Modelo de próxima oferta para clientes de una entidad de crédito utilizando técnicas de aprendizaje automático

Sergio Andrés Henao Quintero

Trabajo de grado presentado para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Hernán Felipe García Arias, Doctor (PhD) en Ingeniería

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	Muñoz Zapata y Martínez Naranjo [1]
Referencia Estilo IEEE (2020)	[1] A. Sergio Henao Quintero, “Modelo de próxima oferta de una entidad de créditos utilizando técnicas de aprendizaje automático”, Trabajo de grado especialización, Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2023.

Especialización en Analítica y Ciencia de Datos, Cohorte III.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Dedicatoria

Quiero dedicarle este trabajo principalmente a mi madre y futuro esposa , quienes desde un principio me dieron su incondicional apoyo , a mi madre Maria Edilma Quintero que me apoyo en las labores del hogar para que pudiera concentrarme en poder trabajar y estudiar y a mi futura esposa Ana Maria Urán por apoyarme con sus conocimientos en los campos del área analítica, además de su sabiduría y apoyo emocional para sacar adelante este proyecto.

Agradecimientos

Infinitos agradecimientos al profesor Hernán Felipe García Arias , por todos los aportes que fueron de gran ayuda para culminar con éxito este proyecto, y que a pesar de la dificultades administrativas que retardaron el inicio de la tutoría , tuvo la mejor disposición para asesorarme de manera clara y precisa.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

Argmin : argumento mínimo. Representa el conjunto de valores o valor que hace mínima una función

max: máximo

Min : Mínimo

RESUMEN

En este proyecto, se aborda el desafío de aumentar la tasa de clientes impactados por ofertas comerciales en una entidad financiera de un 4% a un 25%. Mediante la aplicación de análisis de datos avanzado y modelado predictivo, se construye un modelo analítico para personalizar las ofertas de productos financieros según las necesidades y preferencias de cada cliente.

Se utiliza modelos de aprendizaje automático, específicamente Gradient Boosting, Random Forest y One vs Rest Classifier, para analizar patrones de comportamiento del cliente y predecir su respuesta a diferentes ofertas de productos. Para asegurar la calidad y representatividad de los datos de entrenamiento, se aplican varias técnicas de preprocesamiento de datos y balanceo de clases.

Para evaluar la efectividad del modelo, se usan varias métricas de desempeño, incluyendo precisión, recall, F1_score y AUC. A través de esta investigación, se busca como demostrar como un enfoque de análisis de datos centrado en el cliente puede contribuir a mejorar la satisfacción del cliente, optimizar la eficacia de las campañas de marketing y, en última instancia, mejorar los resultados comerciales de la entidad financiera.

ABSTRACT

In this project, we address the challenge of increasing the customer impact rate from commercial offers in a financial institution from 4% to 25%. Through the application of advanced data analysis and predictive modeling, we build an analytical model to personalize financial product offers according to the needs and preferences of each customer.

Machine learning models, specifically Gradient Boosting, Random Forest, and One vs. Rest Classifier, analyze customer behavior patterns and predict their response to different product offers. Various data preprocessing techniques and class balancing are applied to ensure the training data's quality and representativeness.

We use several performance metrics to assess the model's effectiveness, including precision, recall, F1 score, and AUC. Through this research, we aim to demonstrate how a customer-focused data analysis approach can improve customer satisfaction, optimize marketing campaigns' effectiveness, and ultimately improve the financial institution's business outcomes.

I. INTRODUCCIÓN

En el contexto actual las compañías financieras enfrentan el desafío de ofrecer productos y servicios que se ajusten a las necesidades y expectativas de manera efectiva y personalizada. La competencia financiera es cada vez más intensa y la situación actual económica que está enfrentando el país, obliga a la empresa a buscar estrategias que retengan a los clientes más que buscar a nuevos clientes.

Una de las estrategias más usadas desde la analítica que me permite enfrentar este reto es la creación de un sistema de recomendación , como es el caso de [1] que utiliza un sistema de recomendación hibrida y que combina diferentes técnicas de aprendizaje automático y minería de datos para ofrecer recomendaciones más precisas personalizadas a los clientes y que pueden ayudar a dar un enfoque al proyecto.

Con este panorama en mente, el proyecto de grado tiene como objetivo construir un modelo predictivo que, basado en el conocimiento profundo de los clientes de una compañía financiera, busca identificar cuáles son los productos más afines a sus clientes (11 productos en total). El propósito es permitir al negocio ofrecer productos que se ajusten mejor a las necesidades del cliente, aumentando así su nivel de profundización y fidelización hacia los productos de la compañía.

Inicialmente, el primer paso en la construcción del modelo es la extracción de información financiera, crediticia y transaccional de los clientes, así como de los productos que adquieren en un período de tres meses previo al desembolso del crédito. Además, se consolida la variable de respuesta, que es el producto que el cliente adquirió una vez realizado el desembolso.

Posteriormente, se consolidó la información extraída y se llevó a cabo un preprocesamiento de os datos que incluyó la imputación de datos, la organización del formato de las variables categóricas, análisis descriptivo de algunas variables consideradas importantes para el negocio y la transformación de las variables categóricas en numéricas por medio de la metodología one-hot-encoding.

En la siguiente etapa, se realizó una ingeniería de características, comenzando con la eliminación de variables con alta correlación mediante la correlación de Spearman [2]. Después, se detectaron datos atípicos utilizando la metodología del vecino más cercano (Clúster N-vecino más cercano [3]) y se eliminaron las filas correspondientes a estos datos atípicos. Por último, se llevó a cabo un análisis de entropía para evaluar la pérdida de información entre la data antes y después de eliminar los datos atípicos, verificando que la pérdida de información no fuera significativa.

Se analizó la variable de respuesta y se encontró que la distribución de sus características estaba desbalanceada, lo que llevó a aplicar metodologías de balanceo de datos. La base balanceada se entrenó con tres algoritmos diferentes: Gradient Boosting [4], Random Forest [5] y One vs Rest Classifier[6] (Regresión Logística 1 vs el resto). Cada algoritmo se ajustó con diferentes hiperparámetros y se evaluó utilizando métricas de desempeño. El modelo de Random Forest obtuvo los mejores resultados, con un F1-score del 65% y una precisión del 65%.

El paso más importante fue analizar las predicciones de los diferentes productos y su relación tanto en las predicciones y métricas como en la realidad del negocio. Se realizaron ajustes en el modelo, eliminando ciertas características en base a los análisis y decisiones del negocio, obteniendo asi una mejora en la predicción del modelo en general.

Finalmente, se realizó un análisis de violines para observar el comportamiento de las probabilidades de cada característica (producto) y probar algunos umbrales de predicción que pudieran mejorar las predicciones del modelo, especialmente en aquellas características (productos) cuyas métricas individuales como el recall y el F1-score son bajas.

II. PLANTEAMIENTO DEL PROBLEMA

Actualmente las ofertas en la empresa crediticia se basan en reglas rígidas de negocio , las cuales son construidas a través del conocimiento de los personas encargadas de los diferentes producto ,que en base a su experiencia y sus conocimientos del mercado eligen y filtran a los clientes, y escogen a aquellos a los cuales se les ofrecerán ofertas. Dichas reglas de negocio son construidas sin tener en cuenta el analisis de los datos por lo que están muy limitadas y solo un porcentaje de ellas son ofrecidas. Como resultado, solo se ofrece un pequeño porcentaje de ofertas ,dejando afuera muchos clientes potenciales que podrían estar interesados en ellas.

En concreto, los clientes a los que les ofrecen ofertas en la actualidad representan solo el 4% del total de la base completa de todos los clientes de los clientes activos en la empresa, este enfoque que actualmente se desarrolla, no solamente es limitado si no también en ocasiones equivocado. Alguno de los errores más comunes es ofrecerle a un cliente la misma oferta en un corto periodo tiempo, también en ocasiones se realizan ofertas que no eran de la necesidad y la capacidad de los clientes, disminuyendo su confianza con la empresa.

Por otro lado durante mucho tiempo no se ha podido superar la cantidad de un 4% de ofertas en un lapso de 3 meses, ya que se basa en la capacidad humana de ofrecerla y esto no genera incremento en las ventas.

Otro problemas es que actualmente solo se hace ofertas a los clientes de los productos más comunes y representativos de la empresa, como son los créditos de moto gamma baja o los vehículos usado, dejando de lado otros producto potenciales sin ofrecerse al cliente, lo que evita que estos tengan crecimientos en sus ventas, e inclusive algunos de ellos han tenido un porcentaje muy bajo de ventas en el transcurso de un año o dos años.

IV. OBJETIVOS

A. Objetivo general

Construir un modelo predictivo que permita identificar el próximo producto de financiación más afín a las necesidades de los clientes de una entidad crediticia utilizando modelos de aprendizaje supervisado.

B. Objetivos específicos

- Caracterizar los productos de financiación para los clientes utilizando esquemas de análisis multivariado.
- Entrenar tres modelos de aprendizaje supervisado que permitan realizar predicciones sobre la próxima oferta utilizando enfoques Uno vs todos
- Evaluar el desempeño de los modelos de próxima oferta para clientes de una entidad de crédito en términos de su especificidad y sensibilidad como indicador verificable de la robustez en la predicción

VI. MARCO TEÓRICO

"En los últimos años, los sistemas de recomendación han experimentado un auge notable, especialmente en las industrias de comercio electrónico y streaming. Es evidente al considerar la experiencia del usuario al buscar en Google; los algoritmos proporcionan un listado de recomendaciones ajustadas a los criterios de búsqueda definidos por el usuario.

Los sistemas de recomendación surgieron en la década de los 90 debido a la necesidad de las industrias de recomendar un elemento específico a los usuarios. Este problema ha sido abordado con enfoques variados, incluyendo sistemas de filtrado y soluciones basadas en modelos de aprendizaje automático.

Entre los trabajos pioneros en sistemas de recomendación se encuentra el estudio elaborado por Goldberg y compañía [7] en el que se diseñó un algoritmo para filtrar correos electrónicos basándose en los comportamientos y opiniones de los usuarios. Otro hito significativo fue el sistema de recomendación agregada GroupLensOtros [8], que utiliza análisis de noticias de internet (NetNews). Los usuarios califican las noticias que leen, y esta retroalimentación se utiliza para generar recomendaciones más pertinentes para ellos.

Con la llegada de los años 2000, las soluciones propuestas adoptaron una nueva dirección al incorporar algoritmos basados en aprendizaje automático. Un ejemplo destacado de este enfoque es el trabajo realizado por Sarwar [9], A diferencia del filtrado colaborativo que calcula similitudes entre usuarios, este enfoque se centra en calcular similitudes entre elementos, utilizando técnicas avanzadas de aprendizaje automático.

Mas adelante las investigaciones sobre el uso de algoritmos de aprendizaje automático para la predicción y calificación de elementos fueron potenciadas por el premio Netflix , estos sucedió entre el año (2006 y 2009) donde el objetivo de esta competición era hacer predicciones precisas

de las clasificaciones de las películas. Al día de hoy este tipo de clasificación esta muy en auge y se puede ver que es utilizada en casi todas las plataformas de streaming.

