ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

ANÁLISIS DE EFECTIVIDAD DE CAMPAÑAS EN EMPRESAS DE CONSUMO MASIVO, MEDIANTE EL USO DE MODELOS PROBABILÍSTICOS DE CLASIFICACIÓN BINARIA

Profesor
Ing. Mario González. PhD

Autor
Willy Giovanny Lema Maigua

2023

RESUMEN

El presente proyecto tiene como propósito principal realizar un análisis de la efectividad de campañas, que pueden ser de marketing o promocionales, para el caso de empresas de consumo masivo con enfoque al tipo de negocios retail (supermercados). Esto por medio de la utilización de modelos probabilísticos de predicción binaria que representan el éxito o rechazo de estas campañas.

Los datos para el análisis corresponden a un dataset obtenido en el repositorio público Kaggle, el cual contiene información acerca de una empresa de consumo con registros anonimizados de clientes, datos demográficos, históricos y preferencias de consumo de los mismos, finalizando con información de su aceptación o rechazo a determinada campaña.

Se realiza un análisis de la correlación de las variables frente a la variable de respuesta que permiten entender el comportamiento del consumidor. Así también para predecir la probabilidad de que un cliente dé una respuesta positiva frente a una campaña, se hace uso de los modelos más comunes de clasificación algorítmica como son regresión logística, árboles decisión, random forest, entre otros para poder obtener el modelo y metodología de mayor precisión en la predicción de los datos.

Adicionalmente, los resultados presentados en este proyecto permiten conocer cuáles son los factores que más inciden para el éxito de una campaña en este caso y poder extrapolar estos hallazgos para combinarlos con el conocimiento empresarial para establecer estrategias más generalizadas que permitan a las empresas dirigir mejor sus proyectos y que estos tengan un alcance eficiente ante los consumidores.

Palabras Clave: Predicción de campañas, análisis exploratorio, consumo masivo, supermercados, clasificación algorítmica, random forest, regresión.

ABSTRACT

The main purpose of this project is to carry out an analysis of the effectiveness of campaigns, which can be marketing or promotional, in the case of mass consumption companies focused on the type of retail business (supermarkets). This using probabilistic binary prediction models that represent the success or rejection of these campaigns.

The data for the analysis corresponds to a dataset obtained from the Kaggle public repository, which contains information about a consumer company with anonymized customer records, demographic data, history, and consumer preferences, ending with information on their acceptance. or rejection of a certain campaign.

An analysis of the correlation of the variables is carried out in front of the response variable that allows to understand the behavior of the consumer. Likewise, to predict the probability that a client will give a positive response to a campaign, the most common algorithmic classification models are used, such as logistic regression, decision trees, random forest, among others, in order to obtain the model and methodology. more accurate in predicting the data.

Additionally, the results presented in this project allow us to adequately know the factors that most affect the success of a campaign in this case and to be able to extrapolate these conclusions to combine them with business knowledge to establish more generalized strategies that allow companies to better direct their projects and have an efficient reach to these consumers.

Keywords: Campaign prediction, exploratory analysis, mass consumption, supermarkets, algorithmic classification, random forest, regression.

ÍNDICE DEL CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

La tecnología concerniente al análisis de datos se ha desarrollado a pasos agigantados en el mundo en los últimos años. Con esto muchas empresas se han encaminado en proyectos relacionadas al aprovechamiento de sus bases de datos para generar valor en sus operaciones. Justamente a nivel empresarial una de las aplicaciones más importantes de la ciencia de datos se da en los procesos de toma de decisiones, puesto que esto permite que mediante analítica se busque minimizar el riesgo de actividades empresariales fallidas y por otro lado maximizar los resultados de la implementación de proyectos que busquen mejorar cualquier proceso o actividad que una organización realice.

Las herramientas de la ciencia de datos, machine learning y otras herramientas de análisis han proporcionado a las empresas varias ventajas competitivas que antes no se hubiesen podido realizar. En esto podemos hablar de modelos de predicción de demanda, ventas, mejor control de inventarios, eficiencia logística e incluso la posibilidad de comprender de mejor manera a los consumidores frente a sus requerimientos personalizados de productos o servicios, todo esto haciendo que las empresas puedan hacer frente a todas estas situaciones de manera prevista y a partir de esto todos sus procesos sean más eficientes y rentables.

En el Ecuador todavía no se ha difundido la aplicación de este tipo de metodologías y por ende existen varias áreas de industria que pudieran verse beneficiadas de esta tecnología analítica. En este contexto, es oportuno mencionar que una de las aplicaciones de la ciencia de datos en nuestro país vendría a enfrentar el hecho de que muchas de los emprendimientos que nacen en el país tienden a cerrar después de cortos períodos de funcionamiento. Pero esta problemática no solo trasciende en pequeñas empresas sino también en empresas de mayor tamaño, para las cuáles se ha visto que varios de los proyectos que realizan en función de hacer conocer su marca, servicio o

producto no tienen el impacto positivo que se pretende o en muchos casos simplemente el beneficio que se busca no es de consideración.

Para enfrentar esta problemática empresarial, el análisis de datos se ha venido aplicando en las empresas a través de la utilización de modelos que permiten, entre otras aplicaciones, predecir cuál puede ser el resultado de la implementación de un proyecto de mejorar en la empresa, por ejemplo, la implementación de proyectos de marketing o campañas de servicios o productos que en la actualidad son fundamentales en cualquier organización nueva o con trayectoria.

Tomando el enfoque de las campañas para promocionar un servicio de una empresa y para maximizar el éxito de este, el presente estudio busca analizar el caso de una empresa de consumo masivo que decide implementar un proyecto de marketing para fidelizar a sus clientes por medio de un servicio nuevo. Se usa una base de datos anonimizada del registro de consumo de sus clientes, preferencias y resultados de aceptación o rechazo al ser consultados sobre su interés de acceder a un nuevo servicio para poder construir modelos que busquen predecir este comportamiento. Estos modelos buscan analizar la incidencia de los factores individuales de las personas para entender cuáles son más significativos y poder en el futuro tener mayor efectividad frente a implementaciones de proyectos similares o simplemente mejorar el servicio para estas personas que han respondido positiva y claramente representan una parte importante de la rentabilidad de un negocio.

Se busca analizar los mejores modelos que brinden exactitud para este caso específico, compararlos y extrapolar sugerencias de cuáles de estos serían más adecuados para otras actividades productivas o comerciales que también se dan en el país. De la misma manera esto contribuye al objetivo de contribuir con ejemplos de aplicación efectiva de la metodología analítica de la ciencia de datos en la industria y en las organizaciones.

2. REVISIÓN DE LITERATURA

El marketing y la ciencia de datos son dos áreas interrelacionadas que desempeñan un papel crucial en el mundo empresarial actual. El mundo empresarial es cambiante y la tecnología también, por lo que deben ir de la mano, si industria empresarial deslinda su transcurso de la tecnología, es muy probable que varias deficiencias se presenten afectando no solamente el rendimiento actual de la organización, sino también imposibilitando que proyectos de avance tecnológico efectivos se puedan realizar. (Zamorano, 2018) Las empresas deben ser eficientes desde todos los puntos de vista, por lo que es importante poner énfasis en varios puntos y desde un inicio trazar las metodologías para su avance. (Calva, 2021)

Marketing: El marketing se refiere a las actividades que una empresa realiza para promocionar, vender y entregar productos o servicios a los consumidores. Su objetivo principal es comprender y satisfacer las necesidades y deseos de los clientes de manera rentable. (Marín, 2019) El marketing implica estrategias y tácticas para identificar y alcanzar a los clientes objetivo, posicionar productos o servicios en el mercado, establecer relaciones con los clientes y medir los resultados para tomar decisiones informadas. Incluye áreas como investigación de mercado, segmentación, branding, publicidad, promoción, relaciones públicas, ventas y gestión de clientes. (Salazar, 2019)

Ciencia de datos: La ciencia de datos se centra en la extracción de conocimiento y perspectivas a partir de grandes conjuntos de datos. Utiliza técnicas y herramientas estadísticas, matemáticas y de aprendizaje automático para recopilar, analizar, interpretar y visualizar datos con el fin de obtener información valiosa y respaldar la toma de decisiones. (Espino, 2017) La ciencia de datos involucra la recopilación de datos, el procesamiento y limpieza de los mismos, el modelado estadístico, la creación de algoritmos de aprendizaje automático y la comunicación de los resultados obtenidos. (Giraldo, 2018)

2. 1 RELACIÓN ENTRE EL MARKETING Y LA CIENCIA DE DATOS

El marketing se ha beneficiado enormemente de la ciencia de datos. El análisis de datos ha permitido a los profesionales del marketing obtener una comprensión más profunda de los clientes, identificar patrones de comportamiento, realizar segmentaciones más precisas y personalizar las estrategias de marketing. Al utilizar técnicas de ciencia de datos, los especialistas en marketing pueden medir el rendimiento de sus campañas, evaluar el retorno de la inversión y optimizar sus estrategias para obtener mejores resultados. (Zúñiga, 2023)

La ciencia de datos también puede ayudar a mejorar la experiencia del cliente al proporcionar recomendaciones personalizadas, predecir la demanda futura, identificar oportunidades de mercado y comprender el impacto de las decisiones de marketing en el rendimiento empresarial. En resumen, la ciencia de datos brinda a los profesionales del marketing una base sólida para tomar decisiones más informadas y estratégicas. (Carrasco, 2017)

En conjunto, el marketing y la ciencia de datos se complementan y se potencian mutuamente. La ciencia de datos aporta un enfoque basado en datos y análisis cuantitativo al marketing, lo que permite una toma de decisiones más precisa y fundamentada, mientras que el marketing brinda el contexto y las metas comerciales necesarias para orientar el análisis de datos hacia resultados prácticos y estratégicos.

2.2 ANÁLISIS DE CLIENTES Y SEGMENTACIÓN

Los canales de comunicación han crecido de manera bastante rápida en los últimos años. Con la ayuda de la tecnología es posible que un mensaje puede llegar a expandirse a diferentes poblaciones en periodos bastante cortos de tiempo. En este contexto, las organizaciones comerciales también han aprovechado esta situación para poder comunicar sus servicios a potenciales

consumidores. La utilidad de la ciencia de datos en este sentido es el de poder aprovechar este canal comunicacional para llegar a un público definido.

Esta definición del público, se hace llamar segmentación, y permite que una organización o empresa puede delimitar el público a quien va dirigido un determinado comunicado. (Delgado, 2017). La segmentación, es una estrategia que permite que la comunicación tenga más probabilidades de ser efectiva procurando reducir la implementación de recursos de forma ineficiente para distintos tipos de objetivo. La transmisión de mensajes se puede volver más ajustada con el objetivo de mejorar los resultados que busca una organización en la comunicación de sus mensajes y comunicados.

2. 3 ANALÍTICA AVANZADA EN MARKETING

Las herramientas de la ciencia de datos pueden ser combinada con los conceptos de marketing para buscar los mejores resultados en el tema de análisis operacional de una empresa. Diferentes técnicas de predicción pueden aplicarse para analizar distintos procesos de una organización. Por ejemplo, se puede ocupar técnicas de regresión para el caso de análisis y planeación de ventas, y determinación de precios combinados junto a otros factores. El marketing en el campo empresarial también puede funcionar en casos de búsqueda de fidelización de clientes porque se puede buscar ajustarse a las necesidades de los consumidores y atender estos requerimientos. (Grewal & Kopalle, 2019)

Para las empresas, uno de los objetivos más grandes es que se pueda automatizar el marketing de la organización con la ayuda de datos. Es decir, buscar una implementación de marketing basado en datos. Para esto es importante conocer no solo las herramientas que se tiene a disposición sino también el de tomar en cuenta en cuenta el objetivo de entender patrones de conducta, comportamiento de consumidores entre otros. El marketing en este sentido se encuentra en constante evolución, entonces es necesario seguir con

una retroalimentación frecuente de estos temas para seguir haciendo uso de los mercados más necesarios en este tipo de negocios.

