte de la compañía. Por ejemplo, para un cliente cuya emisión de póliza haya sido en febrero y una prima estandarizada por encima de la media, se observaría un comportamiento como el siguiente:

[2] En el artículo: "Explique su modelo con los valores SHAP" disponible en https://ichi.pro/es/explique-su-modelo-con-los-valores-shap-206942979284541, Menciona que Lundberg y Lee (2016), propusieron el valor SHAP como un enfoque conjunto para explicar el resultado de cualquier modelo de aprendizaje automático.

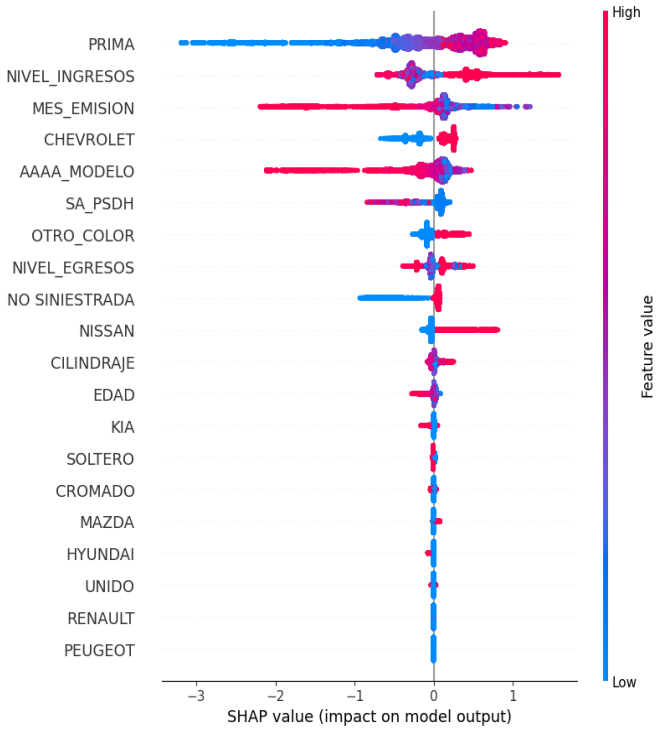
Gráfica 14: Shap-Values para una observación.

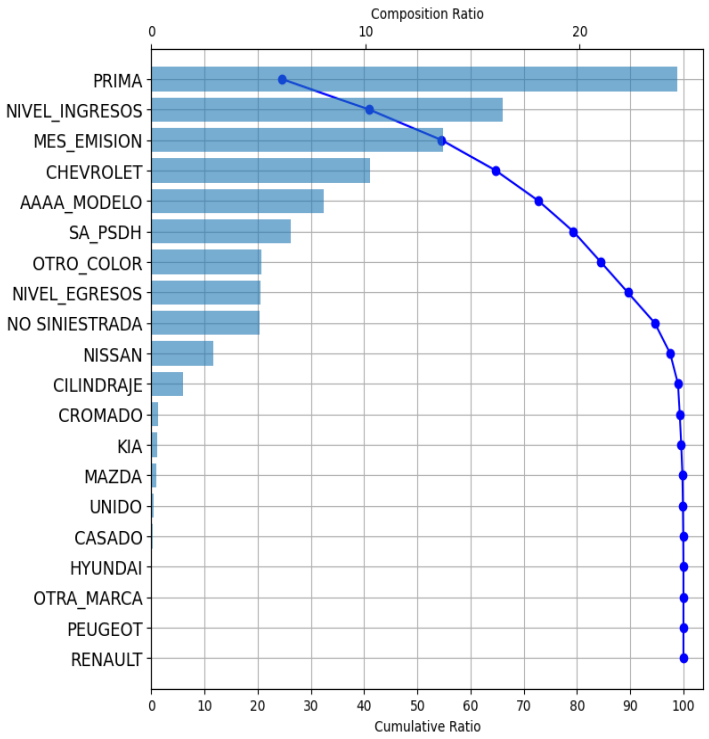


La anterior gráfica permite visualizar que, para la observación indicada, las variables que "empujan" la predicción hacia la derecha (zona roja) con mayor contribución positiva son la prima, el año del modelo y el mes de emisión. Para este ejemplo, la predicción hecha es de 0.71, lo que indica una mayor posibilidad de cancelación sí el año del modelo es 2008, una prima estandarizada de 0.8754 y cuyo mes de emisión haya sido en febrero. En cambio, si los ingresos están en el nivel ocho, esta característica lleva la predicción a la izquierda.