Se podría decir que en todos estos años de investigación de algoritmos , se tendrían ya todos los problemas resueltos, ya que a través del tiempo ha habido muchas mejoras algorítmicas , sino que también se han abordado mucho enfoques sobre el temas , en la diversidad de temas en que estos algoritmos de recomendaciones se han utilizado. Pero estos algoritmo aun tiene limitaciones y las investigaciones han ido mucho más allá como por ejemplo la recomendaciones Consciente de la secuencia [10] , el cual es un enfoque que se basa en la premisa de que el orden en el que los usuarios interactúan con los elementos puede proporcionar información valiosa sobre sus preferencias actuales y futuras.

Un aspecto importante de la recomendación consciente de la secuencia es la recomendación de "próximos elementos". En lugar de recomendar los elementos más populares o los elementos que son similares a los que el usuario ha interactuado en el pasado, este enfoque se centra en recomendar elementos que sean relevantes para explorar a continuación, dada la secuencia de interacciones del usuario. La recomendación consciente de la secuencia ha demostrado ser efectiva en diversas aplicaciones, desde la recomendación de música hasta la personalización de la navegación web.

Una consideración crucial en los sistemas de recomendación es la diversidad de las recomendaciones. Los sistemas de recomendación deberían evitar recomendar siempre los mismos elementos o elementos muy similares entre sí, ya que esto podría llevar a los usuarios a percibir el sistema como monótono o poco útil. En cambio, un buen sistema de recomendación debería ser capaz de presentar una variedad de elementos que sean relevantes para los intereses del usuario.

Ekstrand et al. [11] exploraron este problema en su estudio, donde discuten la importancia de equilibrar la precisión y la diversidad de las recomendaciones. Argumentan que, aunque la precisión (recomendar elementos que los usuarios encontrarán relevantes y disfrutarán) es importante, también lo es la diversidad (recomendar una variedad de elementos que permitan a los usuarios descubrir nuevos intereses).

En el desarrollo de sistemas de recomendación y, en general, en cualquier aplicación de aprendizaje automático, los conjuntos de datos de referencia y las medidas de precisión desempeñan un papel crítico. Los conjuntos de datos de referencia proporcionan un medio para entrenar y validar modelos de aprendizaje automático, mientras que las medidas de precisión ofrecen una forma de evaluar y comparar la efectividad de diferentes modelos y algoritmos (Wagstaff, 2012) [12].

Sin embargo es importante aclarar que cuando se utiliza un algoritmo para hacer una recomendación de manera ofline (cuando no se recomienda en tiempo real , si no que se utiliza de manera especifica en un tema particular) , es decir se retiene una parte de las interacciones de los usuarios (por ejemplo, compras, clics, visualizaciones, etc.) , puede no reflejar con precisión el rendimiento que un algoritmo tendría en la práctica. Aunque este enfoque es ampliamente utilizado, varios trabajos han señalado que puede no ser adecuado para estimar cómo se comportaría un algoritmo en un entorno de producción real. Esto se debe a una serie de factores, como la posibilidad de sesgo de selección en los datos de entrenamiento y la variabilidad en los comportamientos de los usuarios [12].

Por lo tanto, es crucial complementar las evaluaciones fuera de línea con otras formas de evaluación, como pruebas A/B en vivo o simulaciones de interacción con usuarios, para obtener una imagen más completa del rendimiento de un sistema de recomendación.

VII. METODOLOGÍA

El problema del negocio a abordar desde la perspectiva de la analítica de datos es el diseño y desarrollo de una solución analítica que permita incrementar el porcentaje de clientes impactados por ofertas comerciales del 4% actual hasta, al menos, un 25%. Para lograr este objetivo, se requiere la construcción de un modelo de aprendizaje automático capaz de analizar y comprender de manera más efectiva las necesidades y preferencias de los clientes, así como identificar patrones de comportamiento que puedan ser utilizados para optimizar y personalizar las ofertas comerciales. Al adoptar un enfoque más centrado en el cliente, la empresa podrá ampliar su alcance, aumentar la satisfacción del cliente y mejorar sus resultados comerciales.

Algunos artículos proporcionen una visión general de sistemas de recomendación que nos pueden ayudar en el proyecto, como es el caso de [13] que trabajan un modelo de puntuación crediticia basado en un aprendizaje profundo para mejorar la inclusión financiera. Aunque este enfoque se centra en la puntuación crediticia, las técnicas y algoritmos de predicción utilizados pueden ayudar a dar una idea de cómo adaptarlo a productos financieros personalizados. Para esto se la mayor características de estos modelo se caracteriza por tener una variables respuesta, en nuestro caso esta variable respuesta estará compuesta de varias características, donde cada característica representará un producto diferente, por lo que es importante que los modelos analíticos que se evalúen tengan la capacidad de predecir este tipo de variables respuesta.

Uno de los desafíos principales en este proyecto es la construcción de una base de datos adecuada para el entrenamiento y la predicción del modelo. Dicha base de datos debe contener información suficiente para abordar todas las características del cliente que influyen en su decisión de adquirir un producto con la empresa. Es fundamental contar con información completa y relevante para identificar estas características. En el caso de clientes nuevos o aquellos con poca historia debido a un corto período de vinculación, se requiere buscar alternativas para complementar esta información y así lograr un modelo eficientemente entrenado. Este proceso constara de las siguientes fases:

1) Construcción de la data de entrenamiento : Consiste en construir una base previamente, analizada, procesada y finalmente moldeada con ingeniería de características, con el fin de obtener una base que ayude a entrenar el modelo de predictivo.

2) Modelamiento : Consiste en construir el modelo que permita predecir la variable objetivo.

3) Evaluación: Se evalúa el modelo predictivo por medio de una métricas establecidas para establecer que tan eficiente es.

4) Validación: Una vez ha sido evaluado y el resultado es el esperado , se someten datos nuevos al modelo para realizar la predicción.

Descripción de la base de datos

Es importante antes de hacer la descripción de los datos tener presente que la información para emplear en este proyecto fue suministrada por una reconocida entidad de financiación crediticia con la cual el estudiante cuenta con un vínculo laboral actualmente, razón por la cual dicha información no podrán ser almacenada y/o expuesta en repositorios externos y solo podrá ser empleada para fines académicos como la realización de este proyecto, dada la confidencial y sensibilidad de estos.

En ningún momento los resultados obtenidos podrán ser expuestos de tal manera que se permita la individualización de los clientes, Adicionalmente, los informes o algoritmos que puedan generarse durante el desarrollo de este trabajo podrán ser utilizados por la entidad financiera sin ninguna restricción.

El dataset consta de 330833 registros, donde cada registro representa un cliente; las características del dataset se representan por 87 variables , de las cuales 12 son categóricas y 65 son numéricas.

Las variables se componen en grupos de categorías que describen grupos características del cliente, estas son:

Variables sociodemográficas: Describen los componentes fundamentales y sociales del cliente. Algunas de las variables incluidas en el dataset, son la Edad , actividad económica, la región donde vive y el género.

Variables Financieras: Describen el desempeño financiero del cliente. Algunas de las variables incluidas en el dataset son, promedio ingresos ,cantidad de seguros vigentes ,calificación cartera entre otras variables.

Variables Crediticias: Describen el comportamiento crediticio del cliente . Algunas de las variables incluidas en el dataset son , valores desembolsado de un producto de una clase , # número de producto , moras , montos de tarjetas de crédito entre otras.

TABLA I DESCRIPCION DE LA BASE DE DATOS

Tipo de variable	Variable	Descripción
Variables sociodemográficas	Edad	Edad del cliente
	Genero	Genero del cliente
	Actividad Económica	Actividad económica principal
	Departamento	Departamento de donde vive el cliente.

Variables Crediticias	Nivel de riesgo	El nivel de riesgo del cliente con la entidad
	Valor desembolsado (actual)	Valor en pesos desembolsado del producto actual .
	Valor desembolsado(T-3 -T0)	Valor en pesos desembolsado del producto adquirido en el periodo tiempo.
	Mora (Actuales)	Mora del producto actual
	Mora (T-3 -T0)	Mora del producto que tenia el
	Saldo Promedio (T-3 – T0)	cliente en el periodo de tiempo.
	Promedio Monto Tarjetas de Créditos (T-3-T0)	Promedio del saldo de las tarjetas de crédito que tuviera activas en el periodo de tiempo.
	Cantidad de tarjetas activas.(T-3-T0)	Promedio del tipo de la tarjeta De crédito activa en el periodo de tiempo.
	Tipo de tarjeta	Tipo de la tarjeta de crédito
Variables Financieras	Ingresos Mensuales(Actuales)	Ingresos mensuales que el cliente registra actualmente
	Promedio Ingresos Mensuales(T-3 -T0)	Promedio de los ingresos mensuales que el cliente registro en ese periodo de tiempo.
	Calificación Cartera	Representa la capacidad de un cliente en cumplir con sus compromisos financieros
	Calificación Cartera (T-3-T0)	La misma definición de calificación de cartera pero el periodo descrito.

	Afinidad transaccional	Representa el gusto del cliente a nivel.
	producto adquirido (T-3 -T0)	Producto adquirido en ese periodo tiempo
	Cantidad de producto adquiridos(T0- T-3)	Cantidad del producto en el periodo de tiempo
Producto	Producto adquirido (variable respuesta)	Es el producto sobre el cual se analiza las características del cliente, con el objetivo de poder hacerles una oferta en base a estos productos.
	Fecha de desembolso (T0)	Fecha de desembolso en la que el cliente adquirió el producto. Representa el T0
	Fecha de desembolso actual	En el caso en el que el cliente sea nuevo, se toma la fecha del producto actual.

La TABLA I están representadas la mayoría de las variables del dataset . T-3 – T0 representa la historia de la variable en un lapso de 3 meses, antes de adquirir unos de los productos de la variable respuesta. En el caso de las variables producto (T-3 -T0) , cantidad de producto adquirido(T-3 -T0), estas representan en realidad todas la variables asociadas a la adquisición de un determinado producto y la cantidad de este en el periodo de tres meses antes de adquirir el producto asociado a la variable respuesta (se incluyen producto también diferentes a los de la variable respuesta).

Cuando el cliente es nuevo o tuvo un paso esporádico con la entidad, se toma el producto la fecha de desembolso actuales y se le asocian algunas variables actuales.

Exploración de los datos

El siguiente paso es realizar un analisis descriptivo de la base de datos con el fin de entender el negocio de los diferentes productos que se va a ofertar, en especial se hace un analisis de la variable respuesta para entender la distribución de esta y de alguna variables que se consideran importantes para el negocio.

Preprocesamiento de la data

El primer paso para explorar los datos es revisar las variable que tienen datos nulos, y la cantidad de estos, luego se e realiza una imputación de estos, teniendo muy presente su significado en el negocio.

Luego se estandarizan las variables numéricas, ya que hay variables que tienen unos rangos más amplios que otras variables, asi evitamos al momento de construir el modelo de clasificación, la efectividad vea afectado su desempeño de manera negativa.