2.4 BENEFICIOS DE LA IMPLEMENTACIÓN DE LA CIENCIA DE DATOS EN MARKETING

Ciertamente los beneficios que se pueden tener desde el campo de marketing para las aplicaciones empresariales son bastante variados. Se tratar de tener al consumidor como fuente principal de retroalimentación para mejorar y también posicionarlo como el ente al cual toda clase de innovación va dirigida. Al final de todo proceso de marketing, se busca que todos los rendimientos sean optimizados. (Roland, 2020). Si bien los beneficios que la ciencia de datos en temas de marketing representa a una empresa en términos de mejoras, estos pueden ser agrupados en enfoques de mejora relacionados a:

2.4.1 Mejora de la experiencia de los consumidores

La experiencia que un cliente lleva luego de haber adquirido algún producto o servicio puede ser la clave para que este vuelva a realizar la misma acción en un futuro. Es por esto que mejorar la probabilidad de que un cliente tenga una excelente experiencia en un negocio es de vital importancia para una organización. Partiendo de la recolección de datos que se puede tener en torno a los requerimientos y necesidades de un consumidor se poded realizar varias mejoras para satisfacer esto. Todo esto representará una ventaja competitiva para la organización justamente por tener en cuenta lo opinión de los consumidores. (Rubio, 2019)

2.4.2 Mejorar la eficiencia de gastos

Los recursos con los que cuenta una empresa son bastante valiosos, por lo que siempre se busca que estos sean utilizados de la mejor manera y evitando sobre todo que estos recursos no sean bien dirigidos o aprovechados. Para optimizar los gastos, se debe tomar en cuenta el ROI (retorno de la inversión) y verificar

que canales y herramientas pueden ser más útiles para la correcta administración de estos recursos. Si una organización puede implementar una correcta administración de los recursos además de mantener esto como una constante en el tiempo, los beneficios pueden traducirse como rentabilidad dentro de una empresa.

2.4.3 Implementación de las estrategias SEO

La ciencia de datos en el campo del marketing esta relacionado en el objetivo que tienen los científicos de datos de aumentar el tráfico de buscadores frente a este tema. En este contexto, el tráfico de buscadores se traduce al posicionamiento de una marca en las redes lo cual es bastante importante para el sistema actual de comercialización que prevalece en la actualidad. El posicionamiento correcto de una organización también tiene el objetivo de reducir la incertidumbre que existe al momento de realizar una campaña de marketing. El combinar las estrategias de SEO y el posicionamiento puede resultar en un gran beneficio para una empresa desde el punto de vista de la perspectiva que la gente pueda tener acerca de una organización. (Namdhini, 2021)

2.4.4 Lead scoring y implementación empresarial

La implementación de una correcta clasificación de los clientes o también denominada segmentación de clientes es esencial. Esto desde el punto de vista el objetivo de incrementar el tráfico de ventas por la correcta identificación de los consumidores y de las necesidades de los mismos. Los sistemas de puntuación en términos de preferencia de consumo y intereses pueden marcar una diferencia para aumentar la eficiencia operacional de una empresa. Como un paso adicional se puede crear bases de datos a partir de la información de los clientes para agrupar preferencias y se pueda optar por investigar tendencias que se pueden presentar acerca del comportamiento de los clientes. Identificar

estos patrones claramente puede ser relacionado con una mejora en el lead scoring de una empresa. (Saura, 2021)

2.4.5 Entender al público objetivo

El público objetivo es área clave de un negocio puesto que es el ente hacia donde se dirigen todos los esfuerzos de una organización. La organización que se tenga en cuanto a la recolección de datos y la calidad de la misma será la base para poder tener una buena recopilación de datos de las opiniones y percepciones de los usuarios. Esta información es valiosa para poder determinar que áreas se deben mejorar. Así mismo entender a los clientes permite poder efectivizar campañas de promoción, de manera que sea posible obtener una retroalimentación que sea la base de mejoras en cualquier área de una empresa. La retroalimentación en este sentido también es una estrategia no solamente del presente de una organización sino también de la proyección al futuro de esta.

2.4.6 Marketing por canales personalizados

La personalización es un valor agregado de suma importancia en las operaciones de una organización. Esta personalización tiene que ver con manejo de los canales por los cuáles la información llega a los usuarios. Entender cuáles son los medios más efectivos para poder llegar a la gente claramente puede significar la posibilidad de marcar una ventaja competitiva que diferencia a una empresa. Bajo este contexto los canales pueden ser mejor aprovechados para la información llegue de forma directa y con un mensaje que pueda ser mejor receptado por el consumidor. Desde el punto de la ciencia de datos, la información recolectada acerca de preferencias y históricos de consumo puede ser utilizada para obtener la mejor predicción acercas de cuáles serán los movimientos futuros en esta área. (Shobhana, 2022)

3. IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

El objeto de estudio son las respuestas que los consumidores tienen ante campañas de marketing o promocionales que las empresas realizan dentro de sus operaciones. En este caso hacemos referencia al sector retail de supermercados donde la empresa en estudio acerca a sus clientes una oferta promocional y busca una respuesta positiva para que el cliente acceda a la promoción o consuma un determinado producto ofertado.

Se pretende comprender cuáles son los factores que más influyen para un determinado cliente acceda a ser parte de esta campaña y de esta manera tener herramientas que permitan a la empresa dirigir mejor sus productos hacia un segmento de mercado o población objetivo. Con esto también se busca que la asignación de recursos para este tipo de proyectos sea lo más eficiente y rentable para la empresa y se pueda establecer planes de fidelización de clientes.

Con esta investigación, es posible identificar patrones y características específicas de las personas que acceden a este tipo de marketing dentro de esta rama de negocios, lo cual puede ser de mucha ayuda para establecerse como factores de referencia a tomar en cuenta para empresas de este sector productivo ante la implementación de proyectos similares y maximizar la inversión de recursos en este sentido.

Finalmente, la metodología aplicada en este estudio, que comprende la comparación de los mejores modelos de predicción y selección del más adecuado busca brindar las bases de procedimiento para que otros sectores productivos también puedan realizar proyectos similares como lo son en el campo financiero, tratamientos médicos, efectividad en el sector educativo, entre otros.

4. PLANTEAMIENTO DEL PROBLEMA

En el Ecuador las empresas, casi no han aplicado metodologías de segmentación de mercado para los productos que ofrecen por falta de conocimiento o simplemente por lo que representa realizar estudios como estos. Sin saber que justamente realizar esta segmentación puede tener grandes beneficios empresariales como lo tener un mejor entendimiento de los clientes, selección de productos y la reducción de la posibilidad de que el negocio tenga que finalizar sus operaciones y cerrar.

El sector de los supermercados tampoco es la excepción, y si bien hay grandes cadenas que manejan este tipo de operaciones, también existen empresas de menor tamaño en este sector y en otros que no ocupan este tipo de metodologías haciendo que sus operaciones no siempre sean eficientes y rentables. En esto las campañas de marketing y de promoción de parte de las empresas buscan llegar a la gente con un determinado producto o servicio, promocionarlo y incrementar sus ventas e ingresos. El problema es que, debido a la falta de identificación del mercado objetivo específico para las empresas y correcta segmentación de mercados, estas campañas no siempre podrían tener un efecto positivo en el objetivo de mejorar los ingresos o las ventas de una empresa. Las empresas que no tienen claro cómo realizar esta segmentación y al momento de realizar campañas tienen un riesgo más alto de incurrir en pérdidas de recursos y desaprovechamiento de este tipo de proyectos.

Es por esto que se busca realizar un análisis de los datos de una empresa frente a la realización de una campaña de marketing para sus clientes, donde se registra si estos finalmente terminaron accediendo a la promoción o no. Para esto se toman en cuenta distintos factores demográficos y de preferencias de consumo de los clientes y finalmente su respuesta para formular modelos que permitan predecir el éxito o no de estas campañas. Al hacer esto, se podrá observar los factores más representativos que influyen en la decisión de un cliente de aceptar o no una oferta y con esto extrapolar que factores debe una

empresa mejorar para que los clientes tengan una respuesta positiva a estas campañas.

Esta metodología busca crear una referencia para el sector de consumo masivo retail, supermercados, pero también fácilmente aplicable a otras actividades productivas frente a proyectos donde se busque predecir una variable de respuesta tomando en cuenta como datos ciertas características recolectadas del consumidor o de una transacción.

5. OBJETIVO GENERAL

Realizar un pronóstico del éxito o rechazo de una campaña de promoción en una empresa de consumo masivo, con el fin de determinar la correlación entre variables y determinar los principales factores que influyen a la variable de respuesta. Entender la metodología y analizar los modelos de predicción usados como una herramienta que puede ser aplicable a otros proyectos similares en otras ramas productivas.

6. OBJETIVOS ESPECÍFICOS

- Determinar a través de un análisis descriptivo los principales factores que en el caso de estudio influyen en la aceptación o rechazo por parte del cliente, definir características demográficas, socioeconómicos y específicos de los consumidores del estudio.

- Pronosticar la aceptación de los consumidores frente a la campaña realizada por la empresa en estudio, utilizando modelos de predicción clasificatoria binaria y compararlos.

- Establecer soluciones y sugerencias estratégicas que la empresa pueda llevar a cabo frente a los hallazgos del estudio de predicción de éxito o rechazo de la campaña realizada para sus clientes en el presente estudio.

- Realizar un análisis gráfico detallado para el caso de la empresa que permita encontrar patrones y rasgos de sus consumidores para los objetivos comerciales de la misma.

7. JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

Para el presente estudio, se busca realizar un análisis predictivo de éxito de una campaña de marketing en una empresa de comercialización de consumo masivo. Para esto es necesario analizar los factores que se tienen en la base de datos, realizar correlaciones entre las mismas e identificar los factores de mayor incidencia para el modelo predictivo. La importancia de este análisis es de utilidad principalmente para el caso de la empresa, que busca optimizar la utilización de sus recursos enfocados en sus clientes y poder hacer énfasis en la fidelización de estos.

Se abarca esta temática debido a que, en el sector industrial comercial actual, diferentes estrategias de posicionamiento de marca, productos y servicios se han venido utilizando. Sin embargo, una conceptualización de la correcta aplicación de estas estrategias es todavía una deficiencia lo que provoca que en muchos casos las empresas que deciden optar por este tipo de proyectos terminan sin tener los resultados esperados en cuanto a la efectividad de estas campañas y solamente representando egresos que no justifican las inversiones realizadas. En el contexto ecuatoriano, esto sucede aún más regularmente y se puede notar que las empresas tienen dificultades para definir a sus mercados objetivos base por lo que las campañas de promoción de productos y servicios no tengan el éxito deseado.

La ciencia de datos es un conjunto de herramientas que por su naturaleza hace énfasis en los datos que se han recogido históricamente dentro de la empresa y busca también tomar en cuenta a la recolección continua de datos para sus análisis. Esta es una potencial estrategia que permita que la implementación de campañas para servicios y productos sea más efectiva, debido a que se busca

realizar no solamente un construir un modelo como tal sino el de analizar el nivel de incidencia que tienen cada una de las variables. Esto de suma utilidad para una empresa que busca encontrar aquellos sectores en donde se debe mejorar o cambiar estrategias.

El análisis predictivo del presente estudio se realiza tomando en cuenta la variable de salida que es del tipo binario. Es decir, una respuesta afirmativa o negativa de parte del cliente ante la oferta de una campaña promocional. Es decir, se utiliza modelos de predicción y clasificación binaria. De entre estos se puede mencionar a los modelos de árbol de decisión, random forest, extra tres y regresión logística que se utilizan en este estudio.

Estos modelos permiten analizar la variable de salida de un modelo, pero tomando en cuenta la significancia individual y correlacional entre las variables del estudio.

7.1. MODELOS ELEGIDOS PARA EL ESTUDIO

7.1.1 Árbol de Decisión. Son modelos predictivos que se forman a partir de reglas binarias para clasificar las observaciones en función de sus atributos y de esta manera poder determinar la variable de respuesta. Los modelos de árboles de decisión brindan la posibilidad de poder utilizar tanto valores cuantitativos como categóricos, aunque esto también depende del enfoque del estudio. (Amat, 2020). En un modelo de árbol de decisión, las observaciones de entrenamiento se agrupan en los nodos terminales. Esto quiere decir que para predecir una nueva observación se recorre el árbol hasta que este llega a uno de los nodos de finalización.

7.1.2 Random Forest. Es un modelo versátil que puede realizar tareas tanto de predicción como de regresión. El modelo random forest está formado por un conjunto de árboles decisión para los cuales cada uno de ellos se encuentra entrenado con una muestra diferente de los datos de entrenamiento. Se plantea

como una técnica supervisada que busca segmentar el espacio de los predictores en regiones más simples donde las interacciones entre las variables se pueden controlar de mejor manera. Al ser esta un método no paramétrico, no es necesario que se cumpla con ningún tipo de distribución.

7.1.3 Extra trees. También denominados Árboles extremadamente aleatorios, este algoritmo tiene un enfoque central similar al del modelo de Random Forest pero con las diferencias de que al momento de crear un subconjunto de las características predictivas para cada uno de los árboles a crear, se genera un valor aleatorio para cada característica planteada y luego procediendo a escoger el mejor de ellos. Otra diferencia es que el modelo de extra trees muestrea sin reemplazo considerando divisiones aleatorios y no mejores decisiones como se plantea en el modelo de random forest. (Huertas, 2020)

7.1.4 Regresión Logística. Este modelo tiene como objetivo comprobar hipótesis o relaciones causales en los casos en los cuáles la variable dependiente es categórica. La regresión logística está basada en las probabilidades y principios de odd ratio. Esto es equivalente a decir que el modelo toma en cuenta a las variables independientes para predecir la probabilidad de que algo suceda sobre la probabilidad de que no ocurra. (Cárdenas, 2022) La regresión logística es útil para identificar causas de sucesos o fenómenos que ayudan a entender una variable de respuesta específica.