De manera general, como se observa en la **gráfica 14**, se identifica que la probabilidad de cancelación está mayormente impactada por las variables: Prima, nivel de ingresos, mes de emisión, marca y año del vehículo. El gráfico de cascada de la derecha muestra que la variable más importante del modelo es la prima, la cual explica más del 20% del modelo, las primeras cinco características explican por lo menos el 70% del modelo.

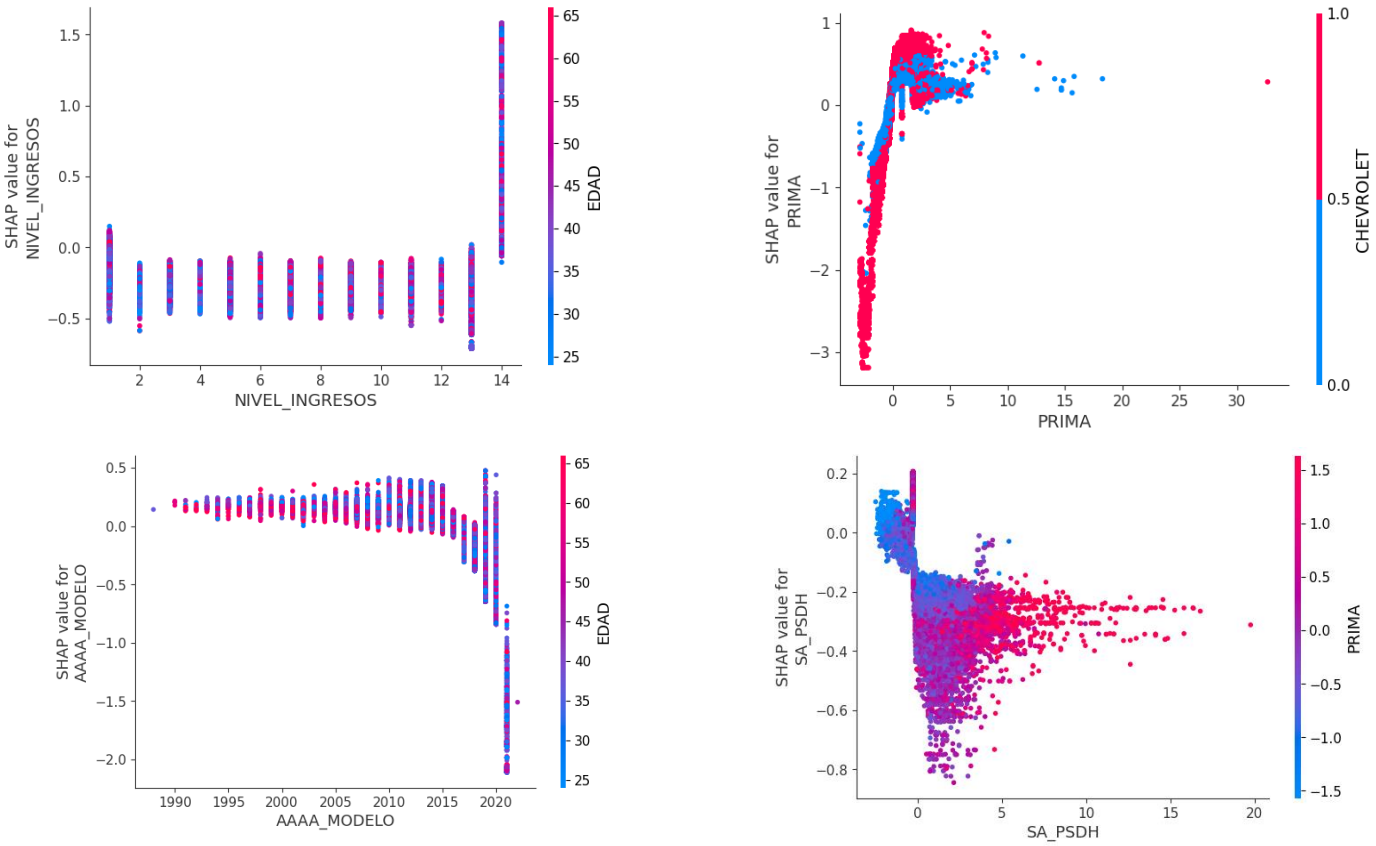
Gráfica 15 Impacto de las variables en el modelo.