Posteriormente todas las variables numéricas se escalan con el método de Min Max escale

$$Xescaled = X - Xmin / Xmax - Xmin$$. [14]

Algunas de las variables categóricas se componen de muchas características , y la estrategia para utilizar one-hot encoding, es graficar y revisar el porcentaje de cada característica para agrupar

aquellas que no represente un gran porcentaje de la variable, asi no creamos una variable por cada características y evitamos en lo posible la maldición de la dimensionalidad al entrenar el modelo.

Ingeniería de características

Lo siguiente es realizar un analisis de correlación entre las variables y observar cuales variables se correlacionan entre sí. Para este proceso se utilizó la metodología de correlación de spearman.

TABLA II CORRELACION DE SPEARMAN

Correlación de Spearman	Descripción
$Rr = 1 - \frac{6 \sum_i d_i^2}{n(n^2-1)}$ [2]	Rr = Coeficiente de correlación de spearman d = Diferencia del rango del elemento n. n = Número de puntos de Datos de las dos variables

Posteriormente al hacer hallazgos de los datos atípicos en la data, utilizamos es el método de local outlier factor(LOF) [3], la cual es una técnica es una técnica de detección de anomalías que se basa en la densidad local de los puntos de datos. LOF asigna a cada punto un valor de "outlier" que indica cuán atípico es en relación con sus vecinos.

Posterior a esto calculamos la entropía [15] de las variables antes y después de eliminar los datos atípicos, esto nos ayuda a evaluar el impacto de la eliminación de datos atípicos en la información contenida en el conjunto de datos. Si la entropía no disminuye significativamente después de eliminar los datos atípicos, esto indica que la pérdida de información es mínima y que la calidad de los datos ha mejorado.

Partición de la data de entrenamiento y test más balanceo de data de entrenamiento

El conjunto de datos se dividirá en una proporción 70-30, para conformar las bases de entrenamiento y testeo, respectivamente. Debido al desbalance de clases presentes en los datos, se crearán 3 bases de entrenamiento diferentes, según las técnicas de muestreo que se exponen en la TABLA III.

TABLA III TECNICAS DE BALANCEO DE LA DATA

Técnica	Forma de selección
Oversampling [16]	Es una técnica más simple en la que se duplican los ejemplos de la clase minoritaria para aumentar su número (todos los productos diferentes a moto gamma).
Smote [17]	Es una técnica de sobremuestreo que genera ejemplos sintéticos de la clase minoritaria. (todos los productos diferentes a moto gamma baja)

Modelos de Clasificación

A continuación se detalla una descripción de los algoritmos empleados para predecir la mejor oferta. Estos algoritmo nos brindan la facilidad de trabajar con una variable respuesta múltiple y nos ayudan de la siguiente manera:

Gradiant Boosting: Este algoritmo capaz de capturar interacciones no lineales y relaciones complejas entre las variables, lo que lo convierte en una herramienta poderosa para predecir las preferencias de los clientes. Además, es resistente al sobreajuste y puede manejar datos ruidosos y con valores faltantes.

Random Forest: Random Forest es altamente flexible y puede adaptarse a una amplia variedad de problemas, incluidos aquellos con múltiples clases y características. Este algoritmo también es resistente al sobreajuste y puede manejar datos ruidosos, faltantes y desbalanceados, lo que lo convierte en una opción robusta para predecir las necesidades y preferencias de los clientes.

One vs Rest Clasiffier: Este algoritmo es una técnica de clasificación multiclase que entrena un clasificador binario para cada clase en el conjunto de datos. En el caso de este proyecto se entrenaría un modelo de regresión logística para cada producto, comparando cada producto con todos los demás. El One vs Rest Clasiffier es especialmente útil cuando se trabaja con variables respuesta compuestas por varias características, como es en este caso, ya que permite predecir de forma eficiente los múltiples productos y las afinidades con los clientes.

TABLA IV MODELOS DE CLASIFICACION

Método Empleado / formulación importante	Características relevantes
Gradiant Boosting [4]	
Optimización numerica	
$P^* = \sum_m^M Pm$	P* es la estimación inicial
Step Descent	
$gm = \left\{\left[\frac{\partial \phi P}{\partial P_I}\right]_{P=p_{m-1}}\right\}$	-g_m es llamado a definir el gradiente “steep descent”
$\rho m = argmin\Phi\left(\boldsymbol{P}_{m-1} - \boldsymbol{P}_{gm}\right)$	$\{Pm\}_1^M$ Son los sucesivos incremento o Boots
Algoritmo 1 : gradiant boosting	
$F_0(x) = \text{argmin}_\rho \sum_{i=1}^{N} L(y_i, \rho)$	Algoritmo de gradiant Boosting , con una optimización inicial , con incrementode boots y utilizando steep descent

For m= 1 to M do:

$$\tilde{y}_i = -\left[\frac{\partial L\left(y_i, F(x_i)\right)}{\partial F(x_i)}\right]_{F(x) = F_{m-1}(x)}, i = 1, N$$

$$a_m = argmin_{a,\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(x_i; a_m)]^2$$

$$\rho_m = argmin_{\rho} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_1) + \rho h(x_i; a_m))$$

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; a_m)$$

End For

End algorithm

Random Forest [5]

random forest convergencia

$$mg(\boldsymbol{X},\boldsymbol{Y}) =_{avk} I(h_k(\boldsymbol{X}) = \boldsymbol{Y}) - \max_{j \neq Y} {}_{avk} I(h_k(\boldsymbol{X}) = j)$$

Donde $h_1(\mathbf{X})$, $h_2(\mathbf{X})$ …$h_k(\mathbf{X})$ son un conjunto de clasificadores y la función I(.) es una función indicadora.

Índice de Gini (Arboles de decisión)

$$Gini(t) = 1 - \sum p_i^2$$

Gini(t) es el índice del nodo a particionar y P_i es la proporción de muestras de una clase para un nodo especifico de t.

Particiona de forma recursiva el conjunto de datos para clasificar. Cada $h_k(\mathbf{X})$ que luego se evalúa en la convergencia del Random forest.

One Vs Rest Clasiffier [6]

Función de minimización de la función de perdida

$$f(x) = \frac{(p(\boldsymbol{x}))^{\frac{1}{q-1}} - (1-p(x))^{\frac{1}{q-1}}}{(p(\boldsymbol{x}))^{\frac{1}{q-1}} - (1-p(x))^{\frac{1}{q-1}}}$$

P(x) = la probabilidad de un localizado en x este en la clase 1.

f(x) = es la función de minimización de la función de pérdida $(1-y*f(\mathbf{x}))^q$ o $|y-f(\mathbf{x})|^q$

One vs Rest Clasiffier

$$f(x) = arg \underset{r \in 1,\ldots,N}{\sin in} \sum_{i=1}^{F} \left(\frac{1 - sign(M_{ri}\, f_i(x)}{2}\right)$$

N= Es el número de clases

F = El número de clasificadores binarios entrenados

Métricas de desempeño

Es importante definir métricas de desempeño y en conjunto con el entendimiento de negocio, medir que tan efectiva es la predicción de los modelos. Las métricas de desempeño que utilizaremos en nuestro modelo son las siguiente:

TP = True Positives (verdaderos Positivos)

TN =True Negatives (Verdaderos negativos)

FP = False Positives (Falsos positivos)

FN = False Negatives (Falsos Negativos)

- **Precisión:** La Precisión ayudara Evaluar que también puede predecir el modelo para los clientes

$$\text{Precisión} = \frac{TP}{TP+FP} \quad (1) \quad [14]$$

- **Recall:** Con el recall se medirá la capacidad para predecir con precisión los productos que los clientes pueden preferir, así como su habilidad para identificar a los clientes interesados en cada producto.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2) \; [14]$$

- **Accuracy :** EL accuracy medirá el porcentaje de casos que el modelo ha acertado.

$$\text{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN} \quad (3)\ [18]$$

- **F1_score:** con esta métrica evaluaremos el rendimiento del modelo en términos de su capacidad para identificar correctamente las diferentes características de los clientes de los diferentes productos y asi evitar en lo posible los falsos positivos y falsos negativos.

$$\text{F1\_score} = 2* \frac{presicion*Recall}{Precision+Recall} \quad (4) \quad [2]$$

- **Curva ROC :** Con la curva ROC podremos graficar la capacidad de los modelos para discriminar los productos a medida que varía el umbral de clasificación. [2]

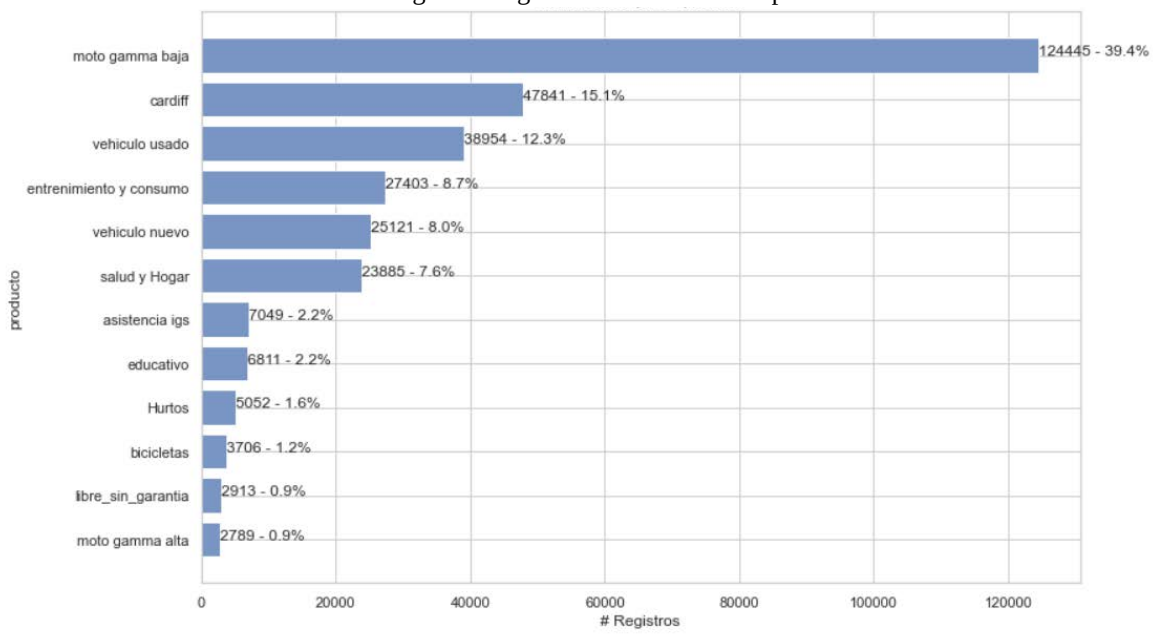
- **AUC (Área bajo la curva) :** El AUC es el área bajo la curva ROC y con ella evaluaremos el rendimiento de los modelo independientemente del umbral de clasificación. [2]

VIII RESULTADOS

Exploración de los datos

Se realiza primero un análisis de la Variable Respuesta , y la distribución de cada una de las características, las cuales representan el comportamiento de elección de los clientes en sus diferentes producto y que además me dan una idea el pre-procesamiento y e ingeniería de características necesarios para el desarrollo de un modelo optimo. Luego se analizan otras variables que son consideradas importantes para el negocio.