7.2. SELECCIÓN DE LA BASE DE DATOS

La base de datos seleccionada para el estudio proviene del repositorio de analítica Kaggle. La misma detalla información recolectada de una empresa de consumo masivo (Supermercado) sobre una campaña promocional que se realizó para sus clientes. Esta base de datos contiene información demográfica de los participantes, preferencia de consumo de estos y sobre todo información acerca de su aceptación o rechazo a la campaña.

La base de datos original contiene 2240 registros clasificados en 22 campos o variables.

7. 3. IDENTIFICACIÓN Y DESCRIPCIÓN DE VARIABLES

El análisis descriptivo y predictivo para el presente estudio se enfoca en la variable dependiente “Respuesta”, que corresponde a la información del tipo binario (1: positivo 0: negativo) de la aceptación del cliente frente a la campaña promocional ofrecida por parte de la empresa de consumo masivo. Las demás variables son consideradas como independientes, y siguen la siguiente clasificación:

Tabla 1: Categorización de variables Base de Datos (campaña promocional)

Variable	**Descripción de campo**	**Tipo de Variable**	**Clasificación**
Id	ID único de cada cliente.	Numérica / int64	Independiente
Año_Nacimiento	Edad del cliente.	Numérica / int64	Independiente
Niv_Educación	Nivel de educación del cliente.	Object	Independiente
Estado_Civil	Estado civil del cliente.	Object	Independiente
Ingresos	Ingresos familiares anuales del cliente.	Numérica / float 64	Independiente
N_Niños	Número de niños pequeños en el hogar del cliente.	Numérica / int64	Independiente
N_Adolescentes	Número de adolescentes en el hogar del cliente.	Numérica / int64	Independiente
Fecha_Cliente	Fecha de alta del cliente en la empresa.	Object	Independiente
Ult_Compra	Número de días desde la última compra.	Numérica / int64	Independiente
C_Vinos	La cantidad gastada en productos vitivinícolas en los últimos 2 años.	Numérica / int64	Independiente
C_Frutas	La cantidad gastada en productos de frutas en los últimos 2 años.	Numérica / int64	Independiente
C_Carnes	La cantidad gastada en productos cárnicos en los últimos 2 años.	Numérica / int64	Independiente
C_ProdsMar	La cantidad gastada en productos pesqueros en los últimos 2 años.	Numérica / int64	Independiente

C_Dulces	Cantidad gastada en productos dulces en los últimos 2 años.	Numérica / int64	Independiente
C_PremiumProds	La cantidad gastada en productos de oro en los últimos 2 años.	Numérica / int64	Independiente
N_CompPromos	Número de compras realizadas con descuento.	Numérica / int64	Independiente
N_CompWeb	Número de compras realizadas a través de la web de la empresa.	Numérica / int64	Independiente
N_CompCatalogo	Número de compras realizadas por catálogo (compra de productos para enviar por correo).	Numérica / int64	Independiente
N_CompTiendas	Número de compras realizadas directamente en tiendas.	Numérica / int64	Independiente
N_VisitasWebMes	Número de visitas al sitio web de la empresa en el último mes.	Numérica / int64	Independiente
Reclamo	1 si el cliente se quejó en los últimos 2 años.	Numérica / int64	Independiente
Respuesta	1 si el cliente aceptó la oferta en la última campaña, 0 en caso contrario.	Numérica / int64	Dependiente

Fuente: *(Repositorio de bases Kaggle) / Elaboración propia*

7.4. PRE-PROCESAMIENTO Y LIMPIEZA DE DATOS

Con la finalidad de contar con información precisa y consistente en el estudio, además de buscar que las predicciones que busca realizar el modelo, se procede procesar la base de datos en las etapas de limpieza y descripción estadística del mismo. Para esto se utiliza los softwares Excel, Open Refine y Python para analizar el comportamiento de las variables en primera instancia.

7.4.1. PREPROCESAMIENTO

Teniendo como objetivo el poder hacer que la base de datos se adapte eficientemente a los objetivos de estudio, se toma a Python como principal software de análisis. Se verifica que la base sea consistente y pueda cargarse desde el archivo origen *.csv* de la campaña de marketing de la empresa de consumo. Se procede verificar también el proceso inicial de renombre de variables, que originalmente se encuentran en idioma inglés, hacia español para hacerlas más entendibles al contexto del análisis. Por otra parte, también se

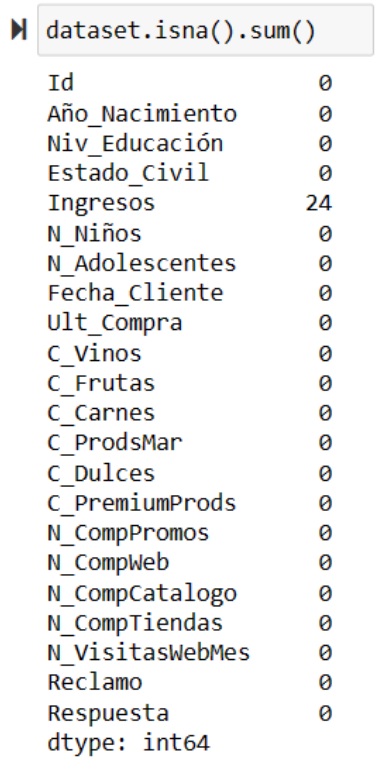
realiza la eliminación de registros duplicados. Todo esto como preámbulo de las siguientes fases del análisis de la base.

Debido a la naturaleza de la base de datos y para los objetivos de estudio, no se realiza cruces de la misma con otras fuentes de datos.

7.4.2. LIMPIEZA

Partiendo de la base de datos original que consiste en 2240 registros dentro de los 22 campos en los cuáles se clasifican los datos de la base, se procede a realizar a la validación de los formatos de los datos, y la eliminación de errores tipográficos en primera instancia.
Luego, con la ayuda de Python se procede a analizar la base para identificar registros vacíos, datos duplicados e inconsistentes. Por lo que se identifica 24 registros inconsistentes en un campo específico.

Figura 1. Verificación de datos inconsistentes para la base de datos (previa limpieza)



```
dataset.isna().sum()

Id                   0
Año_Nacimiento       0
Niv_Educación        0
Estado_Civil         0
Ingresos            24
N_Niños              0
N_Adolescentes       0
Fecha_Cliente        0
Ult_Compra           0
C_Vinos              0
C_Frutas             0
C_Carnes             0
C_ProdsMar           0
C_Dulces             0
C_PremiumProds       0
N_CompPromos         0
N_CompWeb            0
N_CompCatalogo       0
N_CompTiendas        0
N_VisitasWebMes      0
Reclamo              0
Respuesta            0
dtype: int64
```

Fuente: *Python. Análisis Base de datos Campaña de Marketing*

Con el mismo software y con la finalidad de hacer que la base pueda ser analizada de la mejor manera para tener resultados más fiables en cuanto a interacción de variables y modelo de predicción se procede a limpiar estos datos. Con esto la base a analizar consistirá en 2216 registros bajo los mismos 22 campos de clasificación.

Figura 2. Verificación de datos inconsistentes para la base de datos (post limpieza)

```
dataset.isna().sum()
```

```
Id                   0
Año_Nacimiento       0
Niv_Educación        0
Estado_Civil         0
Ingresos             0
N_Niños              0
N_Adolescentes       0
Fecha_Cliente        0
Ult_Compra           0
C_Vinos              0
C_Frutas             0
C_Carnes             0
C_ProdsMar           0
C_Dulces             0
C_PremiumProds       0
N_CompPromos         0
N_CompWeb            0
N_CompCatalogo       0
N_CompTiendas        0
N_VisitasWebMes      0
Reclamo              0
Respuesta            0
dtype: int64
```

Fuente: *Python. Análisis Base de datos Campaña de Marketing*

7.5. ANÁLISIS DESCRIPTIVO / EXPLORATORIO

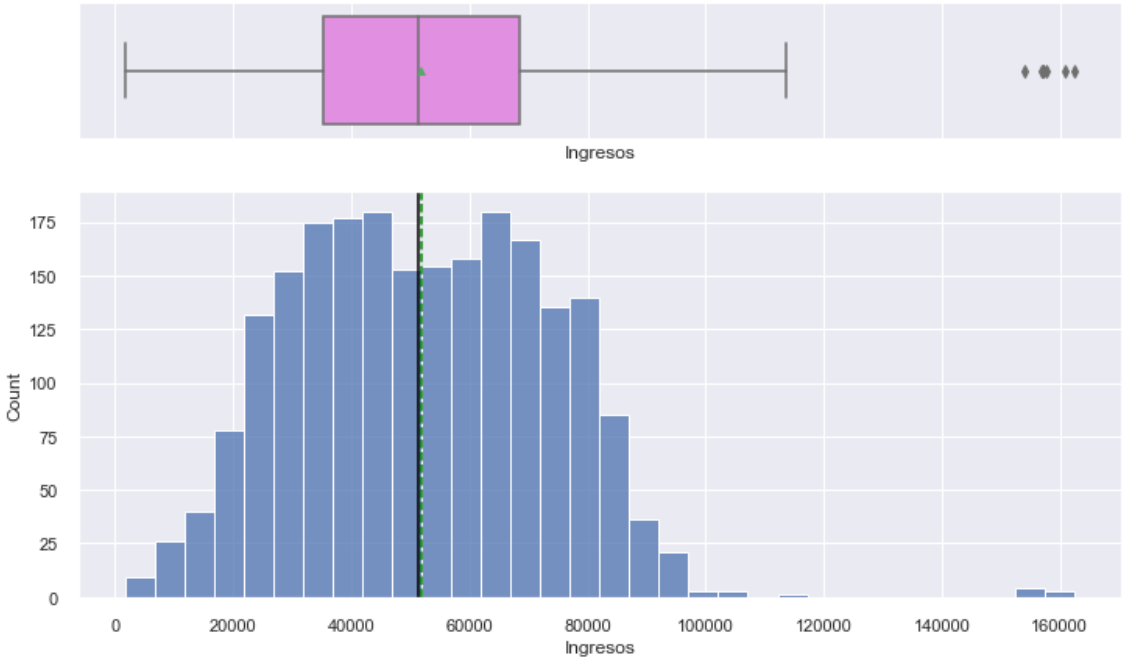
En esta etapa se realiza el análisis descriptivo de la base de datos y las variables contenidas en el mismo. Esto con la finalidad de poder observar tendencias o patrones que contribuyan en el estudio de la base. Adicionalmente se utilizan técnicas de visualización para identificar hallazgos gráficamente y direccionar el estudio hacia los objetivos planteados.

Se muestran las variables del estudio tomando en cuenta en que estas ya han pasado por la fase de preprocesamiento y limpieza para tomarse en cuenta dentro del estudio como tal.

7.5.1. ANALISIS UNIVARIADO

1 | Ingresos

Figura 3. Histograma de la variable “Ingresos”

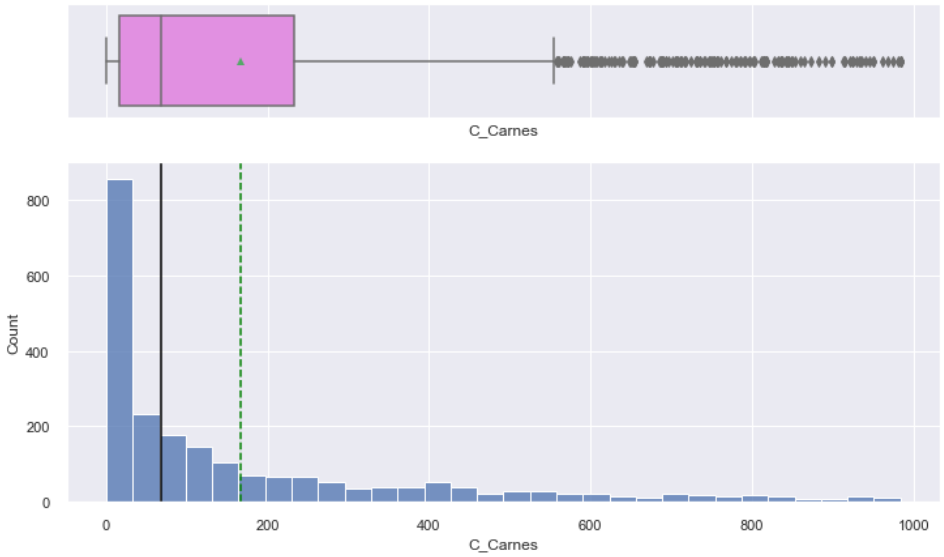


Fuente: *Elaboración propia*

- De la gráfica de ingresos se puede notar que la distribución sigue una tendencia normal con una media aproximada de $50,000.
- Existe una mínima cantidad de valores atípicos, pero se asume que en el contexto real pueden existir estas brechas.