A continuación, se muestran algunas relaciones de dependencia parcial[3] para algunas de las variables más significativas.

Gráfica 16 Relaciones de dependencia parcial SHAP.



DISCUSIÓN DE RESULTADOS.

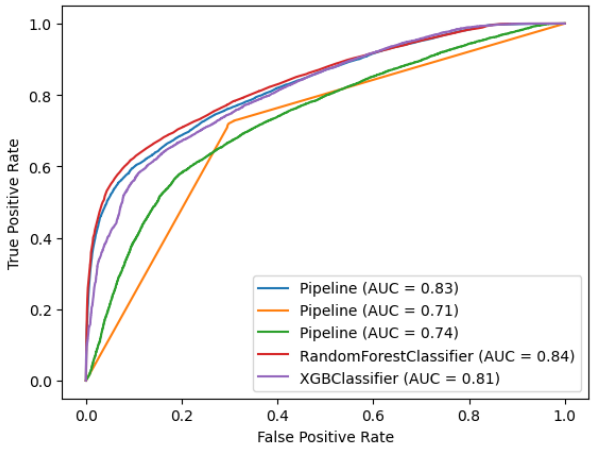
A partir de los resultados anteriores, se observa que el modelo *Random Forest* fue más optimo frente a otros modelos evaluados con las misma base de entrenamiento, Obteniendo un AUC de 84% frente a modelos como arboles de decisión, XG Boost o regresión logística, como se observa en la **gráfica 17**.

Sin embargo aunque el modelo presenta un accuracy del 77% se puede utilizar como una introducción a la detección de alertas tempranas. Al poner el modelo en practica permitirá realizar ajustes con caracteristicas no incluidas y validar las caracteristicas incluidas inicialmente en el modelo, y de esta manera minimizar el error presentado.

[3] Muestra el efecto marginal que tienen una o dos características en el resultado previsto de un modelo de aprendizaje automático. Indica si la relación entre el objetivo y una característica es lineal, monótona o más compleja. (https://ichi.pro/es/explique-su-modelo-con-los-valores-shap-206942979284541)

El modelo de alertas tempranas deberá compensar el error agrupando aquellos usuarios con una probabilidad alta, de acuerdo con un umbral especifico, para este caso probabilidades dadas por las métricas precisión (79%) o sensibilidad (67%).

Gráfica 17 Curvas ROC y métrica AUC varios modelos.



CONCLUSIONES.

De acuerdo con los resultados obtenidos, se concluye que:

1. usando la base actual del estudio de este documento, se evidencia que el modelo *Random Forest* fue más preciso en un 2% que el modelo *XG Boost*.
2. Al comparar la métrica AUC del modelo *Random Forest*, con las métricas de otros modelos (arboles de decisión, regresión logística, XG Boost), se comprueba que el modelo *Random Forest* fue el mejor, de acuerdo con la base de datos actual utilizada.
3. Por medio de los valores SHAP, se pudo determinar que las características más importantes y que influyen en la posible cancelación de la póliza es la prima, nivel de ingresos y mes de emisión, que explican más del 50% del modelo.
4. Conocer las posibles variables que impactan en la decisión de cancelación de un cliente y su respectiva perfilación, puede ser utilizado en campañas de fidelización y retención para vigencias futuras.
5. A pesar de un *Accuracy* bajo (menor a 80%), se puede usar los resultados del modelo para detectar alertas tempranas, y de esta manera minimizar los riesgos de cancelación al identificar los clientes con mayor probabilidad de acuerdo con sus características.
6. Se recomienda utilizar modelos internos, como buenas prácticas, que permitan el calculo del SCR a partir de la experiencia y comportamiento propio de los productos, así en el cálculo de la provisión por riesgo de caída de cartera, modelos soportados por la data y explicados por IA pueden ser útiles y proveer herramientas que permitan ante la SFC validar los modelos internos versus el modelo estándar y si es posible reducir el impacto económico sobre los resultados financieros de la compañía.

7. Finalmente, también se recomienda ajustar dentro del sistema de tarificación algún nivel de riesgo asociado al riesgo de caída de cartera, debido a que las tarifas deben establecerse de modo que se espere que la prima pueda cubrir todos los costos asociados y lograr el beneficio técnico esperado. Esto, de acuerdo con el principio número dos de *Statement of Principles* Regarding Property and Casualty Insurance Ratemaking de la CAS (The Casualty Actuarial Society) que declara: "Una tarifa cubre todos los costos asociados con la transferencia de riesgo", para lograr el equilibrio se debe considerar que la tarifación es prospectiva y que este se debe lograr a nivel global e individual, es decir para cada riesgo.

REFERENTES BIBLIOGRAFICOS

Aguilar Apolo, W. E. (2017). V de Cramer: patrón de relación entre variables, de acuerdo a la frecuencia de datos.

Berquist, R., Cooper, W. P., Hachemeister, C. A., Hall III, J. A., Richards, H. R., Riddlesworth, W. A., ... & Trudeau, D. E. (1979). Statement of Principles Regarding Property and Casualty Loss and Loss Adjustment Expense Liabilities. Astin Bulletin, 10, 305-317. Disponible en: https://propertycasualtyfocus.com/wp-content/uploads/2015/08/state-principles-regarding-property-casualty-insurance-ratemaking.pdf

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Calzón, C. B., Martínez, A. H., & Rodríguez-Piñero, P. T. (2008). Factores de riesgo y cálculo de primas mediante técnicas de aprendizaje. Fundación MAPFRE, Instituto de Ciencias del Seguro. Disponible en: https://app.mapfre.com/ccm/content/documentos/fundacion/cs-seguro/libros/factores-de-riesgo-y-calculo-de-primas-mediante-tecnicas-de-aprendizaje-122.pdf

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Cordero, M. D. L. G., Segovia-Vargas, M. J., & Escamilla, M. R. (2017). Análisis del riesgo de caída de cartera en seguros: Metodologías de "inteligencia artificial" vs "modelos lineales generalizados". Economía Informa, 407, 56-86.

Gutiérrez, M. A., Estany, M. G., & Marín, A. M. P. (2012). Modelos internos en Solvencia II. Su aplicación al cálculo del coeficiente de caída de cartera. Gerencia de riesgos y seguros, 29(112), 38-48.

Espinosa-Zúñiga, J.J.,(2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. Ingeniería Investigación y Tecnología, 21 (03), 1-16. https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Hadidi Nasser (2015). Rate Making-Key Concepts. University of Wisconsin-Stout. AWB conferencia: Actuaries without borders section des actuaries sans frontiers L'AAI.

Lundberg, S., & Lee, S. I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

Pascual Villacampa, M. (2006). Proceso de tarificación en el Seguro del Automóvil. Una perspectiva técnica. Disponible en: http://www.servidor-gestisqs.com/ub/intranet/PDF/tesis_alumnos/Montse_Pascual.pdf

UNESPA (2015). Solvencia II. De un vistazo. Disponible en: http://unespa-web.s3.amazonaws.com/main-files/uploads/2017/07/Solvencia-II.-De-unvistazo.-FINAL.pdf