Fig. 1 Histograma de la variable respuesta



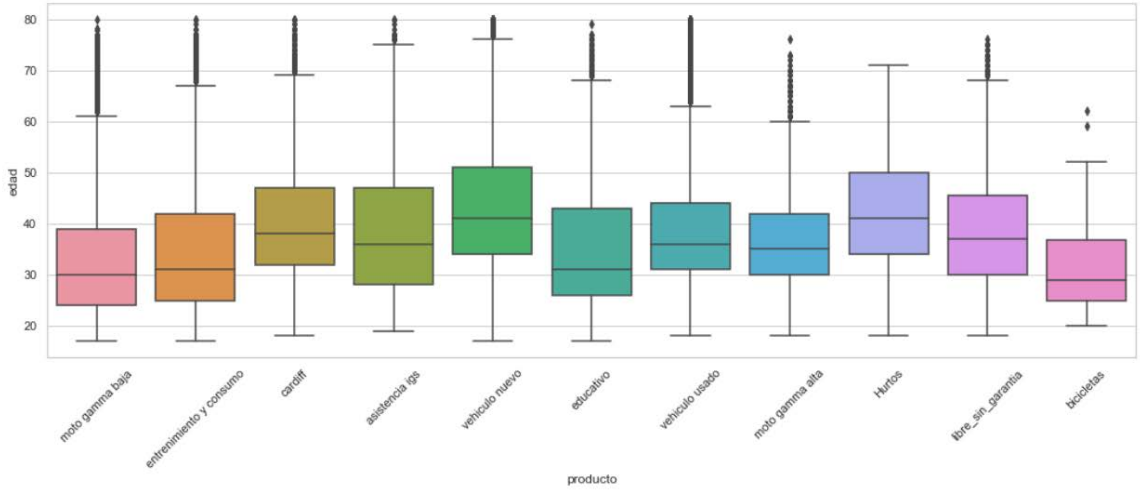
En la Fig. 1 Histograma variable respuesta podemos observar que algunos de los productos tienen mucha más información que otros , en el caso de las motos gamma baja , representa un 39.4% del

total de los datos , mientras que para las motos gamma baja , representan un 0.9% del total de los datos , esto a nivel de negocio representa que la mayoría de los clientes suelen adquirir un crédito de moto gamma baja , mientras que una minoría adquiere un crédito de motos gamma alta o un libre sin garantía.

1) Edad de los clientes

Por medio de un Boxplot se analiza la edad de los clientes , la cantidad de tarjetas activas antes de adquirir algunos de los producto y el promedio de ingreso mensuales de T0—T-3 antes de adquirir alguno de los productos

Fig. 2. edad vs producto adquirido



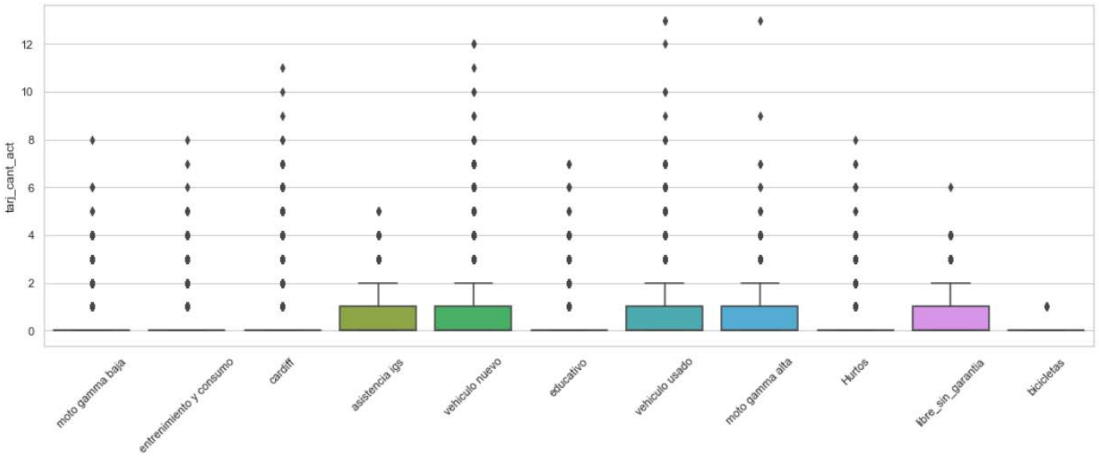
En la Fig. 2 edad vs producto adquirido podemos observar que el producto bicicletas es el producto escogido por los más jóvenes mientras que los vehículo nuevos y los hurtos son adquiridos por las personas en promedio con la edad mayor a os 40 años , también es importante observar que los productos de motos gamma baja y educativo en promedio tiene una edad de adquisición entre los 30 años , por otro lado también los vehículos usados muestran que promedio los clientes adquieren este producto alrededor de los 35 años y se nota una marcada diferencia

con el cliente de vehículo nuevo , el cual es un cliente más experimentado y de mayor adquisición y suele adquirirlo en una mayor edad.

2) Cantidad de tarjetas Activas

Para la cantidad de tarjetas activas (cualquier tipo de tarjeta activa con la entidad) un cliente antes de adquirir un producto con la empresa , en promedio para los productos de moto gamma baja , entretenimiento y consumo , Cardiff, educativo, hurtos y bicicletas los clientes no tienen tarjetas activas , mientras que en promedio antes de adquirir productos como libre sin garantía, asistencia Igs , vehículo nuevo, vehículo usado, motos gamma alta , los clientes tenía al menos entre 1 y 2 tarjetas activas. En algunos casos particulares los clientes pueden tener hasta 12 tarjetas activas como es el caso de los vehículos usados y nuevos.

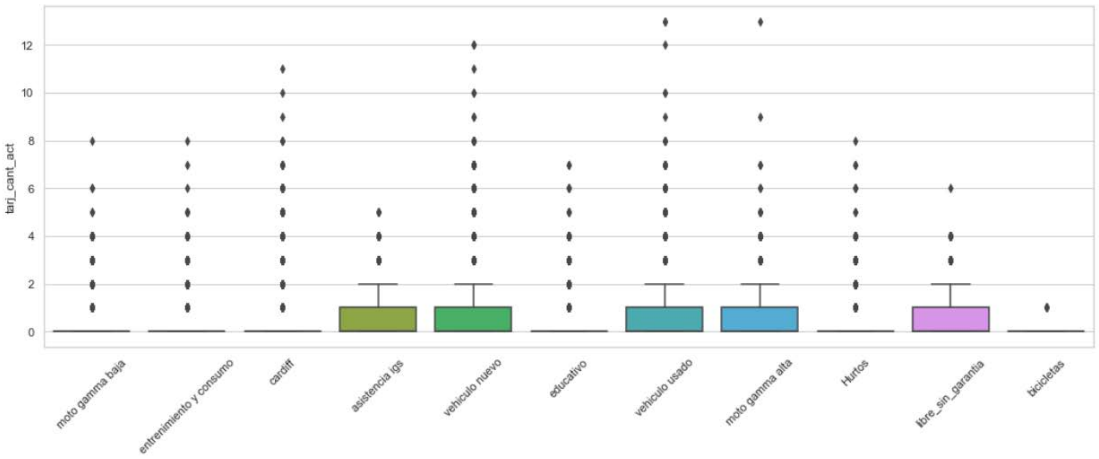
Fig. 3. Producto vs cantidad de tarjetas activas



3) Ingresos promedio mensuales

En el caso del promedio ingreso de los clientes antes de adquirir alguno de los producto en la Fig. 4 Productos vs ingresos mensuales, es claro que el promedio de ingreso de los clientes que adquieren un vehículo nuevo o usado , una moto gamma alta o un libre inversión , es muy superior a los demás productos , en general esto tiene mucho sentido , ya que adquirir un crédito para alguno de estos productos requiere una alta capacidad de endeudamiento y de ingresos , caso contrario de las motos gamma baja o entretenimiento y consumo.

Fig. 4. Productos vs ingresos promedio mensuales



Preprocesamiento de los datos

En la TABLA V se observa las variables que tienen datos nulos, todas estas variable a excepción de la fecha de desembolso, son variables categóricas.

TABLA V PORCENTAJE DE DATOS NULOS

Variable	% porcentaje nulo
Actividad económica	47%
Afinidad trans	47%
Fecha desembolso antes	42%
Última G	16%

segmento	1%

Con el conocimiento del negocio las variables Actividad económica , Afinidad transaccional, Ultima G y segmento se imputarán con la categoría No Informa. En el caso de la fecha de desembolso , no es indispensable en la construcción del modelo, por lo que se tomó la decisión de eliminarla.

1) . *Estandarización*

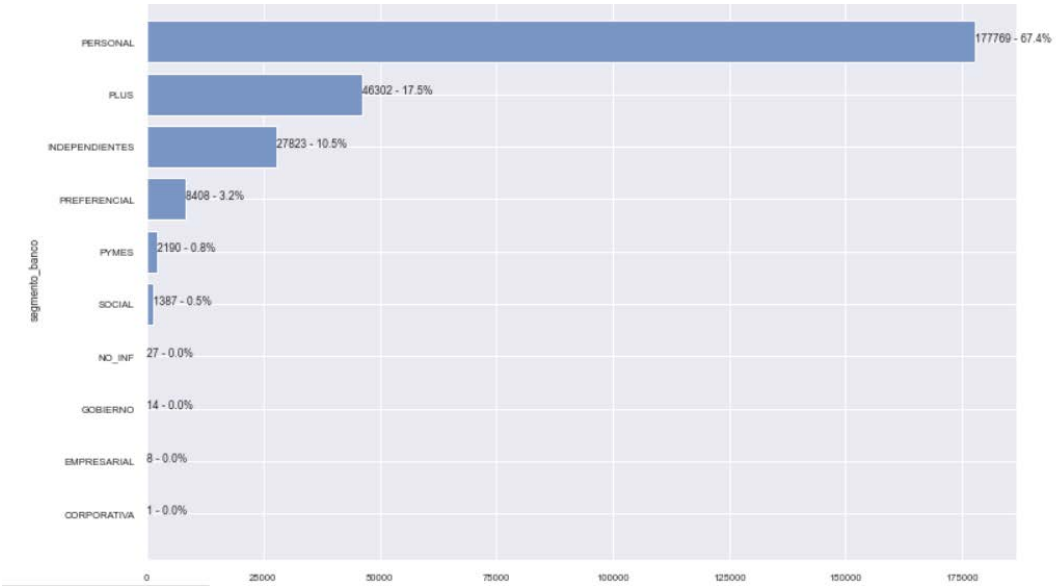
Para que las variables numéricas tengan una escala similar , se hicieron dos procedimientos: el primero consiste que a las variables correspondiente a los desembolsos, saldos e ingresos mensuales del cliente se dividieron por 1.300.000 , el cual es un aproximado del salario mínimo legal vigente , esto con el objetivo de que esta variable estas variables con valores muy gigantes se representen en una escala más ajustada a los demás datos.