2 | **C_Carnes** (consumo)

Figura 4. Histograma de la variable “C_Carnes”

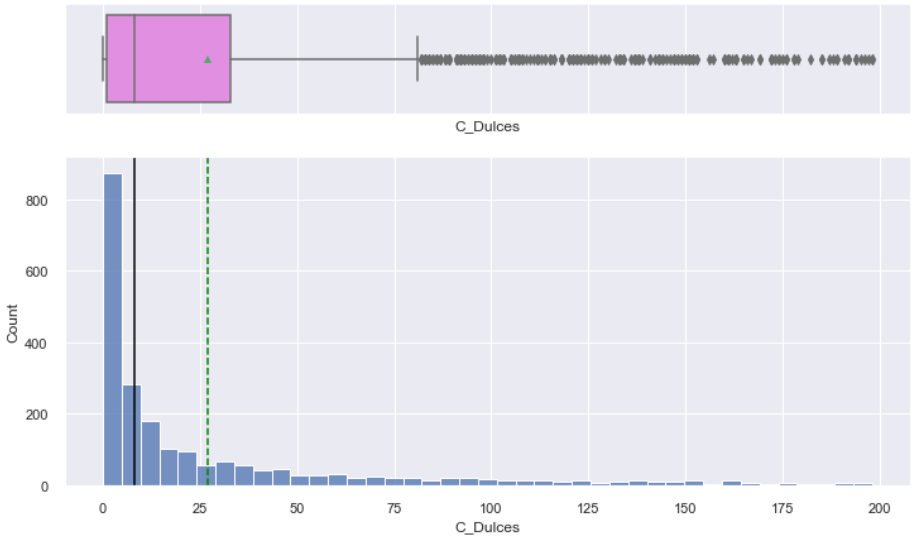


Fuente: *Elaboración propia*

- Se muestra la distribución del consumo de productos cárnicos de parte de los consumidores en el estudio.

3 | **C_Dulces** (consumo)

Figura 5. Histograma de la variable “C_Dulces”

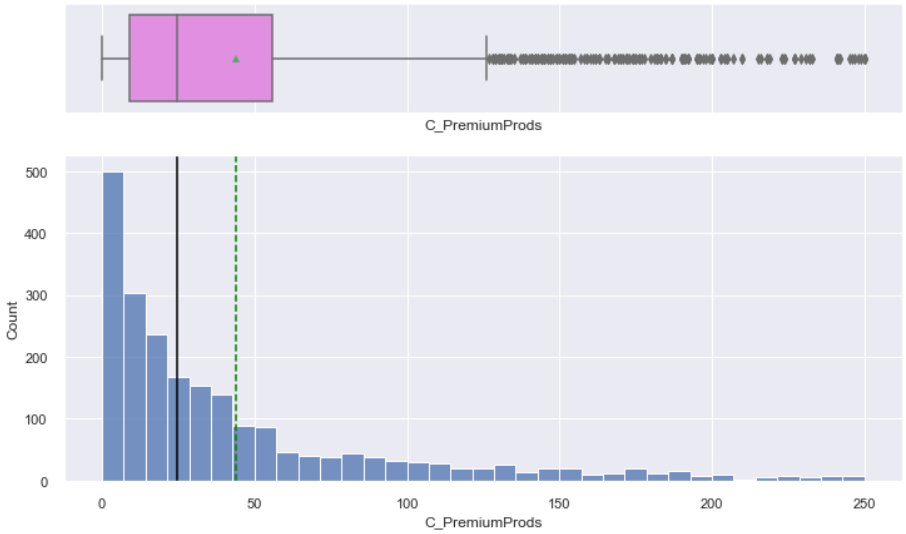


Fuente: *Elaboración propia*

- Se muestra la distribución del consumo de productos de la categoría “Dulces” en la base de datos.

4 | **C_PremiumProds** (consumo)

Figura 6. Histograma de la variable “C_PremiumProds”

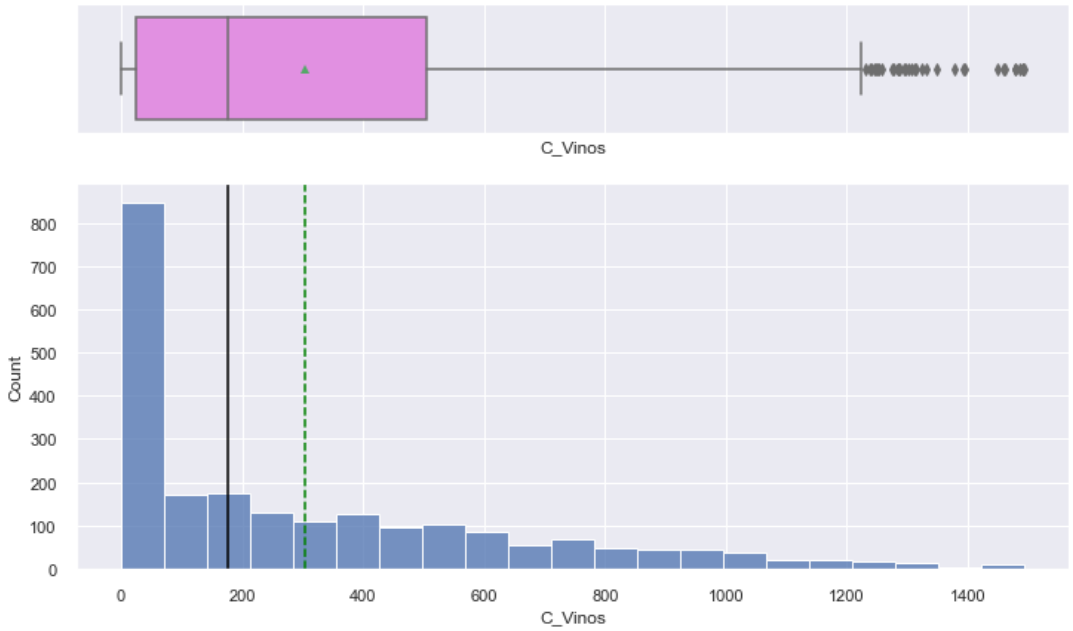


Fuente: *Elaboración propia*

- Se muestra la distribución del consumo de productos de la categoría de Productos premium en la base de datos. Se observa también un sesgo a la izquierda con valores que siguen una distribución típica para este tipo de productos.

5 | **C_Vinos** (consumo)

Figura 7. Histograma de la variable “C_Vinos”

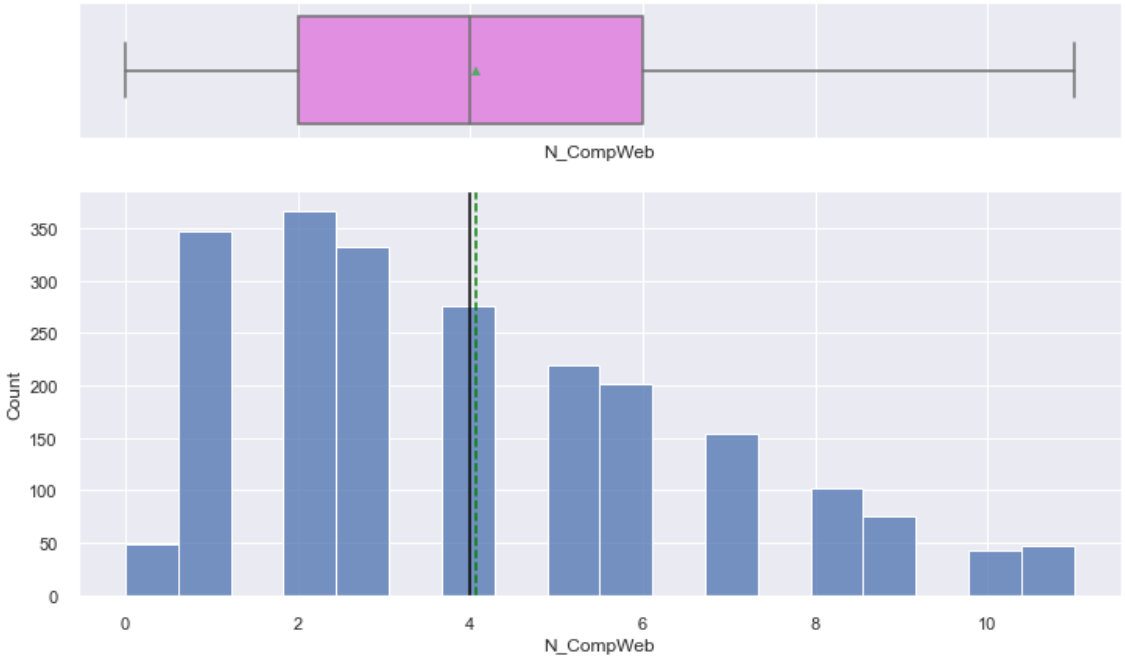


Fuente: Elaboración propia

- De la gráfica se puede observar que la distribución se encuentra sesgada a la derecha.
- La mediana de la distribución es menor a 200, más del 50% de clientes han gastado menos de 200 en vinos.
- Hay algunos datos que pudiesen ser atípicos en la derecha, pero no son tan pronunciados considerando que para este caso pudiera ser un escenario del contexto real, así que se los mantiene.

6 | N_CompWeb

Figura 8. Histograma de la variable “N_CompWeb”

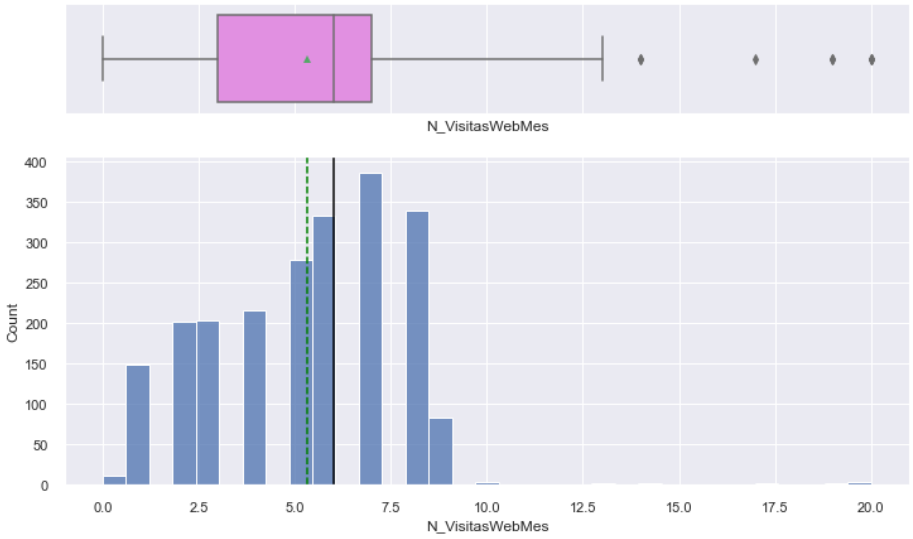


Fuente: *Elaboración propia*

- De la gráfica se puede observar la distibución del número de compras web por parte de los clientes. Se observa también una mediana de 4 compras por consumidor.

7 | N_VisitasWebMes

Figura 9. Histograma de la variable “N_VisitasWebMes”



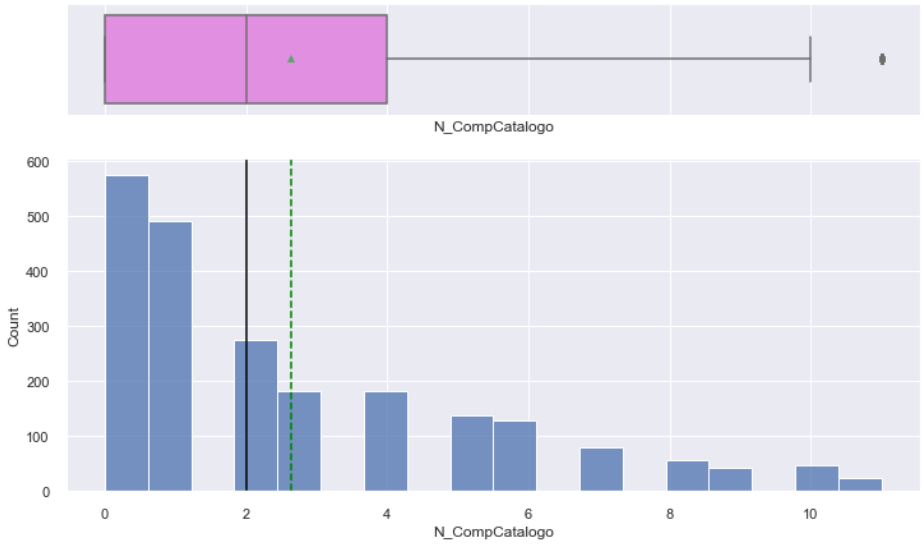
Fuente: *Elaboración propia*

- La distribución para el número de visitas en el mes por vía Web está sesgada y tiene algunos puntos atípicos en la derecha.

