Modelo de pronóstico de deserción de los tarjetahabientes al adquirir una tarjeta de crédito y ser cancelada a los 90 días.

Model for forecasting cardholder's attrition when acquiring a credit card and being canceled after 90 days.

Ramón Siddartha Riveros Pulgarín,

Fundación Universitaria Los Libertadores.

rsriverosp@libertadores.edu.co

RESUMEN

Este proyecto se centra en el análisis de una inquietud que se planteó en el área de riesgos y de cliente relacional, con respecto a la probabilidad de deserción que se puede presentar con un cliente o tarjetahabiente al adquirir una tarjeta de crédito y ser cancelada en los próximos 90 días. Con este contexto se determina el análisis y buscar la mejor predicción en la clasificación del modelo y dejar como base los cimientos para en futuros estudios relacionados con la educación financiera en Colombia en los sectores económicos con falta.

Palabras clave: TDC – Tarjeta de Crédito, Tarjetahabiente, Machine Learning, Random Forest, Regresión logística, Validación Cruzada, Curva ROC y AUC

ABSTRACT

This project focuses on the analysis of a concern that arose in the area of risk and relational customer, with respect to the probability of attrition that may occur with a customer or cardholder when acquiring a credit card and being canceled in the next 90 days. With this context, the analysis is determined and the best prediction is sought in the classification of the model and the foundation for future studies related to financial education in Colombia in lacking economic sectors is established.

Keywords: TDC – Credit Cart, Machine Learning, Random Forest, Logistic Regression, Cross Validation, Curve ROC and AUC

INTRODUCCIÓN

Las Tarjetas de Crédito son un medio de pago en donde al cliente se le asigna un cupo de crédito que se empieza a consumir haciendo compras en establecimientos de comercio o bien retirando dinero (avances). Al momento de la compra o del avance, al cliente se consulta a cuantas cuotas desea diferir dicha transacción, pudiendo escoger entre 1 y 36 meses. Cada mes se debe pagar lo correspondiente a intereses y abonos a capital. A medida que va pagando, se empieza a liberar de nuevo el cupo de crédito de tal forma que el cliente pueda seguir utilizando su TDC. Cuando las TDC utilizan todo su cupo de crédito, no es posible seguir transando con ellas por lo que el cliente debe recurrir a solicitar un aumento de cupo o bien abonar capital a la deuda para liberar cupo. Algunas tarjetas de crédito tienen atados

cuotas de manejo (generalmente se pagan trimestralmente) y algunos cargos recurrentes (débitos automáticos). Igualmente, hay algunas tarjetas de crédito que el cliente no mueve (no transa) ya que muchos clientes las mantienen no como medio de pago sino como un medio para hacer frente a alguna eventualidad.

Desde este contexto este trabajo pretende desarrolla un modelo de deserción que prediga la probabilidad que un cliente cancele su TDC en los siguientes tres meses.

REFERENTES TEORICOS

Contexto de las tarjetas de crédito

En el año de 1998, en Latinoamérica el único país que tenía un nivel de tarjetas por habitante activo cercano a la unidad era Argentina. Por su parte, México y Colombia tenían un nivel de alrededor de 0.6, Chile de casi 0.5 y Brasil y Venezuela de 0.34. Puede afirmarse entonces que en lo que respecta a penetración de las tarjetas plásticas en la PEA, Colombia estaba 30% por encima del promedio latinoamericano en 1998. (Banrep,2004). En razón a la información que se produce sobre los detalles de las transacciones, el incentivar pagos con tarjetas representa una alternativa interesante y novedosa en la búsqueda al mejoramiento de la eficiencia del sistema tributario. (Jaramillo, 2004).

Durante los últimos años diferentes sectores del gobierno y la academia han discutido sobre la necesidad de ampliar el acceso a los servicios financieros para la mayoría de hogares posibles y así lograr unas mejores condiciones en términos de oportunidades y bienestar de la población.