Posteriormente se escalizaron todas las variables con la metodología de min - max escaler en un rango de [0-1]

2) *Conversión de variables categóricas*

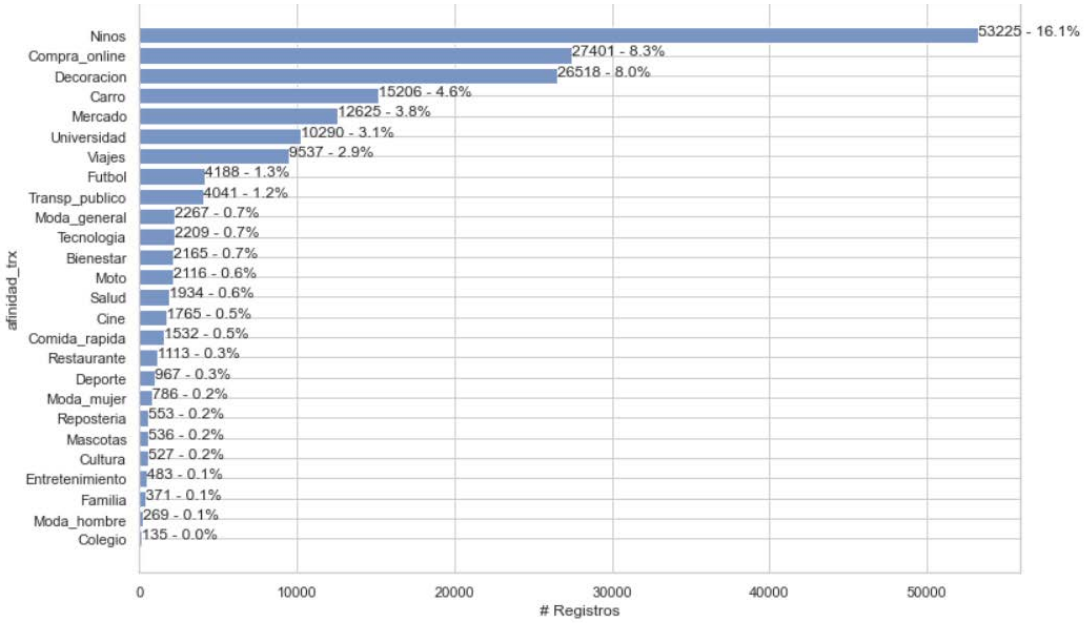
En la Fig. 5 Histograma segmento banco podemos visualizar que la variable segmento en las características Personal, Plus, Independientes y preferencial componen casi en la totalidad las variable segmento , por lo tanto el proceso que se realizó consiste , en aplicarles a las otras características unas transformación y agrupación en la otra característica llamada **“otras”**

Fig. 5. Histograma segmento banco



Otro ejemplo es la en la cual podemos visualizar las características de la variables afinidad transaccional en donde paras la características Niños, compra online ,decoración , carro , mercado universidad y viajes, las cuales representan casi el 95% de toda la información de la variable, no se transforman, mientras las demás características de la variable se transforman y se agrupan en otra característica llamada **“otras”**

Fig. 6. Histograma afinidad transaccional



Al realizar este proceso con las demás variables categóricas disminuimos la cantidad de variables que se generan al aplicar el método de one-hot-encoding sin perder la información que brindan en el momento de realizar el modelo predictivo , y disminuimos el riesgo de sufrir la maldición de la dimensionalidad , ya que más variables mucho más dispersión entre ellas. Como ejemplo si se dejan todas las categorías por cada característica de la variable afinidad transaccional y el segmento banco, pasaríamos 87 variables tener 124 variables , mientras que de esta manera pasamos de 87 variables a 97 variables.

Ingeniería de características

1) *correlación de spearman*

Se eliminan todas las variables que tengan un coeficiente de correlación de spearman mayor >0.8 y menor -0.8, lo cual elimina en total unas 29 variables con este valor.

2) *Analisis de datos atípicos y entropía*

Se utiliza la técnica de Local outlier factor(LOF) con la métrica de Manhatthan para hallar datos atípicos y se encontró que el 14% de las filas tienen datos tienen datos atípicos por lo que se tomó la decisión de eliminarlos , posterior a esto calculamos la entropía [15] de las variables antes y después de eliminar los datos atípicos, esto nos ayuda a evaluar el impacto de la eliminación de datos atípicos en la información contenida en el conjunto de datos. Si la entropía no disminuye significativamente después de eliminar los datos atípicos, esto indica que la pérdida de información es mínima y que la calidad de los datos ha mejorado.

Como resultado solo hubo una variable en la cual se encontró una disminución muy significativa.

TABLA VI RESULTADO ENTROPIA VALOR AVALUO

Variable	Entropía original	Entropía LOF
Valor avaluó	0.00185	0.0095

Se hablo con el negocio de este resultado y de la importancia de la variable para ellos , y se consideró no trabajar con esta variable.

Bases de entrenamiento y testeo

El resultado final de las variables durante el preprocesamiento e ingeniería de características son el punto inicial para la creación de las bases de entrenamiento y testeo, asi se construye la base de entrenamiento y testeo.

TABLA VII BASES DE ENTRENAMIENTO TESTEO

Categorías de las ofertas	observaciones	Muestra 70%	Muestra del 70% con Balanceo oversampling	Muestra del 70% con Balanceo Smote	Testeo 30%
Moto gama baja	104983	70525	70525	70525	34458
Cardiff	39570	26534	70525	70525	13036
Vehículo Usado	32409	21648	70525	70525	10761
Entretenimiento y consumo	23418	15657	70525	70525	7761
Salud y Hogar	21012	14103	70525	70525	6909
Vehículo Nuevo	19856	13322	70525	70525	6534
Educativo	5697	3797	70525	70525	1900
Asistencia Igs	5171	3459	70525	70525	1712

Hurtos	4151	2756	70525	70525	1395
Bicicletas	3314	2180	70525	70525	1134
Libre sin Garantía	2210	1453	70525	70525	757
Motos gama alta	2789	1398	70525	70525	740

Modelamiento de la data

En el TABLA VII se observa cómo queda la distribución de la base de entrenamiento de cada categoría de la variable respuesta sin balancear , con la técnica de balanceo oversampling y la técnica de balanceo Smote. Posteriormente se entrenan múltiples modelos de gradiant Boosting , random forest y one vs rest clasiffier con las bases de entrenamientos balanceadas y sin balancear y se elige el mejor modelo con respecto a las métricas propuestas.

TABLA VIII METRICAS DE EVALUACION DE LOS MODELOS DE CLASIFICACION

Modelo / Métricas	Precisión	Recall	Accuracy	F1_score	AUC
Gradiant Boosting sin balancear	0.42	0.40	0.42	0.41	0.64
Gradiant Boosting con balanceo Oversampling	0.62	0.6	0.62	0.61	0.82
Gradiant Boosting con balanceo Smote	0.61	0.64	0.63	0.62	0.83
Random forest sin balancear	0.44	0.44	0.42	0.44	0.68
Random forest con balanceo oversampling	0.64	0.62	0.63	0.64	0.87

Random forest con balanceo smote	0.66	0.66	0.65	0.66	0.89
One vs Rest Classifier sin balancear	0.42	0.41	0.43	0.42	0.64
One vs Rest Classifier con balanceo oversampling	0.63	0.62	0.63	0.63	0.83
One vs Rest Classifier con balanceo smote	0.64	0.64	0.63	0.64	0.83

De la TABLA VIII se observa que hay una diferencia positiva en los modelos entrenados que previamente ya están balanceados , ya sea con la técnica de oversampling o Smote , adicionalmente hay una similiraridad entre los resultados de la métricas entre los tres modelo , siendo mejor el random forest con balanceo Smote por muy poco a los demás modelos.

El siguiente paso es analizar las métricas y el comportamiento de las ofertas con la matriz de confusión en cada una de las categorías de la variable respuesta en el modelo entrenado random forest con balanceo Smote. *"Los posteriores analisis se basan en este modelo"*.

TABLA IX METRICAS DEL RANDOM FOREST CON BALANCEO SMOTE

Categorías de las ofertas	Precisión	Recall	F1_score	AUC
Moto gama baja	0.87	0.89	0.88	0.96
Cardiff	0.50	0.51	0.51	0.88
Vehículo Usado	0.49	0.57	0.53	0.89
Entretenimiento y consumo	0.48	0.49	0.49	0.92
Vehículo Nuevo	0.56	0.57	0.57	0.93
Salud y Hogar	0.44	0.42	0.43	0.90
Educativo	0.50	0.40	0.45	0.93
Asistencia Igs	0.56	0.49	0.52	0.91

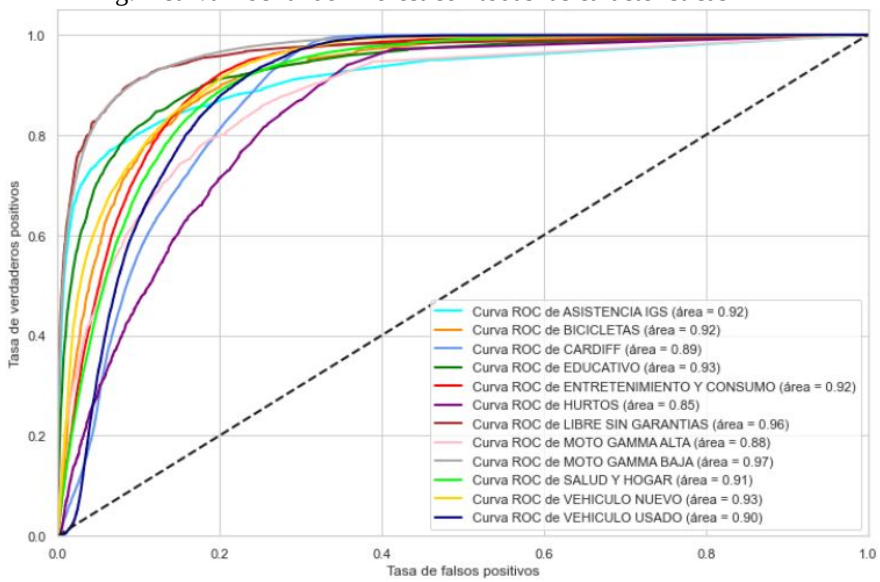
Hurtos	0.05	0.02	0.03	0.85
Bicicletas	0.31	0.16	0.21	0.91
Libre sin Garantía	0.6	0.36	0.45	0.96
Motos gama alta	0.21	0.10	0.16	0.87

En la TABLA IX se observa la medición individual de las métricas , precisión , recall y F1-score de cada una de las características (productos de la empresa) de la variable respuesta. Podemos ver que el producto de moto gama baja es la característica que mejores resultados en sus métricas tiene , mientras que hurtos y motos gama baja por el contario muestran unos resultados muy bajos las métricas , es decir el modelo no logra predecir correctamente una oferta de este productos para los clientes.