- No se tratará esto puesto que puede representar una tendencia real de mercado.

8 | N_CompCatalogo

Figura 10. Histograma de la variable “N_CompCatalogo”



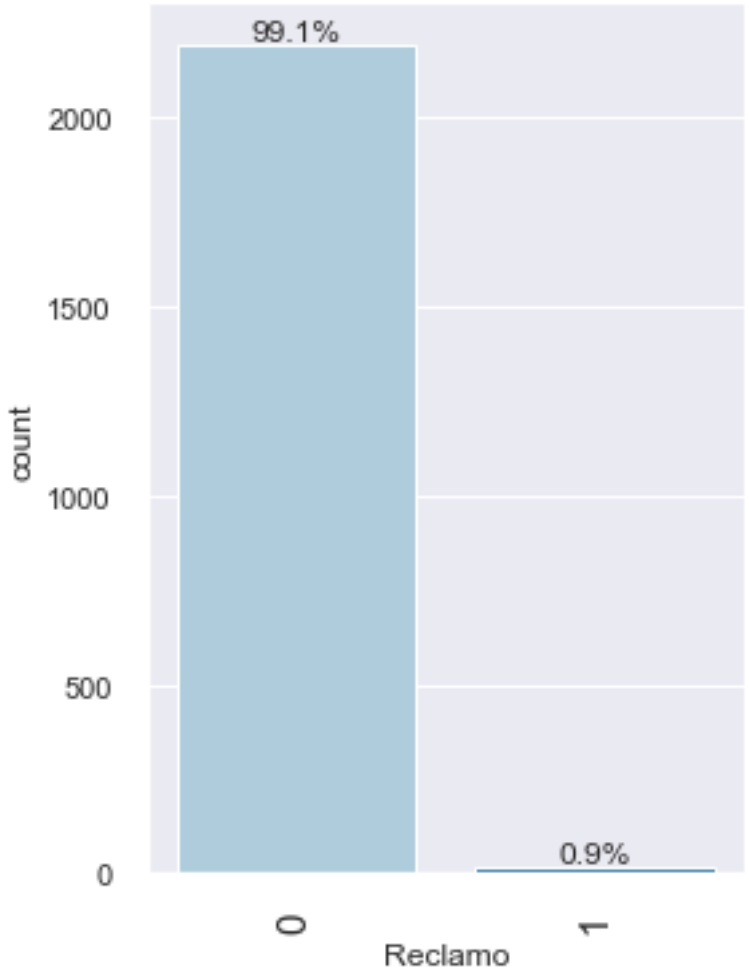
Fuente: Elaboración propia

- Se puede observar que la mayoría de observaciones son 0 para las compras por catálogo.

- La mediana de la distribución es 2. El 50% de los clientes tienen 2 o menos compras de catálogo.

9 │**Reclamo**

Figura 11. Histograma de la variable “Reclamo”

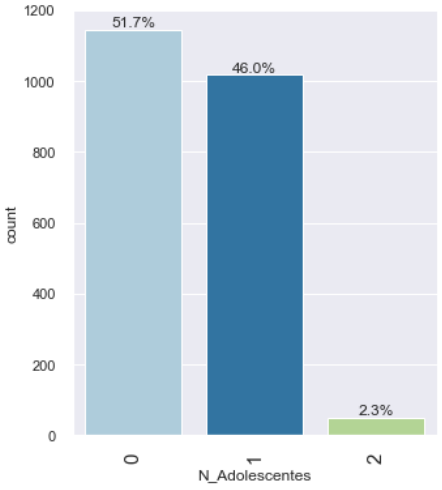


Fuente: *Elaboración propia*

- Cerca de 99% de los clientes no han tenido reclamos en el periodo de la base de datos (2 años). Esto puede ser porque la empresa tiene buen servicio al cliente o porque los clientes no tienen una manera certera de hacer llegar estos reclamos.

10 | N_Adolescentes

Figura 12. Histograma de la variable “N_Adolescentes”

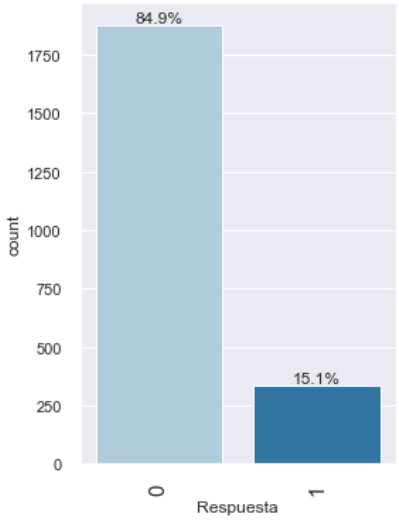


Fuente: *Elaboración propia*

- La mayoría de los clientes no tienen un adolescente en casa.
- Hay muy pocos clientes (2.3%) que tienen más de 1 adolescente en casa.

11 | Respuesta

Figura 13. Histograma de la variable “Respuesta”

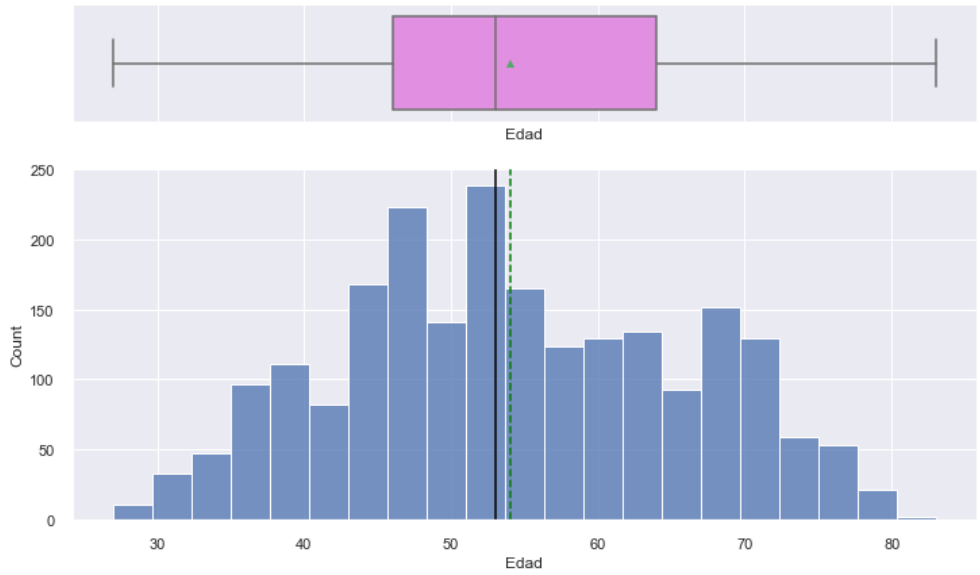


Fuente: *Elaboración propia*

- Cerca del 85% de los clientes respondieron NO a la última campaña.
- Esto indica que la distribución de las clases en el objetivo es no balanceada.
- Solo 15% de los clientes respondieron SI a la última campaña.

12 | **Edad**

Figura 14. Histograma de la variable “Edad”



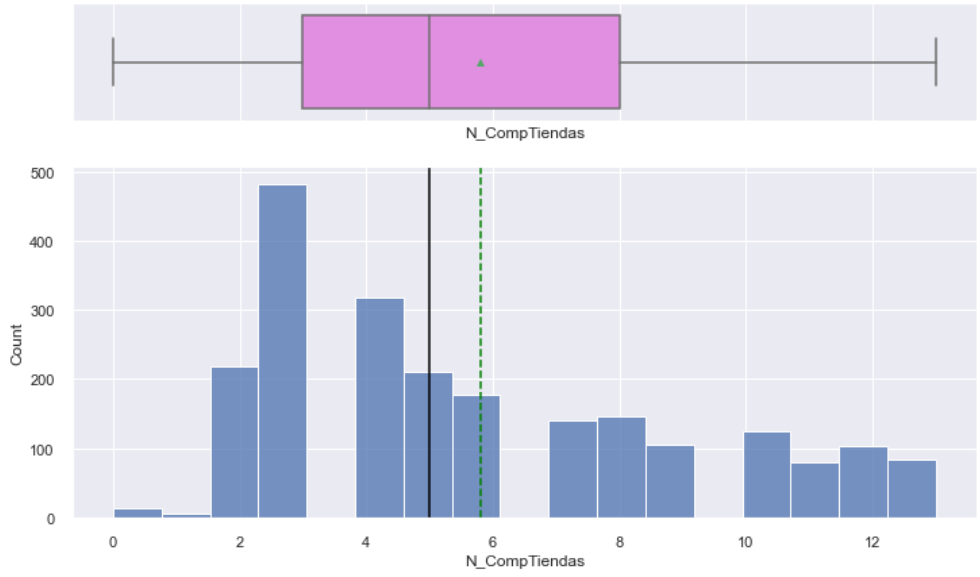
Fuente: *Elaboración propia*

- Se puede observar que no hay valores atípicos en el boxplot.

- La edad tiene una distribución aparentemente normal con mediana y media bastantes iguales.

13 | **N_CompTiendas**

Figura 15. Histograma de la variable “N_CompTiendas”



Fuente: *Elaboración propia*

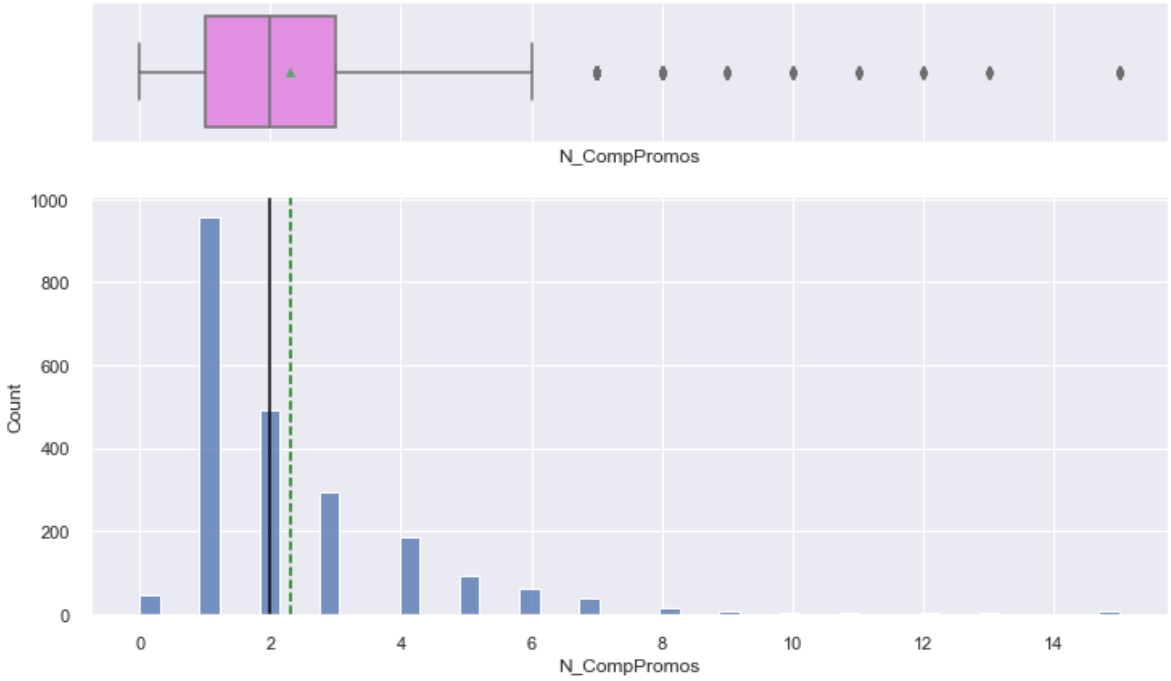
- Hay muy pocas observaciones con menos de 2 compras en la tienda.

- La mayoría de los clientes tienen 4 o 5 compras de la tienda.

- No hay valores atípicos en esta gráfica de la variable.

14 | N_CompPromos

Figura 16. Histograma de la variable “N_CompPromos”



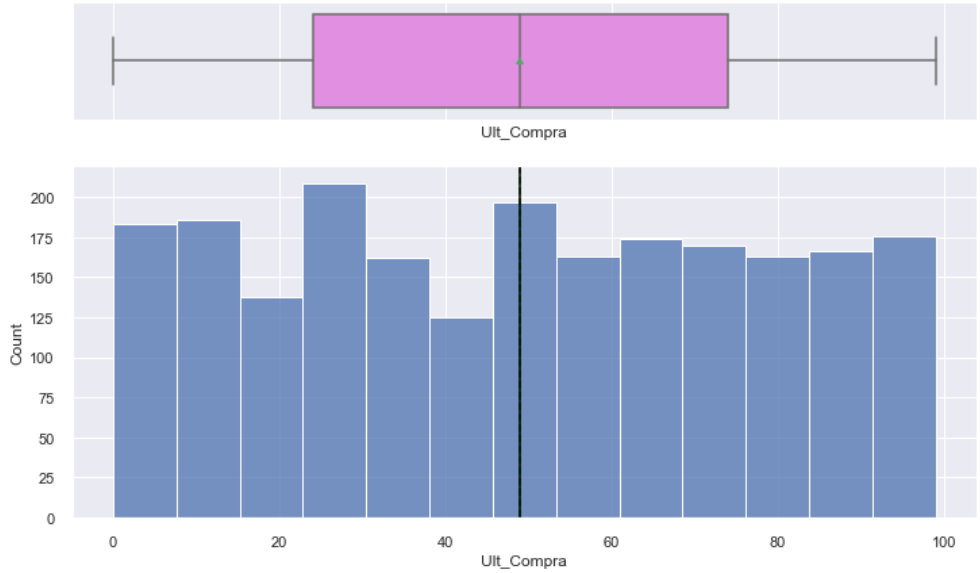
Fuente: *Elaboración propia*

- La mayoría de clientes tienen 2 o menos compras con descuentos promocionales.