Con este fin han surgido diferentes iniciativas del gobierno para ampliar la población bancarizada, entendida como la proporción de los individuos con acceso al uso de los servicios financieros en Colombia. Como consecuencia de esto, el mercado de las tarjetas de crédito en Colombia ha tenido grandes cambios en los últimos años. Ha pasado de tener 8.240.506 tarjetas de crédito vigentes en diciembre del 2010 a 13.752.401 en diciembre del 2015 según los informes de la Superintendencia Financiera de Colombia, lo que representa un crecimiento del 67%. Pero conforme aumenta el número de tarjetas de crédito, también aumenta el número de plásticos cancelados; durante el 2010 se cancelaron 1.351.101 plásticos, mientras que en el 2015 el número de tarjetas de crédito canceladas fue 2.070.176, lo que representa un incremento de cancelaciones del 53% (Superintendencia Financiera de Colombia, 2016).

En la actualidad en Colombia, el uso de la tarjeta de crédito se ha hecho participe de las actividades diarias de los hogares colombianos, partiendo de una utilización de compra más esencial hasta la compra de grandes elementos o servicios (tecnología, viajes, vestimenta, entre otros). Según la descripción anterior, el origen del endeudamiento en los hogares colombianos se porque los usuarios de estos productos no son conscientes del uso que se debe de dar y no llegar a incrementar una deuda de forma desencadena.

La falta de formación financiera del tarjetahabiente y la no promulgación de la economía financiera por las entidades públicas y privadas, han permitido un crecimiento desacelerado en la adquisición de las tarjetas de crédito, adquiriendo más de un producto con la misma entidad u otras entidades, lo que puede aumentar la capacidad de pago del tarjetahabiente.

El hecho de que los tarjetahabientes se endeuden más allá de sus capacidades puede significar un problema en el flujo de gastos del hogar. No existe una cultura de responsabilidad de

consumo, esta es la principal causa por la que los tarjetahabientes tienden a endeudarse por un valor mayor del que puede costear.

Con el fin de mitigar el riesgo de endeudamiento en las familias colombinas, es necesario realizar un cálculo del nivel máximo de endeudamiento, la capacidad de pago, los gastos recurrentes y el endeudamiento con otras entidades financieras.

El cálculo de capacidad de pago debe dejar una holgura de manera tal que mantenga un remanente disponible para atender cualquier imprevisto sin comprometer el cumplimiento de las obligaciones presentes y futuras. En el presente la oferta y demanda de servicios y beneficios en tarjetas de crédito de entidades que emiten el producto no prestan una asesoría básica o personalizada del cómo será el uso del producto.

El gasto mensual promedio del tarjetahabiente está constituido por rubros como alimentación, vivienda, servicios básicos, vestimenta, transporte, salud educación, esta información.

El cupo asignado a la tarjeta de crédito depende de una serie de variables, en las cuales se detalla el nivel de ingresos, egresos, propiedades, vehículos, entre otros, cierto tipo de información se toma en la buena fe tarjetahabiente sin verificar ese tipo de información es verdadera o no, aunque en la mayoría de los casos son datos son verídicos y el cupo asignado son en ocasiones hasta más del doble de los ingresos que se puedan tener lo que con lleva aumenta el nivel de riesgo en la capacidad de pago.

Como existe una política comercial de muchas entidades en ofrecer los servicios de tarjeta de crédito sin cuotas de manejo, se considera que los usuarios de conveniencia que reciben dicho incentivo son subsidiados por los usuarios rotativos, pues siempre existirá un costo positivo (recursos reales utilizados) en el otorgamiento de los créditos (Jaramillo, 2004).

Existen dos tipos de usuarios de tarjetas de crédito: el primer usuario es el que rotan su crédito, quienes no pagan la totalidad de sus saldos cada mes, haciendo uso del crédito de la tarjeta como una fuente de financiación a un plazo mayor a un mes y el segundo usuario el que cancelan la totalidad de sus saldos cada mes, no teniendo que pagar intereses sobre los mismos (Asobancaria, 2007), con respecto a este comportamiento del tarjetahabiente un producto financiero se puede se clasifica tres niveles amplios de satisfacción, el rendimiento de una persona no cumple con las expectativas, el consumidor (no estaría satisfecho), el rendimiento en concordancia con las expectativas (el consumidor está satisfecho); y si el rendimiento de una persona excede a las expectativas (el consumidor se mostrará satisfecho, complacido o entusiasmado), y por ende los consumidores satisfechos son leales por más tiempo, compran más, son menos sensibles a los precios y se expresan en términos favorables con respecto a la empresa. (Kotler, 1996).