En la Fig. 7 curva Roc random forest con todas las características se observa que en general las características tienen una área bajo la curva mayor o igual a 0.85 , siendo hurtos la de menor valor con un área bajo la curva de 0.85, es decir el modelo tiene una capacidad muy alta en todas las ofertas de discriminarlas entre quienes son afines a ellas y quienes no. Se observo que en la TABLA IX algunas de las ofertas tienes una métricas de medición muy bajas , lo que significa que aunque el modelo discrimine bien, está siendo muy conservador o liberal a la hora de clasificar , por lo que es una muy buena opción quizás cambiar los umbrales de predicción.

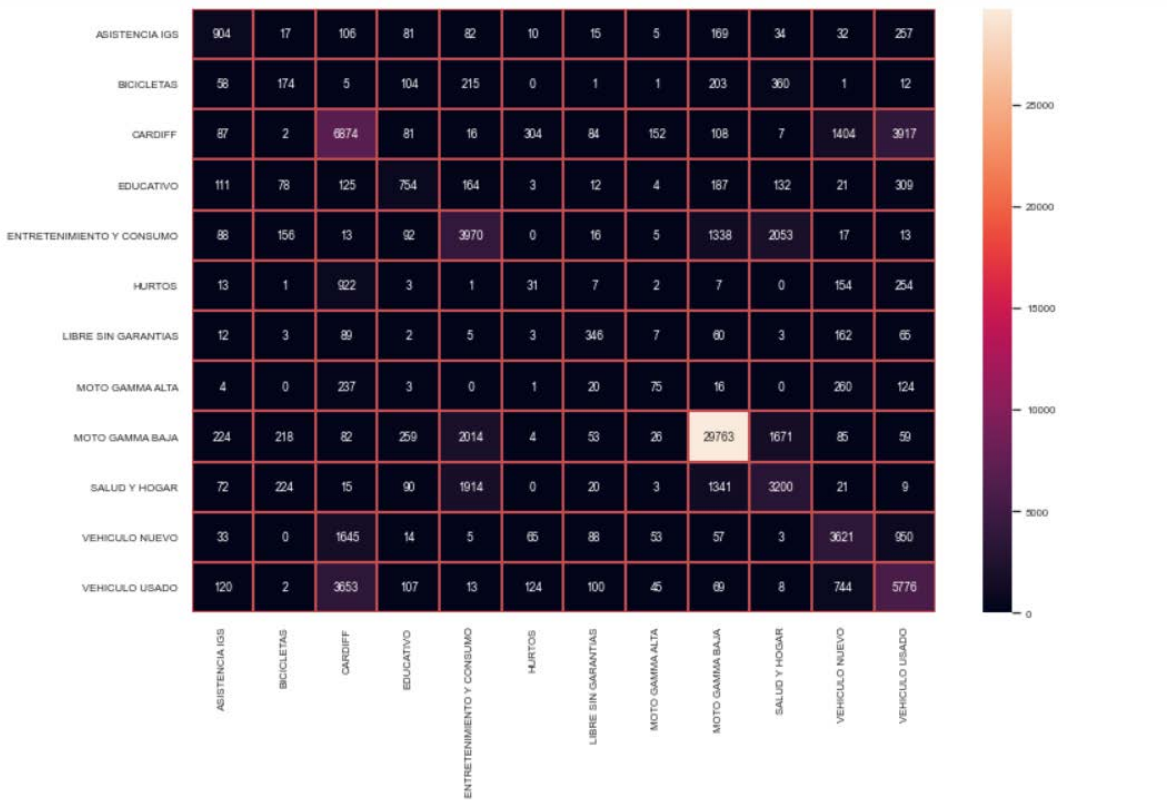
El Modelo se entrenó con diferentes semillas , y cada uno de los entrenamientos que se realizó no modifico significativamente la métricas de evaluación , lo que indica que el modelo es robuzto y consistente a pesar de los diferentes conjuntos de entrenamiento y Prueba.

Fig. 7 curva Roc random forest con todas las características



En la Fig. 8 . Matriz de confusion Random forest con balanceo smote se pueden observar que tan bueno ha sido el desempeño del modelo en cada característica.

Fig. 8 . Matriz de confusion Random forest con balanceo smote



Del grafico se destacan los siguientes aspectos :

- Los clientes que adquirieron un producto de Hurtos los está clasificándolo como clientes de Cardiff y estos a su vez están siendo clasificados en un gran porcentaje como clientes que adquieren un vehículo nuevo o usado.
- Una gran cantidad de clientes de salud y hogar se están clasificando como clientes de entretenimiento y consumo y también un buen número de estos se clasifican como clientes de motos gama baja.
- La mayoría de los clientes de motos gama baja se están clasificando como clientes de vehículos usado y vehículos nuevos.
- Los clientes de bicicletas se están clasificando como clientes de motos gama baja , entretenimiento y consumo y salud y hogar

En base a los anteriores descubrimientos se realiza un analisis de los resultados con la realidad del negocio y se encontró que un 100% de los clientes que tienen un producto hurtos o Cardiff ya adquirieron una oferta de vehículo, ya sea usado o nuevos; por lo que se excluyen del modelo.

Para el negocio , los productos de entretenimiento y consumo y salud y Hogar están realmente dentro de una misma categoría por lo que se considera mejor incluir todos los clientes de salud y hogar dentro del producto de entretenimiento y consumo.

Los créditos del producto de bicicleta que se realizan son muy similares a los créditos de moto gama baja , por lo que decide desde el negocio no tenerlos en cuenta en este modelo , y se plantea otras soluciones más logísticas y empíricas desde el conocimiento de los administradores del producto.

Con estas anteriores decisiones que se tomaron en base al analisis del resultado del modelo y al conocimiento del negocio de la empresa se volvió a entrenar el modelo.

En la TABLA X se observa como al eliminar las características de hurtos, Cardiff, la métrica de los productos de vehículos usados y vehículos nuevos mejoran con respecto a las -

TABLA X METRICAS DE EVALUACION 2 ITERACION MODELO

Categorías de las ofertas	Precisión	Recall	F1_score	AUC
Moto gama baja	0.88	0.83	0.85	0.94
Vehículo Usado	0.62	0.67	0.65	0.95
Entretenimiento y consumo	0.69	0.75	0.72	0.92
Vehículo Nuevo	0.61	0.63	0.62	0.95
Educativo	0.32	0.34	0.34	0.87
Asistencia Igs	0.52	0.54	0.54	0.93
Libre sin Garantía	0.54	0.57	0.55	0.98
Motos gama alta	0.41	0.14	0.20	0.90

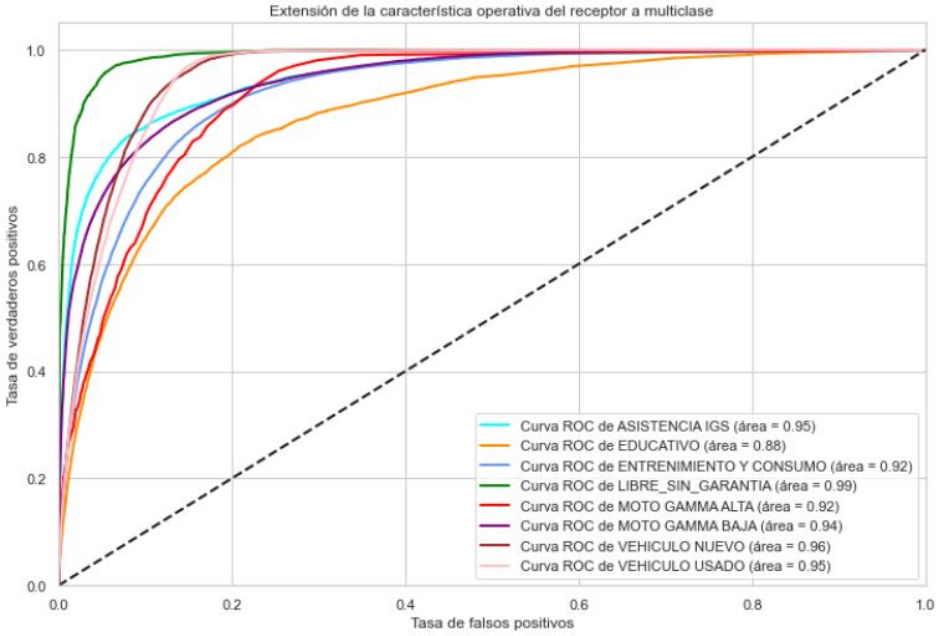
TABLA XI METRICAS DE EVALUACION GENERAL 2 ITERACION

Modelo /Métricas	Precisión	Recall	F1_score	AUC
Random Forest con balance smote	0.74	0.73	0.74	0.91

métricas de la TABLA IX; al incluir dentro de una misma categoría de los productos de entretenimiento y consumo y Salud y Hogar , las cuales mejoran de manera muy positiva sus

métricas. Fig. 9 . curva Roc random forest sin las características eliminadas se observa que igual que el modelo entrenado con todas las características , este modelo tiene unas medidas de AUC en cada una de las ofertas muy altas , es decir tiene una muy alta capacidad de discriminar cuales clientes son más afines a las ofertas y cuáles no.

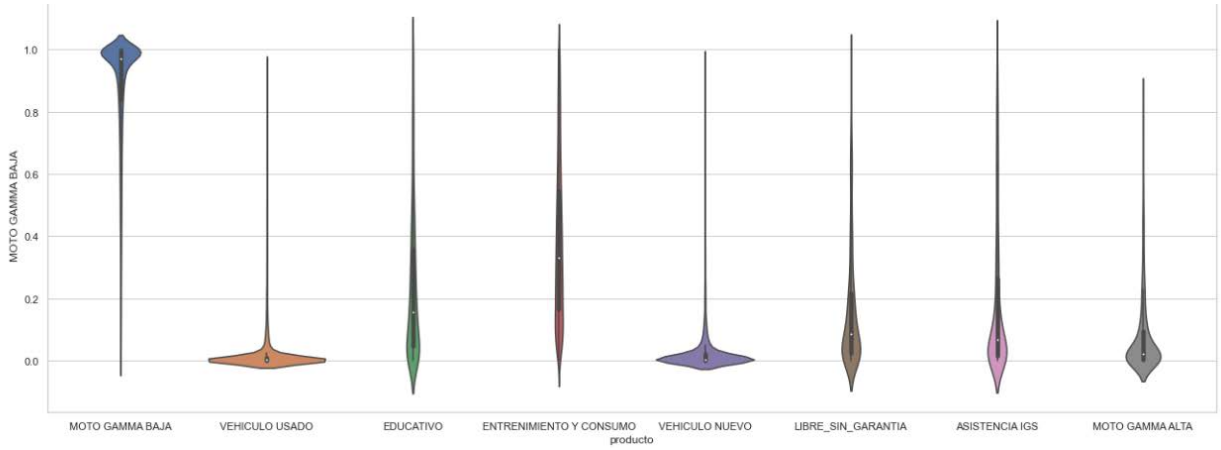
Fig. 9 . curva Roc random forest sin las características eliminadas



De los resultados en todos los modelo realizados balanceando la data entrenamiento han tenido algo importante en común, es que las métricas de precisión , recall , F1 tienen una diferencia marcada con la medida del AUC , tanto como a nivel general del modelo , como a nivel individual de cada una de las características de la variable respuesta. La discrepancia entre estas métricas podría deberse a varios factores, entre ellos, el desequilibrio de las clases de la variable repuesta , o la elección del umbral de decisión para la clasificación. Por lo tanto se realiza un analisis de los umbrales de clasificación de cada una de las características del modelo por medio de graficas de violines.