- Podemos ver que existen algunos valores extremos en la derecha de la distribución. Para la presente variable esto puede ser una tendencia de mercado.

15 | **Ult_Compra**

Figura 17. Histograma de la variable “Ult_Compra”



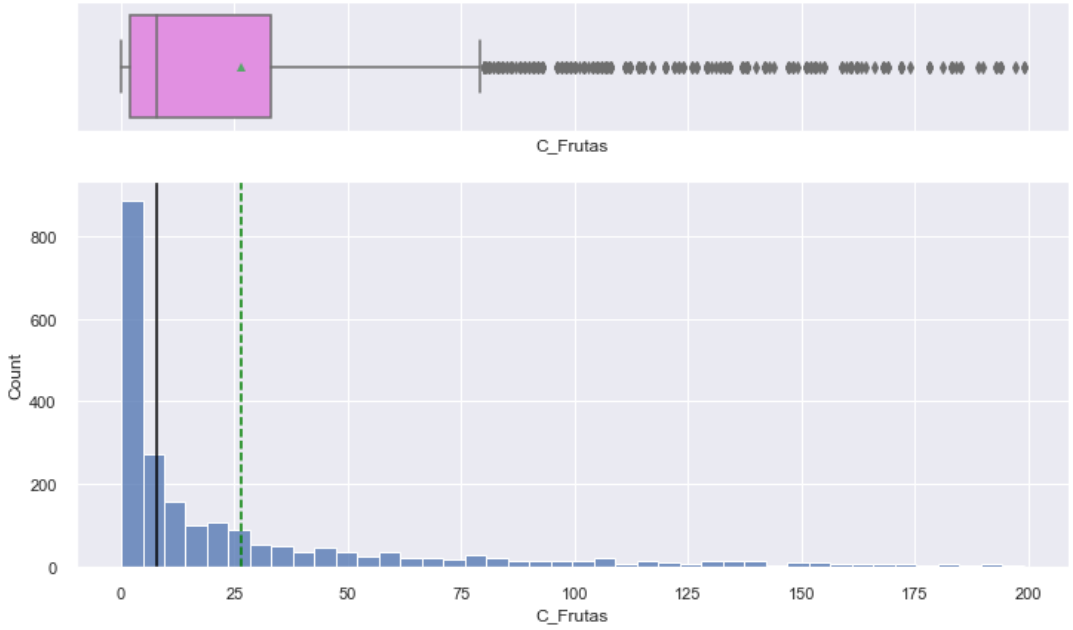
Fuente: *Elaboración propia*

- No hay valores extremos atípicos en la distribución de esta variable.

- La distribuución es bastante siméttrica y uniformemente distribuída.

16 | **C_Frutas** (Consumo)

Figura 18. Histograma de la variable “C_Frutas”



Fuente: Elaboración propia

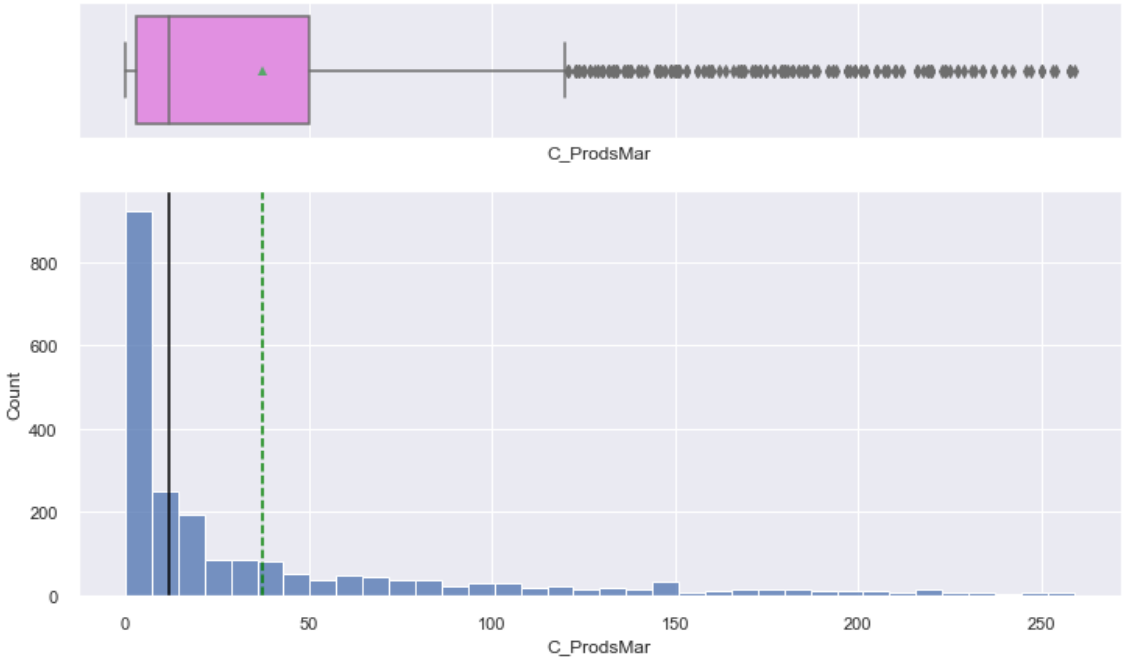
- La distribución para "C_Frutas" esta sesgada a la derecha.

- Como la mediana de la distribución es menor a 20, más del 50% de clientes han gastado menos de 20 en frutas.

- Existen algunos valores extremos a la derecha de la gráfica, pero no se las tratará porque esto puede ser un escenario de la vida real en cuanto a gastos de dinero.

17 | **C_ProdsMar** (Consumo)

Figura 19. Histograma de la variable “C_ProdsMar”



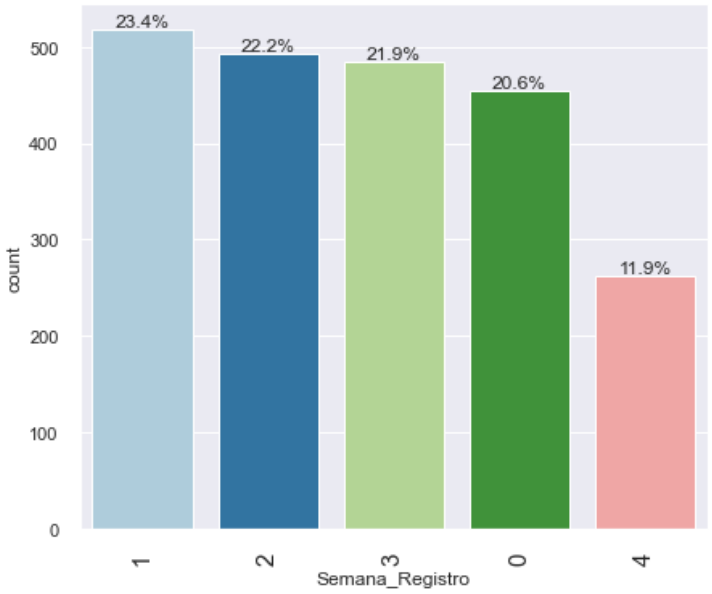
Fuente: *Elaboración propia*

- La distribución para la cantidad de dinero gastada en productos del mar está sesgada a la derecha.

- Existen algunos valores extremos a la derecha, pero no se los tratará debido a que esto representa una tendencia de mercado real.

18 | Semana_Registro

Figura 20. Histograma de la variable “Semana_Registro”

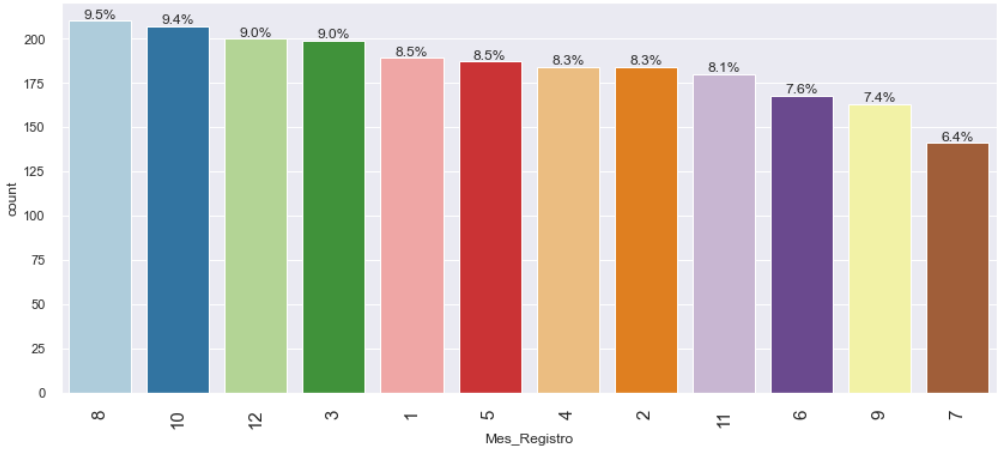


Fuente: *Elaboración propia*

- La gráfica muestra que el número de registros se reduce en el fin de mes. Esto puede ser debido a que la gente cobra a principio de mes usualmente.

19 | Mes_Registro

Figura 21. Histograma de la variable “Mes_Registro”

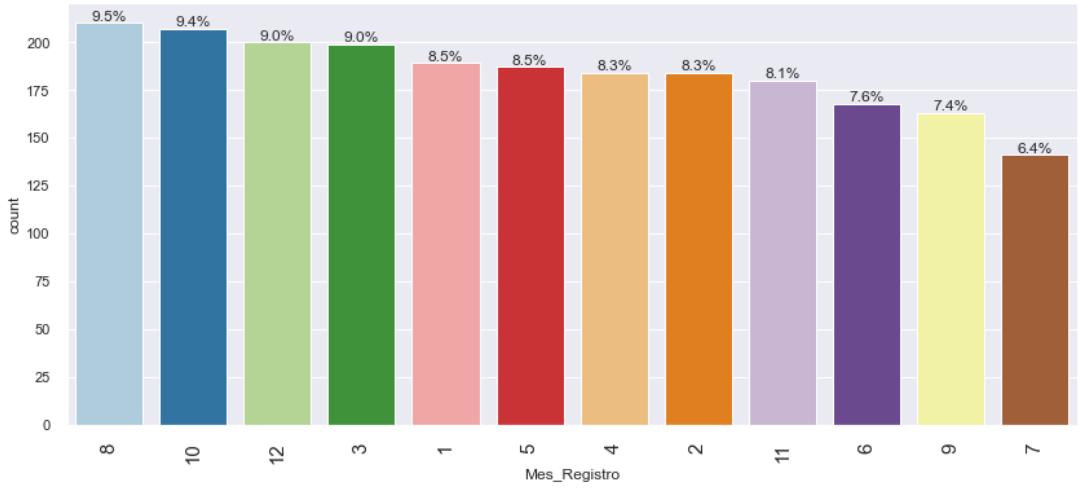


Fuente: *Elaboración propia*

- La gráfica muestra que el número más grande de registros se da en los meses de Agosto, Octubre, Diciembre, Marzo.
- Desde el mes de Junio los registros decrecen.

20 | **Trimestre_Registro**

Figura 22. Histograma de la variable “Trimestre_Registro”

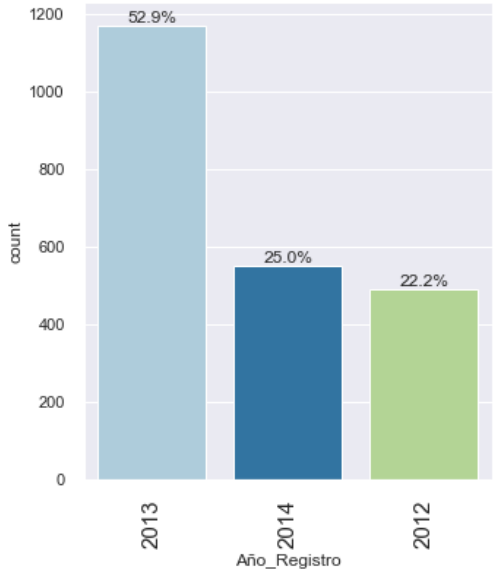


Fuente: *Elaboración propia*

- No hay diferencia notoria entre los registros entre trimestres.
- El número de registros es solamente un poco mayor en los trimestres 1 y 4. Esto puede ser debido a festividades de estos períodos.