En Colombia el acceso al crédito de consumo mediante una tarjeta de crédito es cada vez mayor a lo largo y ancho del país, debido al desarrollo económico que se ha presentado en diferentes sectores y el acelerado surgimiento de entidades con marcas propias de productos financieros, permitidos y respaldados por la superintendencia financiera de Colombia, lo que con lleva a que las entidades financieras presentan un gran competidor a la hora de ofertar y mantener a sus clientes, ofreciendo servicios de calidad, beneficios, presencia, calidad del servicio, entro otros factores influyen en que un cliente tome la decisión de usar o no el producto, solicitarlo o cancelarlo como es el objetivo de este proyecto que es demostrar con un modelo pronostico la probabilidad de deserción de los tarjetahabientes al adquirir una tarjeta de crédito y ser cancelada en los próximos 90 días al ser obtenida.

Con el resultado generado por el modelo pronostico se buscará formular estrategias que permitan mitigar el riesgo de deserción del tarjetahabiente.

Definición, uso y cancelación de la tarjeta de crédito

Las Tarjetas de Crédito son un medio de pago en donde al cliente se le asigna un cupo de crédito que se empieza a consumir haciendo compras en establecimientos de comercio o bien retirando dinero (avances). En su afán de incrementar su participación de mercado, las diferentes entidades emisoras de tarjetas de crédito se han dedicado en los últimos años a otorgar créditos de consumo fácilmente y sin mayores requisitos para lograr una mayor captación de consumidores. (Guillen y Noriega, 2013).

El adquirir un producto financiero, sobre todo un producto activo, es de suma importancia el conocer las diversas tasas que afectan a dicho producto, de esta manera podemos comparar entre productos similares ofrecidos por las distintas entidades financieras, y sobre todo poder analizar las posibilidades de pago de acuerdo a nuestra capacidad de endeudamiento para no incurrir en pagos de mayores intereses ni sobreendeudamiento, lo cual afecta directamente la economía personal y por ende la familiar. (Romero, 2014).

En efecto, la cultura financiera también guarda una estrecha relación con el tipo de educación al que acceden las personas, siendo esta muy diferenciada puesto que existen grupos de personas con mejores niveles de educación y otros niveles de desventaja socioeconómica y de educación y tienen menores niveles de educación financiera.

Igualmente, hay algunas tarjetas de crédito que el cliente no mueve (no transa) ya que muchos clientes las mantienen no como medio de pago sino como un medio para hacer frente a alguna eventualidad.

Machine learning – Análisis de información

El machine learning es una rama interdisciplinaria entre las ciencias del saber que es la estadística e informática, permitiendo ser aplicada en diferentes ámbitos del conocimiento y llevan a explorar áreas como la salud, la aviación, el sistema financiero, entre otros, en cada uno de estos la aplicabilidad ha venido ganado terreno y ha permitido desarrollar herramientas como la inteligencia artificial, la cual ha demostrado un crecimiento cada vez mayo. (APD, 2019).

La aplicabilidad del machine learning ha concebido generar algoritmos más robustos a la hora de identificar una variedad de patrones en el conjunto de datos que son analizados.

El aprendizaje supervisado es una vertiente del machine learning, en la cual se toma un conjunto de datos etiquetados, permitiendo ser utilizados para el entrenamiento del modelo y a su vez dejando como base la utilización de nuevos conjuntos de datos sin ser estos etiquetados en el proceso, según la técnica aplicar se pueden obtener modelos de clasificación o regresión. (APD, 2019).

Para el análisis de datos se aplicará diferentes métodos estadísticos como es la regresión logística que me permite estimar la variable de respuesta en función de las variables continuas, (Rodrigo, 2016), el modelo de Random Forest es un conjunto (colección) de árboles de decisión de tipo ensacado que entrena varios árboles en paralelo y utiliza la decisión mayoritaria de los árboles como la decisión final del modelo de bosque aleatorio. (Misra - Li, 2020).