En la Fig. 10. Gráfico de violines Moto Gama Bajase visualiza la distribución de las probabilidades en donde cada cliente de la categoría de motos gama se está clasificando en cada uno de los demás productos, por ejemplo los clientes de moto gama baja que se clasifican correctamente dentro de su categoría suelen hacerlo con probabilidades entre 0.8 y 1, mientras que aquellos que siendo moto gama baja el modelo los clasifica incorrectamente como motos gama alta sus probabilidades están en su mayoría dentro del rango de 0-0.2. El resto de los gráficos de violines de las demás variables se pueden visualizar en la zona de anexos.

Fig. 10. Gráfico de violines Moto Gama Baja



Los diferentes gráficos de violines dan una visual de los diferentes umbrales de clasificación de cada clase de la variables respuesta , con la cual se minimiza un poco el número de iteraciones en la escogencia de umbrales de clasificación que mejoren el rendimiento el modelo.

Después de escoger diferentes umbrales de predicción se obtienen los que mejores optimizan el modelo. Estos se visualizan en la TABLA XII

TABLA XII UMBRALES DE PREDICCION

Clase /Umbral de clasificación	Probabilidad
Asistencia Igs	0.15
Educativo	0.25
Entretenimiento y consumo	0.3

Libre Sin Garantía	0.2
Moto Gama Alta	0.15
Moto Gama Baja	0.75
Vehículo Nuevo	0.5
Vehículo Usado	0.5

TABLA XIII METRICAS DE EVALUACION CON UMBRALES DE PREDICCION

Categorías de las ofertas	Precisión	Recall	F1_score	AUC
Moto gama baja	0.96	0.88	0.30	0.94
Vehículo Usado	0.76	0.83	0.79	0.95
Entretenimiento y consumo	0.64	0.78	0.70	0.92
Vehículo Nuevo	0.78	0.63	0.70	0.95
Educativo	0.56	0.43	0.51	0.87
Asistencia Igs	0.20	0.65	0.30	0.93
Libre sin Garantía	0.58	0.47	0.52	0.98
Motos gama alta	0.31	0.28	0.30	0.90

TABLA XIV METRICAS GENERAL DEL MODELO CON UMBRALES DE PREDICCION

Modelo /Métricas	Precisión	Recall	F1_score	Accuracy	AUC
Random Forest con balance smote	0.83	0.77	0.79	0.77	0.91

De la TABLA XIII se observa que las métricas precisión , recall, f1-score en todas las clases a excepción de la clase asistencias igs , esto es debido a que las probabilidades en donde cada cliente del producto asistencias igs se está clasificando en cada uno de los demás productos

de manera muy uniforme entre 0-1 , es decir no importa el umbral probabilidad escogido , este no va a discriminar adecuadamente entre los que pertenecen a la clase y los que no pertenecen a la clase.

El grafico de violín de asistencia igs se puede visualizar en anexos.

Variables más representativas del Modelo

Las variables más importantes en la iteraciones y cambios que se hicieron en los modelos son principalmente 7 variables y el porcentaje de aporte que hace cada una para explicar la mejor oferta son las siguientes:

TABLA XV VARIABLES MAS REPRESENTATIVAS

Variables Representativa	Porcentaje de explicación
Promedio del Valor desembolsado	23%
Promedio ingresos mensuales	12%
Edad	9%
Numero obligaciones	8%
Promedio saldo tarjetas activas	7%
Promedio de Días de mora	6%
Cantidad histórica de seguros	5%

Cómo se puede observar en la TABLA XV, estas variables representan el 70% del poder predictivo para predecir el la variable objetivo del modelo, es decir que el resto de variables aunque

en menores rango , representan el 30%. Por lo tanto se realizó un ejercicio en donde se eliminaron una , dos , tres , cuatros y hasta 5 variables menos importantes, con lo cual se entrenaron 5 diferentes modelos con las diferentes eliminaciones; los resultado obtenidos son que las métricas de medición del modelo (precisión, recall, F1-score y accuracy) se ven disminuidas en un 1 o dos puntos a los máximo , lo que indica que aunque son de menor importancia tienen una influencia en el resultado final.

X. CONCLUSIONES

1. Se puede observar que las variables como el 'Promedio del Valor desembolsado', 'Promedio de ingresos mensuales' y 'Edad' son las más determinantes en la decisión del modelo. Esto puede interpretarse como una indicación de que los clientes que gastan más tienen mayores ingresos y son de cierta edad, están más dispuestos a aceptar ofertas de productos financieros.

2. el 30% del poder predictivo está distribuido en un número de variables más amplio, todavía influyen en el resultado final del modelo, aunque a un grado menor. La eliminación de estas variables menos importantes en distintos grados solo disminuyó ligeramente las métricas del modelo, lo que indica que su contribución, aunque pequeña, es significativa para la predicción final.

3. La efectividad de los algoritmos de aprendizaje automático, en particular el Gradiant Boosting, Random Forest y One vs Rest Classifier, fue evaluada en este trabajo. Los resultados mostraron que Random Forest con balanceo Smote obtuvo las mejores métricas de rendimiento.

4. El ajuste de los umbrales de predicción para cada categoría mejoro significativamente la precisión de las predicciones a excepción de los productos de igs; esto a nivel de negocio es muy positivo ya que para el negocio es mas costoso equivocarse en un vehículo nuevo o los demás productos que equivocarse en un producto de igs , el cual es el producto de menor costo en la empresa.

5. Este trabajo ha demostrado que es posible construir un modelo predictivo que puede identificar el próximo producto de financiación más afín a las necesidades de los clientes de una entidad crediticia. Aunque no todos los productos tienen las mejores predicciones como el caso de las motos gama baja , el modelo se ajusta adecuadamente a las necesidades de la empresa para un primer lanzamiento y desarrollo, en el que posteriormente a medida que haya más iteraciones y resultados ofrecidos por el modelo , los productos como las moto gama baja pueden ir mejorando su efectividad.

6. El estudio también evidenció la importancia de una cuidadosa selección y gestión de las características en el modelado predictivo. La eliminación de algunas categorías con pocos datos mejoró las métricas del modelo, lo que enfatiza la necesidad de conectar adecuadamente la analítica y el analisis de los datos con la realidad del negocio.

XI. RECOMENDACIONES

El modelo debe ser monitoreado de forma continua para asegurar que sigue siendo efectivo a medida que cambian los comportamientos de los clientes y las condiciones del mercado, los cuales también influyen en los umbrales de decisión implementados para cada categoría , que aunque han demostrado ser efectivos en mejorar el rendimiento del modelo, estos umbrales necesitan ser revisados y actualizados para reflejar estos cambios.

Los resultados han demostrado que es posible predecir la oferta de producto que un cliente probablemente encontrará más atractiva. Con base en estos hallazgos, sería útil explorar aún más la personalización, tal vez al nivel individual del cliente, para mejorar aún más la satisfacción del cliente y los resultados comerciales. Adicionalmente seria mas productivo en un futuro hacer analisis y exploración de los datos de los clientes de manera mas online, asi la entidad crediticia podría considerar la implementación de sistemas que tengan en cuenta la secuencia de las ofertas y la cronología de las interacciones de los clientes.

Siguiendo con la misma línea del pensamiento online , otra recomendación a considerar es hacer retroalimentaciones del modelo de manera online , es decir se podría considerara por ejemplo los clics del usuario, las visitas a la página del producto, y las interacciones recientes con la plataforma, para ajustar las recomendaciones y predicciones; también se podría analizar características mas complejas como la interacción del cliente con otras páginas o búsquedas que haga de productos similares a los de la empresa, aunque para esto también es importante tener en cuenta el costo de las bases de datos que tengan este tipo de variables.

REFERENCIAS

[1] H. R. Zhang, F. Min, X. He, y Y. Y. Xu, “A hybrid recommender system based on user-recommender interaction”, *Math. Probl. Eng.*, vol. 2015, nº 1, 2015.

[2] K. P. Murphy, *Machine Learning*. Cambridge, Massachusetts, 2012.

[3] M. M. Breunig, H. P. Kriegel, R. T. Ng, y J. Sander, “LOF: Identifying Density-Based Local Outliers”, *SIGMOD 2000 - Proc. 2000 ACM SIGMOD Int. Conf. Manag. Data*, nº March 2014, pp. 93–104, 2000.

[4] J. Friedman, “Greedy Function Approximation:A Gradient Boosting Machine”, *IMS*, 2001.

[5] J. Friedman, “Random Forest”, *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[6] R. Rifkin, “In Defense of One-Vs-All Classification In Defense of One-Vs-All Classification”, nº June 2014, 2004.

[7] D. Goldberg, D. Nichols, B. M. Oki, y D. Terry, “Using collaborative filtering to Weave an Information tapestry”, *Commun. ACM*, vol. 35, nº 12, pp. 61–70, 1992.

[8] P. Resnick, P. Bergstrom, y J. Riedl, “GroupLens : An Open Architecture for Collaborative Filtering of Netnews”, pp. 175–186, 1994.

[9] B. Sarwar, “Item-based Collaborative Filtering Recommendation Algorithms Item-Based Collaborative Filtering Recommendation Algorithms”, nº August, 2001.

[10] andD. J. Quadrana, M., P. Cremonesi, “Sequence-Aware Recommender Systems.””, *ACM Comput. Surv.*, pp. 1–36., 2018.

[11] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, y J. A. Konstan, “User perception of differences in recommender algorithms”, *RecSys 2014 - Proc. 8th ACM Conf. Recomm. Syst.*, pp. 161–168, 2014.

[12] K. L. Wagstaff, “Machine learning that matters”, *Proc. 29th Int. Conf. Mach. Learn. ICML 2012*, vol. 1, pp. 529–534, 2012.

[13] Z. Zhang, K. Niu, y Y. Liu, “A deep learning based online credit scoring model for P2P lending”, *IEEE Access*, vol. 8, pp. 177307–177317, 2020.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1ª ed. 2006.

[15] C. E. Shannon, “Technical .jo-urnal”, vol. XXXIII, nº March, pp. 799–826, 1948.

[16] Haibo He and Edwardo A. Garcia, “Learning from Imbalanced Data”, *IEEE Trans. Knowl. Data Eng.*, vol. 21, pp. 1263–1284, 2009.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, y W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique”, *J. Artif. Intell. Res.*, vol. 16, nº June, pp. 321–357, 2002.

[18] R. T. Trevor Hastie , Jerome Friedman, *The Elements of Statistical Learning*. 2009.