21 | **Año_Registro**

Figura 23. Histograma de la variable “Año_Registro”

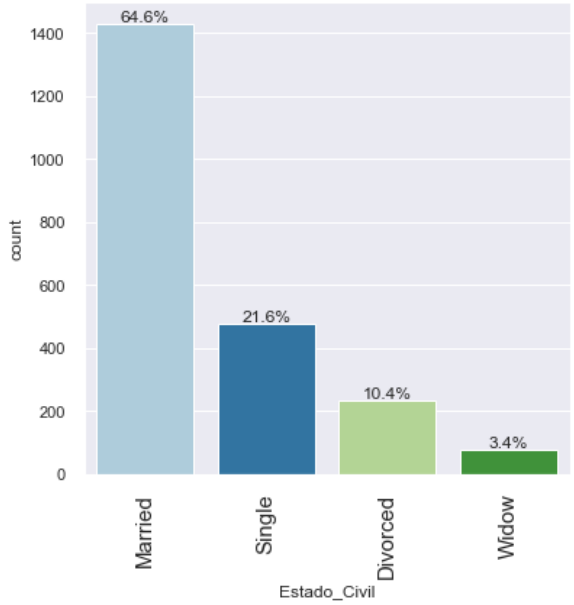


Fuente: *Elaboración propia*

-El mayor número de clientes registrados se dió en el año 2013.

22 | **Estado_Civil**

Figura 24. Histograma de la variable “Estado_Civil”

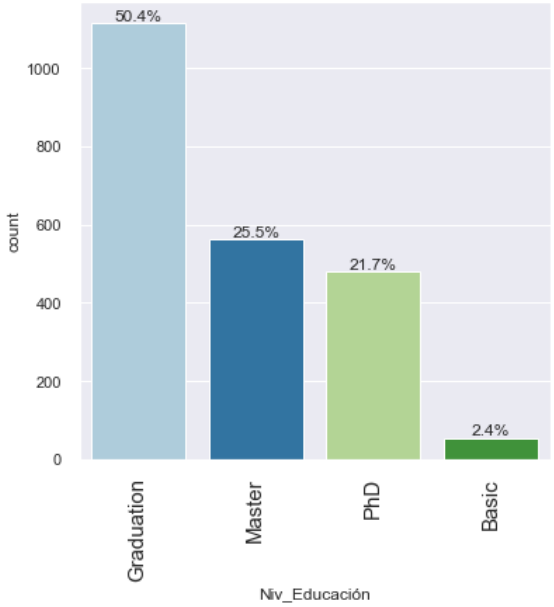


Fuente: *Elaboración propia*

• La mayoría de clientes están casados o para el objetivo del presente estudio unidos. Esto representa alrededor de 64% de los clientes.

23 | **NIv_Educación**

Figura 25. Histograma de la variable “Niv_Educación”



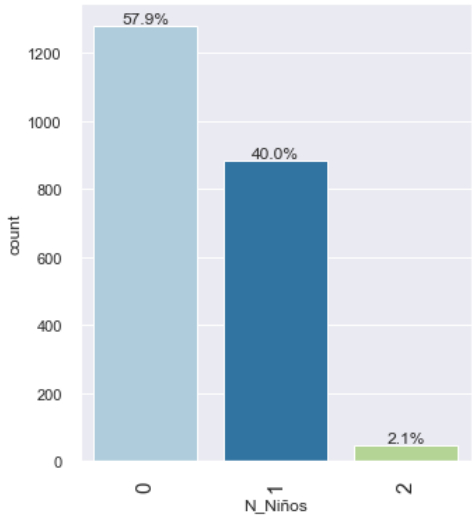
Fuente: Elaboración propia

- El nivel de educación académica de cerca de 50% de la población del estudio es de Tercer Nivel (Graduation)

- Solamente el 2.4% de personas con educación básica.

24 | **N_NIños**

Figura 26. Histograma de la variable “N_Niños”

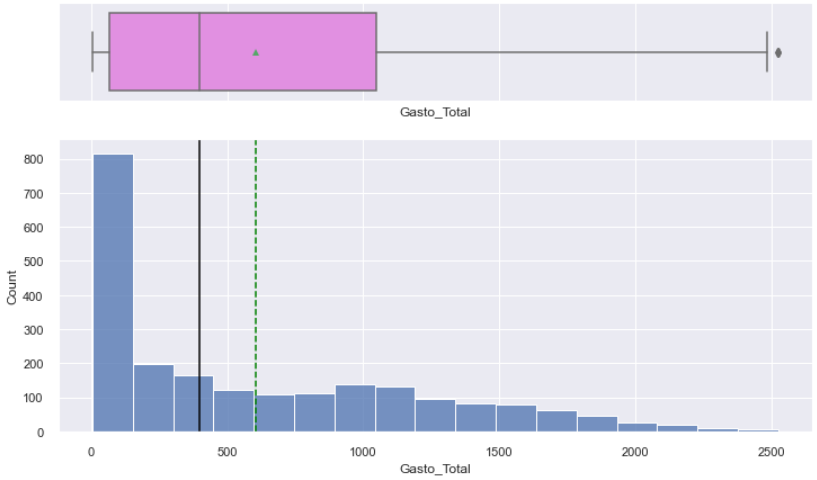


Fuente: *Elaboración propia*

- El 40% de los clientes tienen 1 niño. El 57.9% no tiene niños en casa.
- Hay un porcentaje pequeño de clientes (2.1%) que tiene más de un niño en casa.

25 | **Gasto_Total**

Figura 27. Histograma de la variable “Gasto_Total”



Fuente: *Elaboración propia*

- Se indica la gráfica del gasto total de los clientes, considerando la sumatoria de los gastos por categoría.

7.5.2. ANALISIS BI-VARIABLE

Con el objetivo de mostrar la relación entre variables y su comportamiento frente a la variable de respuesta se presenta un análisis bivariado que permite mostrar esta interacción.

VARIABLES DE CONSUMO VS VARIABLE DE RESPUESTA

Figura 28. C_Carnes (count)

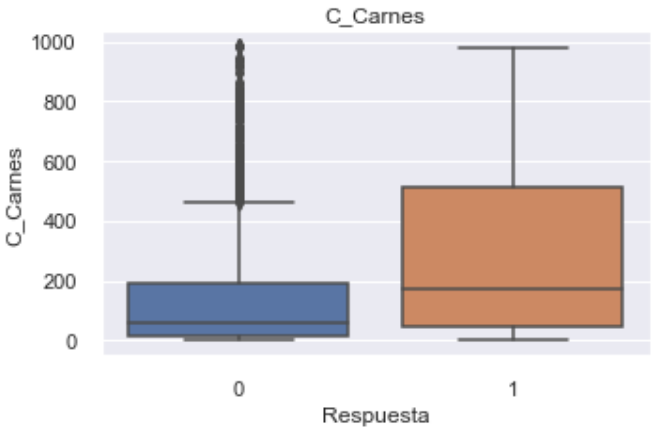


Figura 29. C_Frutas (count)

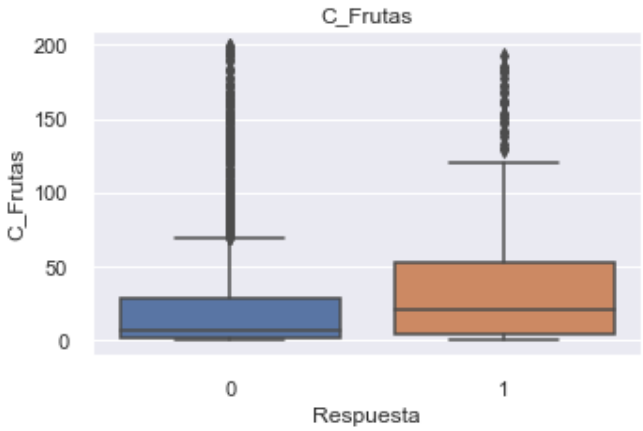


Figura 30. C_ProdsMar (count)

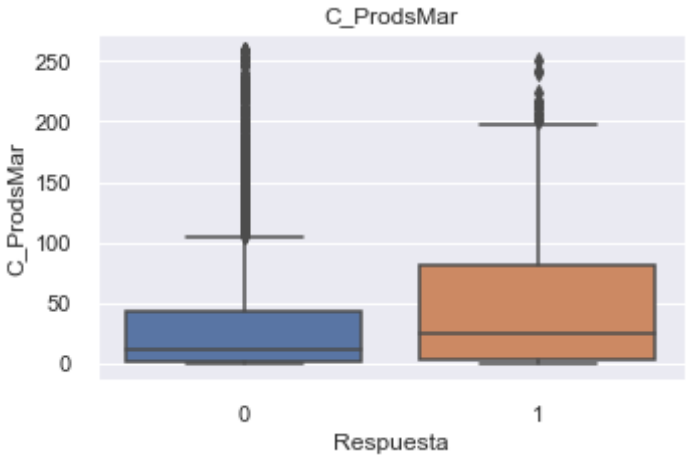


Figura 31. C_Dulces (count)

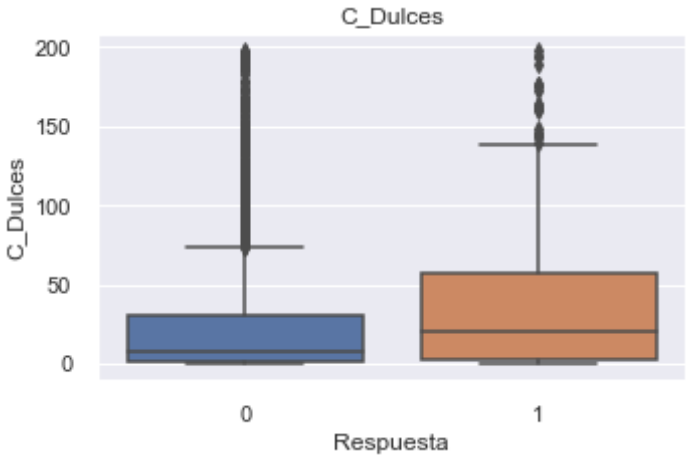


Figura 32. C_Vinos (count)

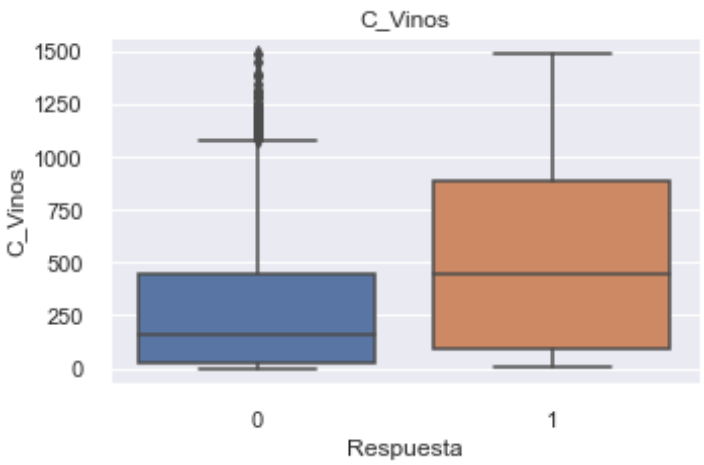
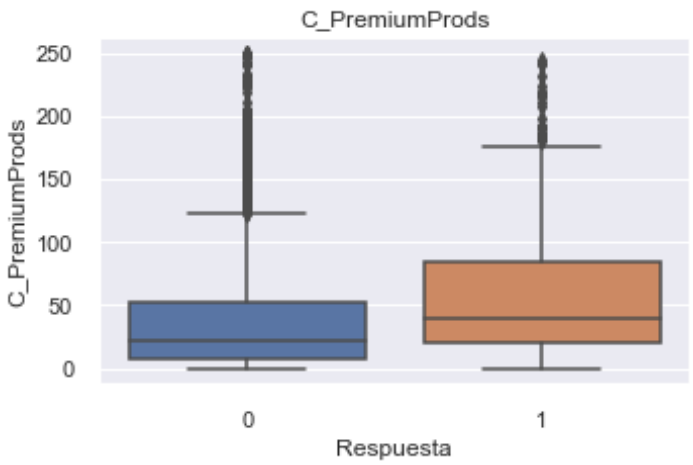


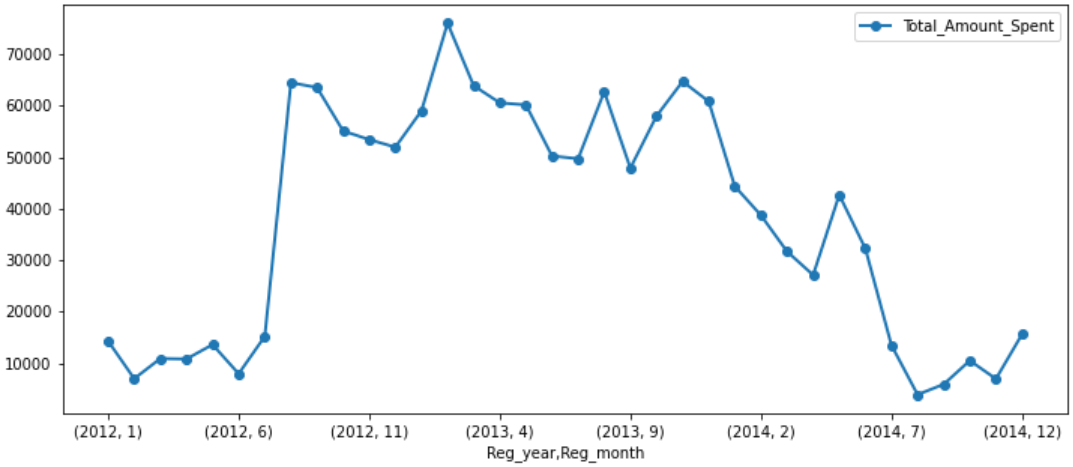
Figura 33. C_PremiumProds (count)



Fuente: *Elaboración propia*

- Cada gráfico muestra que el cliente que gasta más en cualquier producto tiene más probabilidades de aceptar la oferta y tener una respuesta positiva a la campaña.