Marco Metodológico

En la etapa Análisis: En las entidades financieras en Colombia se presenta una serie de planteamientos investigativos a favor o no del producto que será expuesto al público, para este tipo observaciones se planteó en el estudio actual, la identificación la probabilidad de deserción de los tarjetahabientes al adquirir TDC y ser esta cancelada en periodo de 90 días, para esto se utilizó el modelo de regresión logística que permite dar un contexto estadístico a la variable de interés, se utilizará herramientas de análisis del modelo como es la curva de ROC, el AUC de la curva y además una validación cruzada en el modelo de regresión cruzada para medir su efectividad.

En la etapa de diseño: Se plantea realizar una estandarización de las variables predictoras con el fin de mitigar la ausencia de datos.

$$z = \frac{x - \mu}{\sigma}$$

En la etapa de implementación: En esta etapa se plantea implementar las siguientes herramientas para el análisis de los datos.

Modelo de regresión logística: Permitirá obtener el mejor modelo para analizar su predicción.

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Matriz de confusión: Permite evaluar el desempeño de clasificación del modelo de regresión logística alcanzado, por medio de los siguientes ítems:

- Exactitud: Es el número de predicciones correctas
- Razón verdaderos positivos – Recall: Casos positivos identificados correctamente
- Razón falsos positivos: Casos negativos clasificados incorrectamente positivos.

- Razón verdaderos negativos: Casos negativos identificados correctamente
- Razón falsos negativos: Casos negativos clasificados incorrectamente negativos.
- Precisión: Casos predichos positivos.

Curva de ROC: Es la representación gráfica del desempeño obtenido por el clasificador, en el cual se validará la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR).

AUC: Permite identificar la medición definida por la curva de ROC.

Desarrollo del Modelo

El objetivo principal de este proceso de investigación es realizar una demostración por medio de un modelo pronostico que mida la deserción del tarjetahabiente de una entidad financiera en Colombia al adquirir la TDC, en los próximos 90 días.

Con el fin de tener una base de datos acorde al estudio planteado, se ejecutaron diferentes conversatorios con técnicos de las áreas de riesgos y clientes relacional y se obtuvieron 1, que permitieron realizar una depuración de las variables que podrían aportar de forma significativa al modelo de pronóstico, el resultado obtenido fueron 18 variables que se clasifican a continuación.

Tipo de Variable	Cantidad
Dicotómicas	5
Continuas	6
Atributiva	2
Nominal	3
Discreta	2

Tabla 1. Clasificación de variables

Para dar un contexto al comportamiento de los registros con un total de 9182, de los cuales 2590 que representan el 28% de los clientes inactivos.

Cliente Activo – 1: Han utilizado la TDC durante 90 días o más.

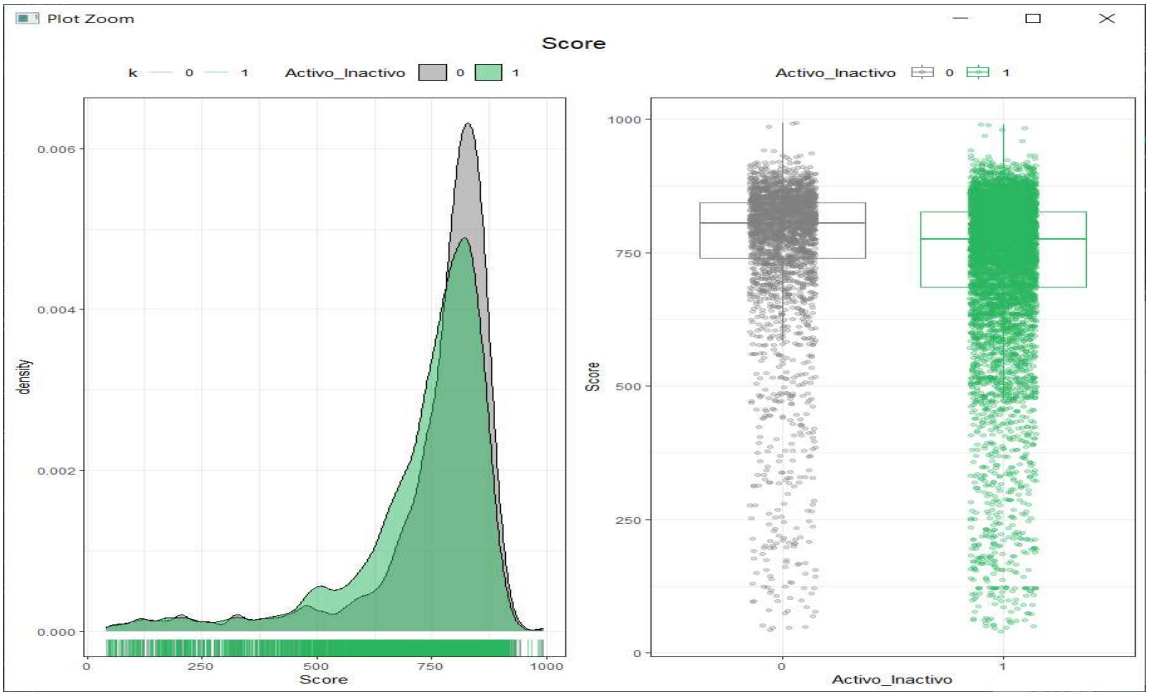
Cliente Inactivo – 0: No Han utilizado la TDC durante los 90 días o más.