ANEXOS

TABLA XVI. METRICAS DE EVALUACION DEL GRADIAN BOOSTING CON BALANCEO SMOTE

Categorías de las ofertas	Precisión	Recall	F1_score	AUC
Moto gama baja	0.89	0.86	0.88	0.92
Cardiff	0.50	0.53	0.51	0.86
Vehículo Usado	0.49	0.54	0.51	0.85
Entretenimiento y consumo	0.47	0.51	0.49	0.89
Vehículo Nuevo	0.56	0.55	0.55	0.88
Salud y Hogar	0.43	0.44	0.45	0.90
Educativo	0.47	0.40	0.43	0.89
Asistencia Igs	0.52	0.53	0.53	0.90
Hurtos	0.06	0.02	0.03	0.84
Bicicletas	0.20	0.15	0.17	0.89
Libre sin Garantía	0.45	0.46	0.46	0.92
Motos gama alta	0.20	0.10	0.13	0.86

TABLA XVII. METRICAS DE EVALUACION DEL ONE VS REST CLASSIFIER CON BALANCEO SMOTE

Categorías de las ofertas	Precisión	Recall	F1_score	AUC
Moto gama baja	0.84	0.86	0.85	0.94
Cardiff	0.48	0.50	0.49	0.86
Vehículo Usado	0.52	0.56	0.54	0.88
Entretenimiento y consumo	0.43	0.48	0.45	0.90
Vehículo Nuevo	0.54	0.56	0.55	0.91
Salud y Hogar	0.46	0.45	0.45	0.90
Educativo	0.43	0.43	0.43	0.91
Asistencia Igs	0.49	0.53	0.51	0.88
Hurtos	0.06	0.03	0.04	0.82
Bicicletas	0.15	0.14	0.14	0.89
Libre sin Garantía	0.41	0.50	0.47	0.87
Motos gama alta	0.20	0.1	0.15	0.87

Fig. 11 . Gráfico de Violines Vehículo Nuevo

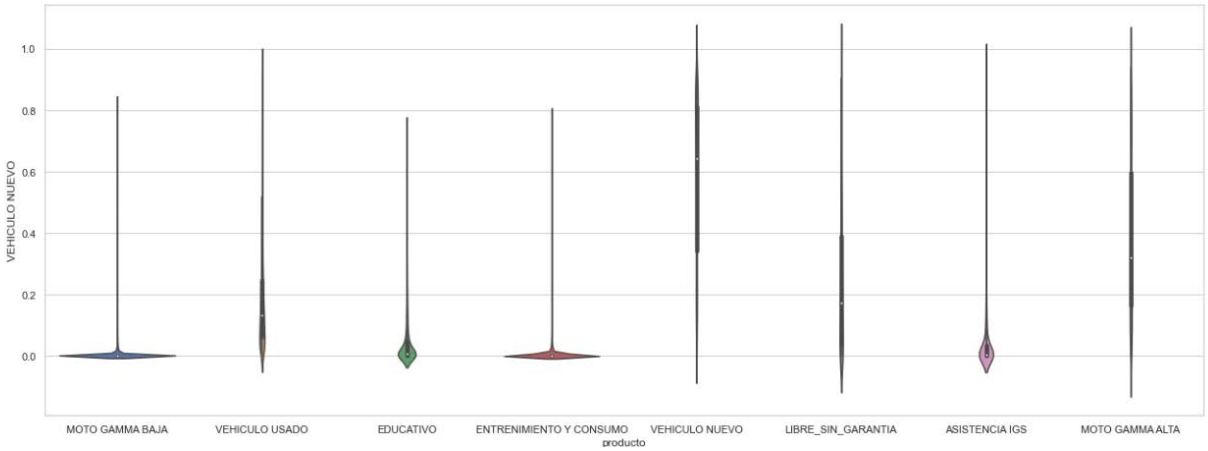


Fig. 12. Gráfico de violines Entretenimiento y Consumo

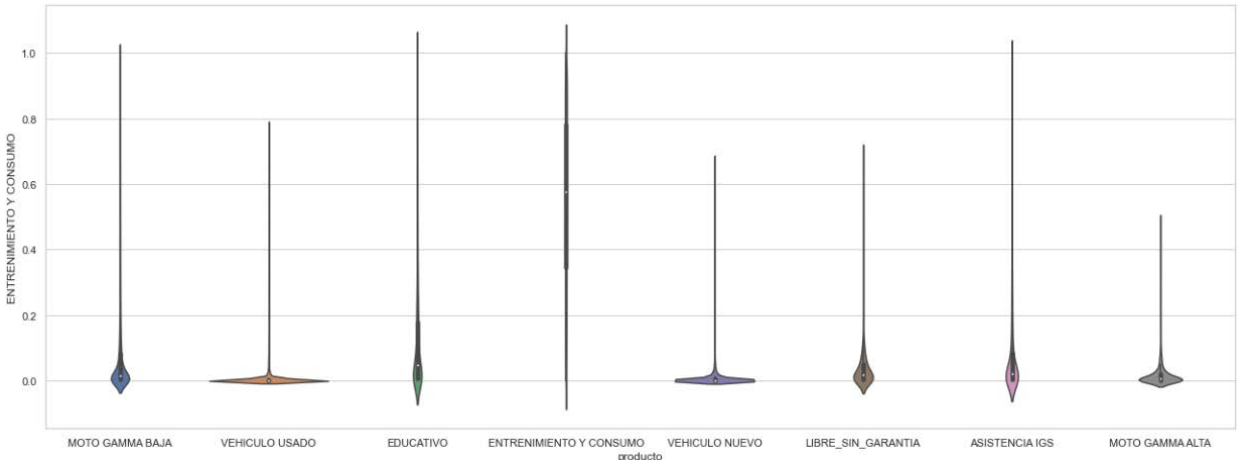


Fig. 13. Gráfico de violines Libre Inversión

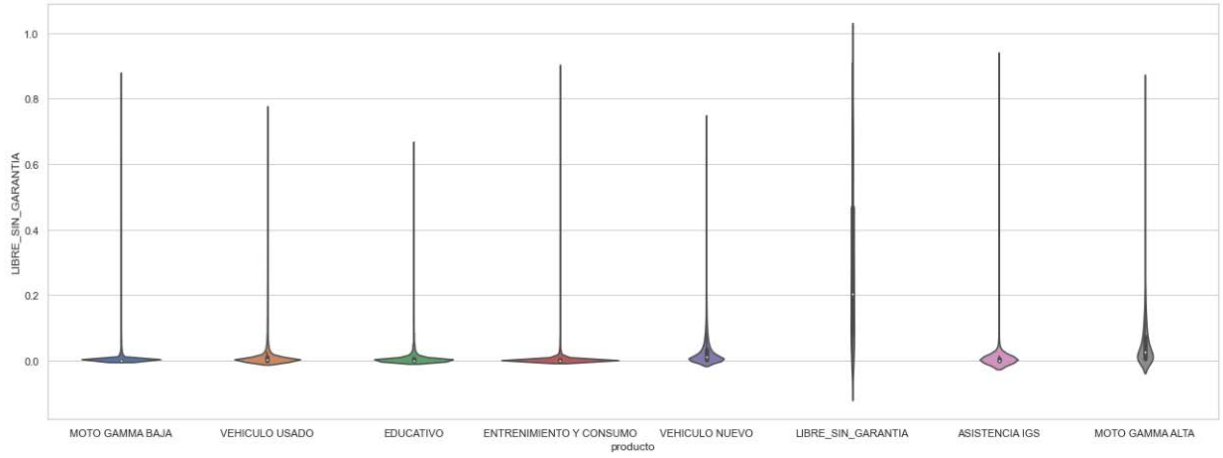


Fig. 14. Gráfico de violines Asistencia Igs

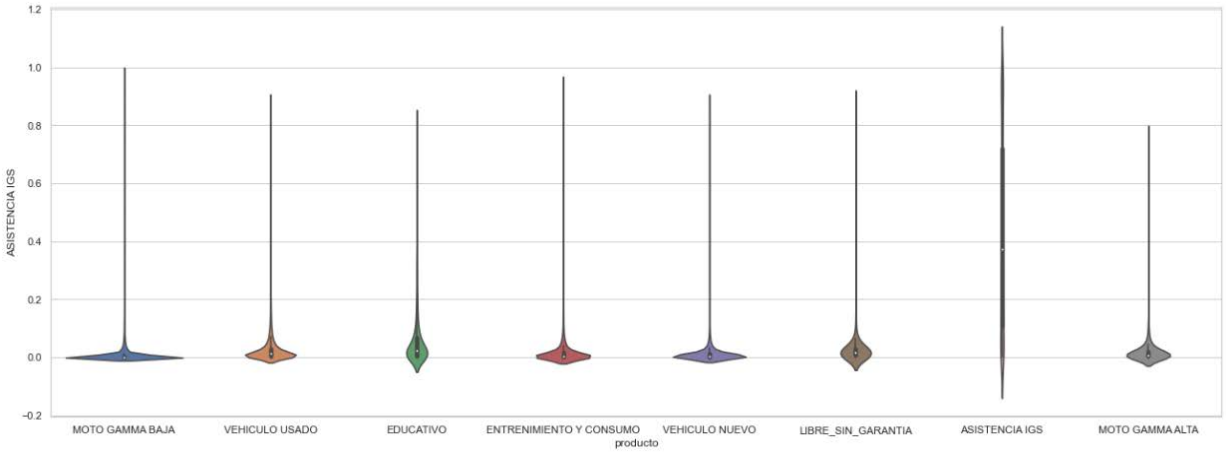


Fig. 15. Gráfico de violines Educativo

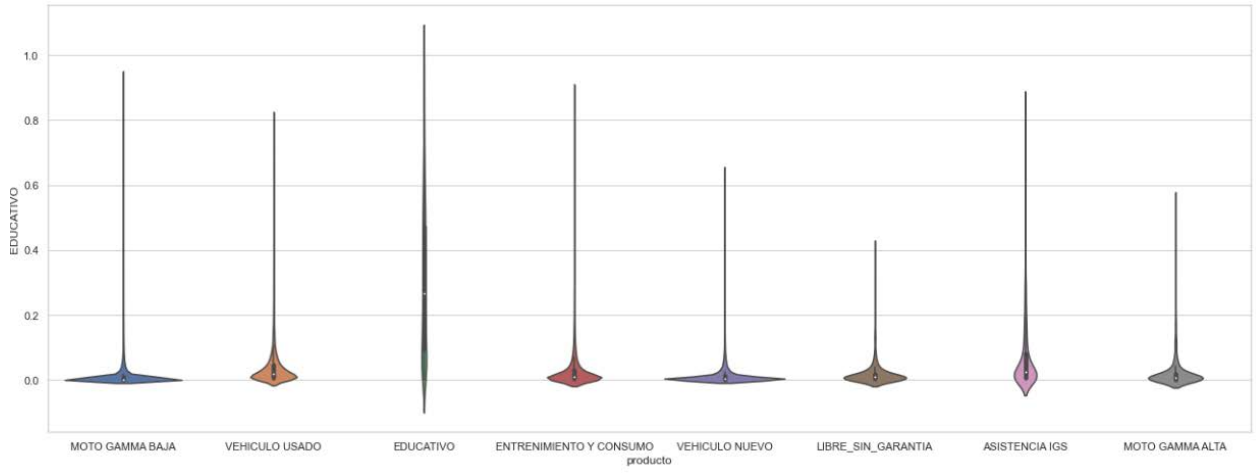


Fig. 16. Gráfico de violines Moto Gamma alta