Grupo Edades	Cliente Activo	%
GRUPO 18 A 29	633	7%
GRUPO 30 A 39	2494	27%
GRUPO 40 A 49	1822	20%
GRUPO 50 A 59	1136	12%
GRUPO 60 A 90	507	6%
TOTAL	6592	72%

Grupo Edades	Cliente Inactivo	%
GRUPO 18 A 29	7	0%
GRUPO 30 A 39	480	5%
GRUPO 40 A 49	424	5%
GRUPO 50 A 59	561	6%
GRUPO 60 A 90	1118	12%
TOTAL	2590	28%

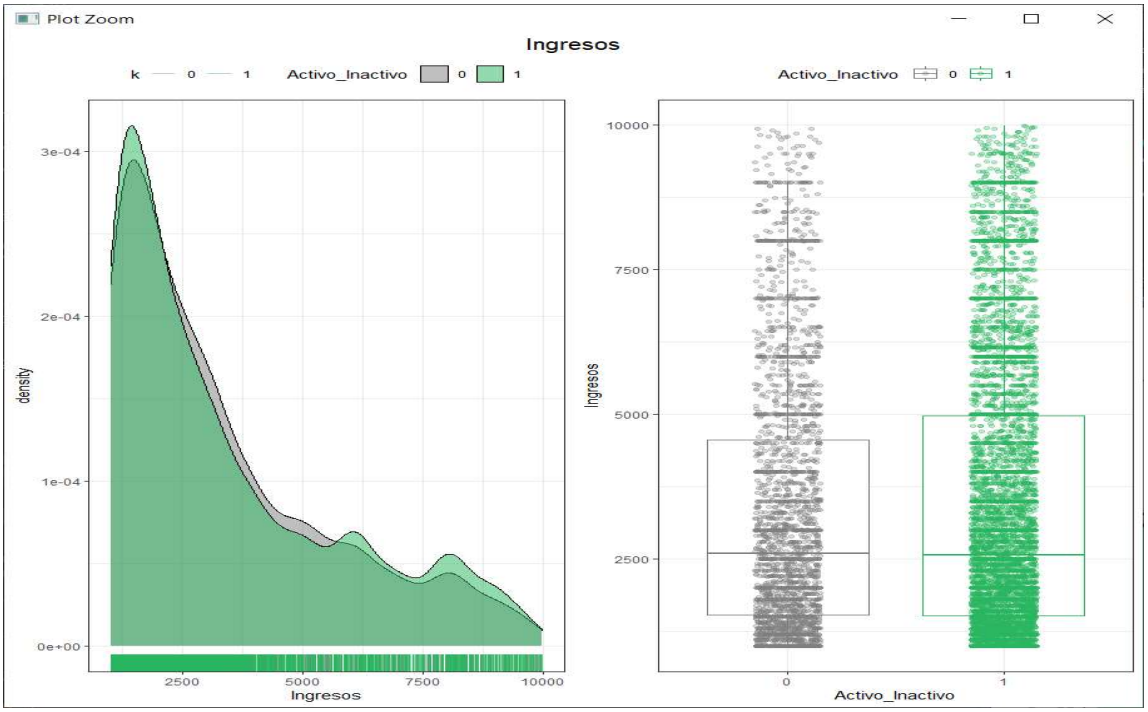
Tabla 2. Grupo de edades

En esta clasificación se identifica que el mayor comportamiento de clientes activos esta un rango entre 30 a 59 años, con respecto a los clientes inactivos donde menos tiene comportamiento de usabilidad es entre los 60 y 90 años.



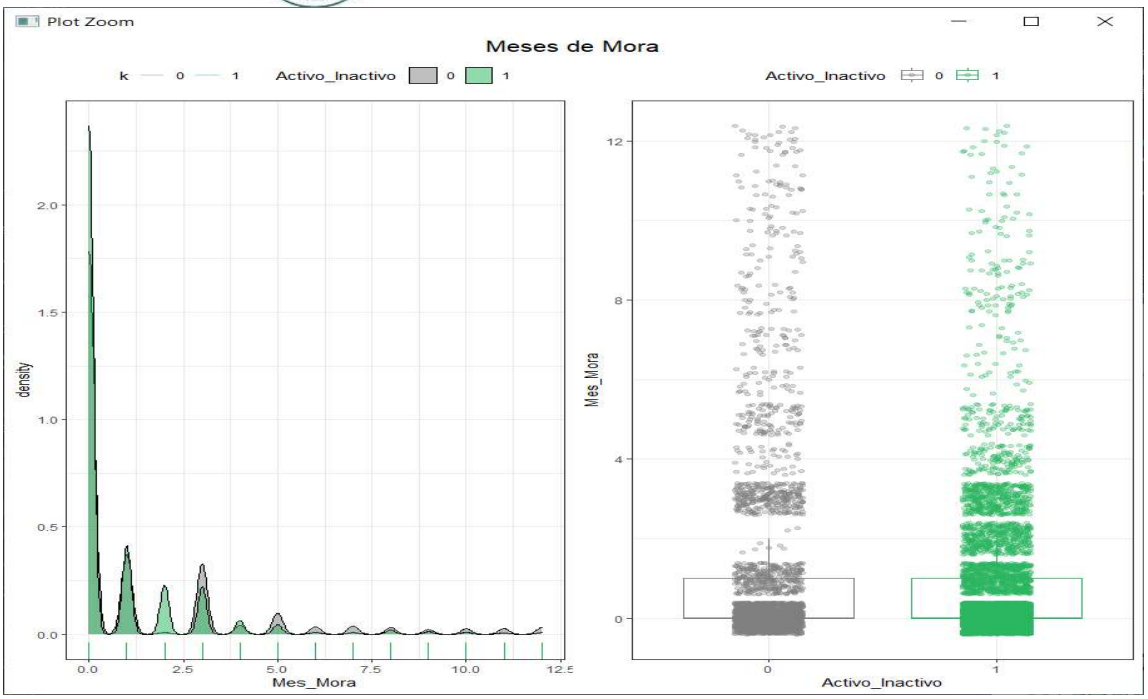
Grafica 1. Relación Score vs Cliente Activo e Inactivo

Con respecto al análisis de riesgo definido por la entidad financiera que es de 500 puntos en el score interno se evidencia en la población para ambos estados superar en gran medida el lumbral definido por la entidad.



Grafica 2. Ingresos vs Cliente Activo e Inactivo

Con respecto al ingreso reportado por los clientes en la entidad se observa un rango entre 1 millón y 5 millones.



Grafica 3. Meses de Mora vs Cliente Activo o Inactivo

MES MORA	Cliente Activo	%
0	4569	50%
1	796	9%
2	444	5%
3	190	2%
4	326	4%
5	124	1%
6	28	0%
7	19	0%
8	17	0%
9	28	0%
10	23	0%
11	12	0%
12	16	0%
TOTAL	6592	72%

MES MORA	Cliente Inactivo	%
0	1624	18%
1	340	4%
2	8	0%
3	298	3%
4	38	0%
5	90	1%
6	32	0%
7	35	0%
8	29	0%
9	20	0%
10	24	0%
11	24	0%
12	28	0%
TOTAL	2590	28%

Tabla 2. Meses de Mora

Se observa un comportamiento favorable en la altura de la mora con una disminución considerable en los clientes activos, pero en los clientes inactivos se evidencia un 62% frente a los 2590 clientes que no la han utilizado.

Co respecto al modelo pronósticos se empleó la regresión logística, aplicada a las 18 variables definidas para la base de datos; con el fin de realizar una validación del modelo se efectuó una partición de la base de datos de un 80% para entrenamiento del modelo y un 20% para las pruebas del modelo. Se entrenaron un total de 5 modelos y en cada uno de ellos se buscó las variables que mayor significancia a porta al modelo y en el cual se obtuvo al final las siguientes variables:

Valores estadísticamente significativos del modelo

GENERO	INGRESOS	MES_MORA	SCORE	EDAD	SALDO DISPONIBLE
2e-16 ***	0.0021 **	4.8e-10 ***	0.0201 *	2e-16 ***	0.0406 .

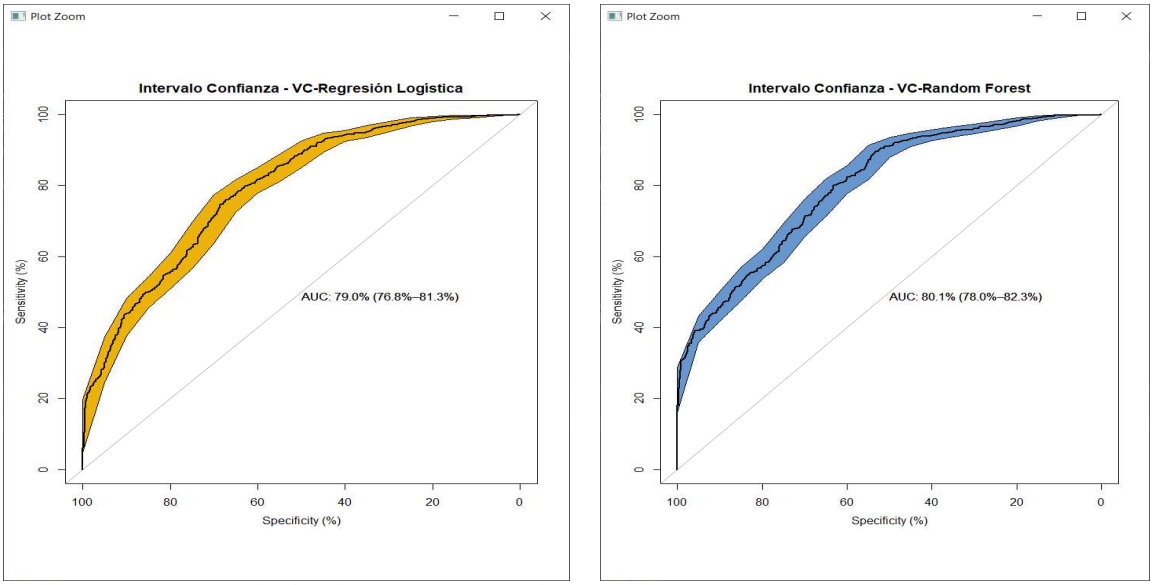
Genero	Al cambiar de categoría de hombre a mujer se presenta una disminución del 51.85%, de no conservar la TDC en los próximos 90 días a su adjudicación.
Ingresos	El ingreso o salario registrado por el cliente no es determinístico para conservar o no la tarjeta antes o después de los 90 días.
Mes de Mora	Al realizar un incremento de un punto en los meses de mora, se presenta una disminución del 8.14% en los clientes de igual características, para no conservar la TDC en los próximos 90 días de su adjudicación.
Score	Al realizar un incremento de un punto en el score, se presenta una disminución del 0.05% en los clientes de igual características, para no conservar la TDC en los próximos 90 días de su adjudicación.

Edad	Al realizar un incremento de un punto en la edad, se presenta una disminución del 7.68% en los clientes de igual características, para no conservar la TDC en los próximos 90 días de su adjudicación.
Saldo Disponible	El saldo disponible para conservar o no la tarjeta antes o después de los 90 días.

Tabla 3. Variables modelo Final – Regresión Logística

Con respecto al resultado obtenido con el entrenamiento del modelo final, se realizó una validación cruzada entre la regresión logística y el modelo Random Forest con una cantidad de 1000 ramas y se observó que ambos se aproximan al AUC definido en la curva de ROC.

Modelo - Precisión	Sin Validación Cruzada	Con Validación Cruzada
Regresión Logística	77.94%	79.0%
Random Forest	79.19%	80.1%



Grafica 4. Curva ROC Regresión logística vs Curva ROC Random Forest

CONCLUSIONES

En el desarrollo del proyecto, realizar reuniones interdisciplinarias con los técnicos de las áreas de riesgos y de clientes, permitieron obtener un rumbo al planteamiento a investigar.

La pérdida de tarjetahabientes no solo genera un costo de oportunidades sino en ganancias sino a la emisión de nuevas tarjetas de crédito.

En la relación con los meses de mora, se identificó que en los tres primeros meses de adquirir una tarjeta de crédito se presenta un mayor porcentaje de inactividad principalmente en el mes 0 con un 18% de la población, esto significa un porcentaje de la participación en el momento de realizar la captación y sostenibilidad del cliente con sus procesos de finalización.

El resultado generado por los modelos de clasificación es evaluado con la herramienta de medición de la curva ROC, en la cual se identifica la mejor clasificación entre los modelos dando a identificar el Random Forest permite tener una clasificación mayor del 80.1% contra el 79% obtenido del modelo de regresión logística.

Con el resultado obtenido del modelo de Random Forest que presento una probabilidad mayor de clasificación con respecto al modelo de regresión logística, se analizaran nuevas variables que permitan identificar otros tipos de objetivos como la falta de educación financiera en los sectores socio-económicos para poblaciones vulnerables, que importancia puede tener en las familias una tarjeta de crédito en tiempos de crisis económicas.

REFERENCIAS BIBLIOGRÁFICAS

Digital y Revista APD. (2019). ¿Qué es Machine Learning y cómo funciona?, España, Redacción APD, . https://www.apd.es/que-es-machine-learning/

Rodrigo. J.A. (2016). Regresión logística simple y múltiple, España, ciencia de datos, https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple#bibliograf%C3%ADa/

Misra. S, Li. H, (2020). Regresión logística simple y múltiple, EUA, Science Direct, https://www.sciencedirect.com/topics/engineering/random-forest

Kotler, P. (1996). Análisis, Planeación, Implementación y Control. Recuperado de https://anafuenmayorsite.files.wordpress.com/2017/08/libro-kotler.pdf

L, Jaramillo. (2004). Las tarjetas de crédito en Colombia: evolución e impacto sobre el consumo y el recaudo tributario. Recuperado de https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/1282/Repor_Julio_2004_Arbelaez_y_Zuleta.pdf?sequence=2&isAllowed=y

Romero, P. (2004). Influencia de la cultura financiera en los clientes del banco de crédito del Perú de la ciudad de Chiclayo, en el uso de tarjetas de crédito, en el periodo enero – julio 2013. (Tesis Pregrado), Chiclayo, Universidad Católica Santo Toribio De Mogrovejo.

Fernandez, S. (2011), Regresión logística, España, Universidad autónoma de Madrid, http://www.estadistica.net/ECONOMETRIA/CUALITATIVAS/LOGISTICA/regresion-logistica.pdf

Yiu, T. (2019), Understanding Random Forest, Tower data Science, EUA, https://towardsdatascience.com/understanding-random-forest-58381e0602d2

DataCamp. (Productor). (2016). R tutorial: Cross-validation[You], https://www.youtube.com/watch?v=OwPQHmiJURI

Kassambara. (11/03/2018). Regression Model Validation, Statistical tools for high-throughput data analysis, England, http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/

The Data Science Show. (Productor). (2018). Cross Validation using caret package in R for Machine Learning Classification & Regression Training [You], https://www.youtube.com/watch?v=Zd9GRoQjKvo