Figura 34. Variables de año y mes de registro vs variable de respuesta



Fuente: *Elaboración propia*

- El gráfico muestra claramente que la cantidad total gastada ha disminuido a lo largo de los años.
- El gráfico muestra el mayor aumento en la cantidad gastada de agosto a septiembre de 2012.

VARIABLES CATEGÓRICAS (DESCRIPCIÓN Y CONTEO)

Figura 35. Niv_Educación (count)

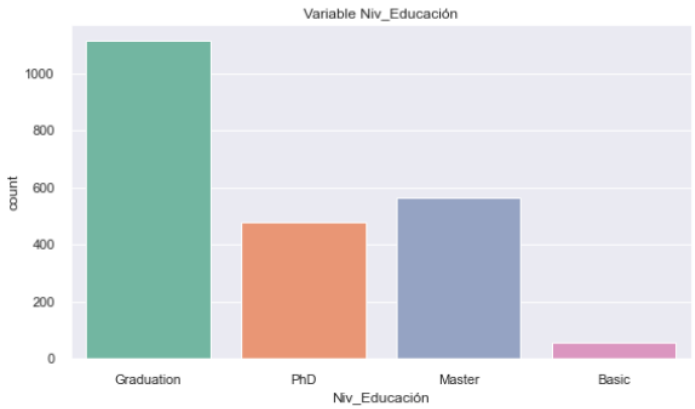


Figura 36. Estado_Civil (count)

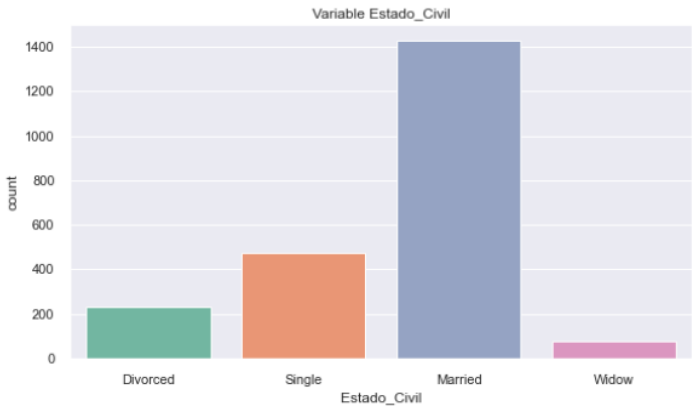


Figura 37. N_Niños (count)

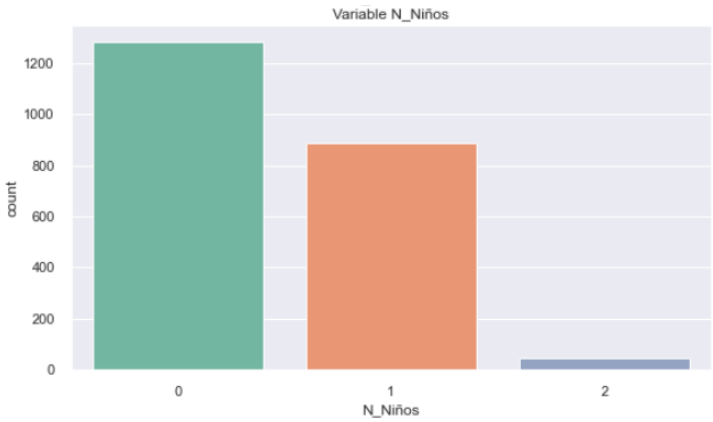


Figura 38. N_Adolescentes (count)

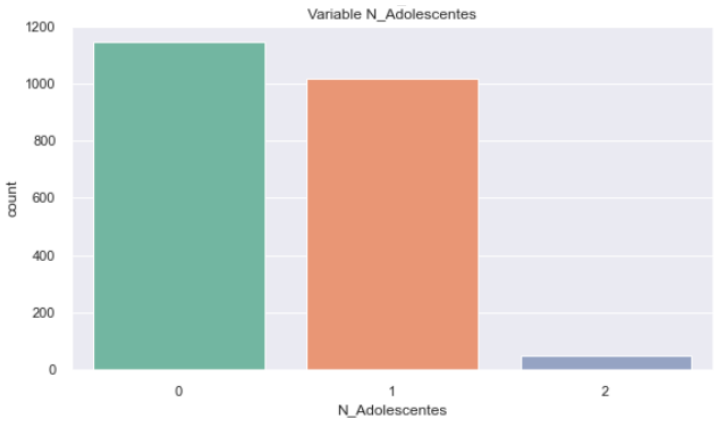


Figura 39. N_CompTiendas (count)

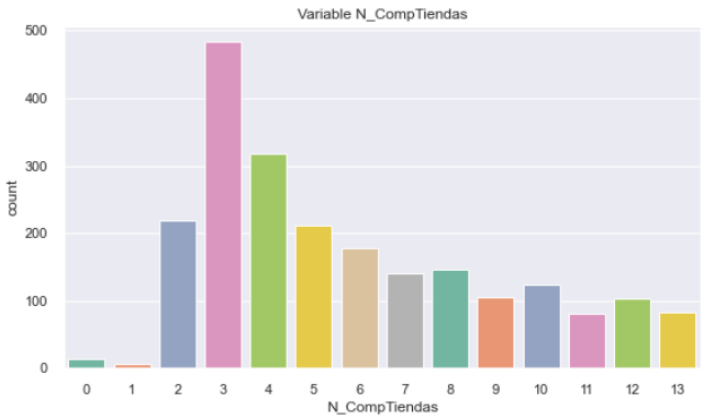


Figura 40. N_CompPromos (count)

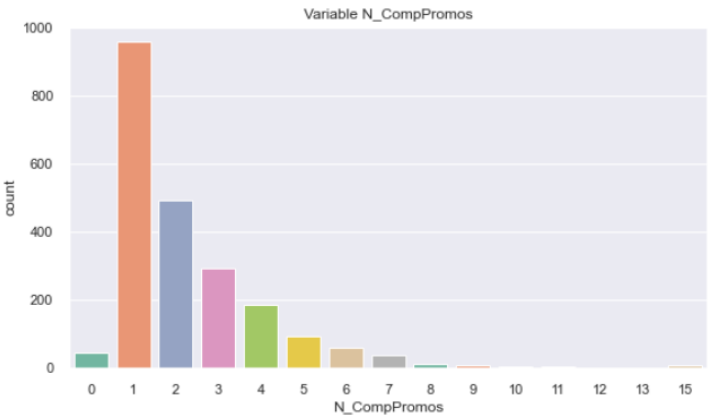


Figura 41. N_CompWeb (count)

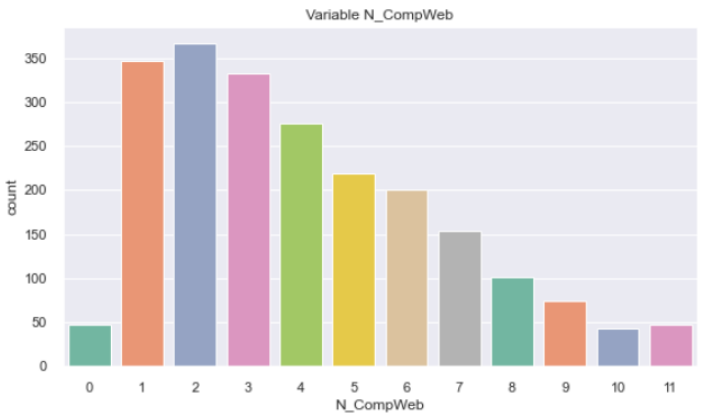


Figura 42. N_CompCatalogo (count)

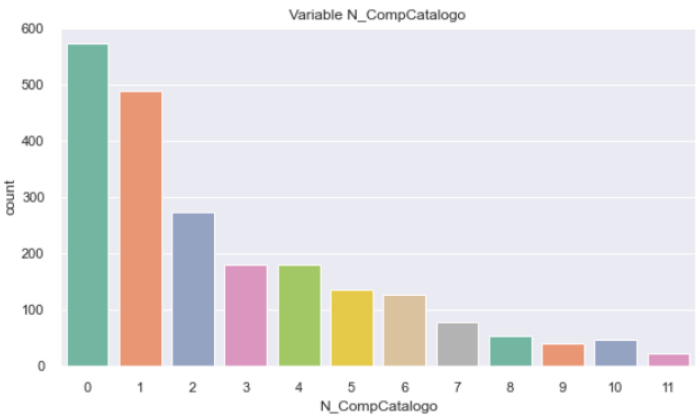


Figura 43. N_VisitasWebMes (count)

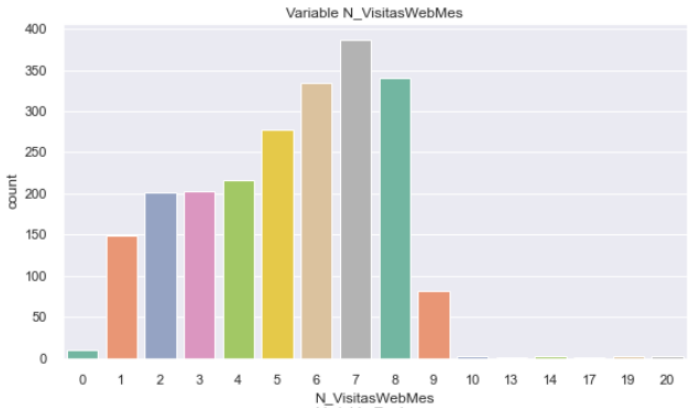
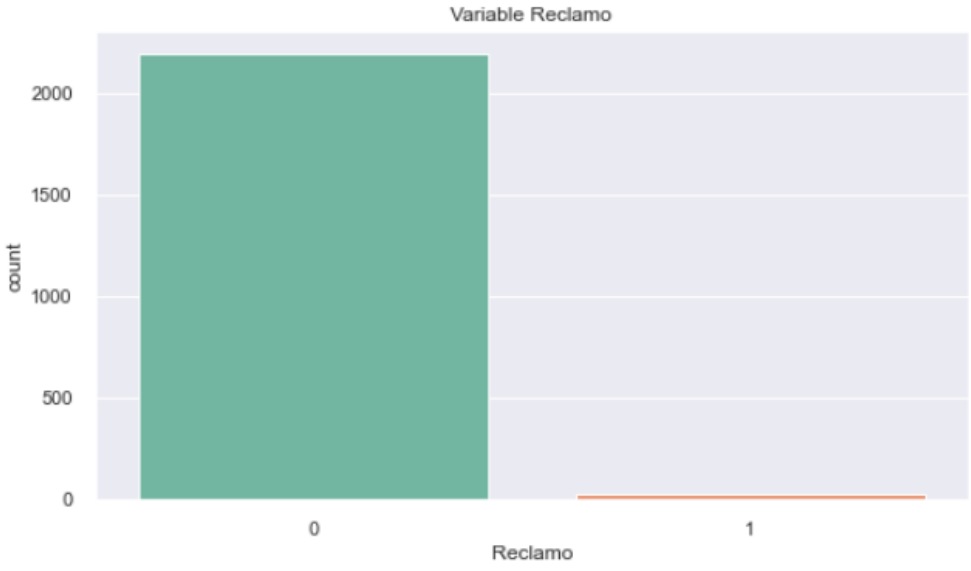


Figura 44. Variable de Respuesta (count)

Figura 45. Reclamo (count)



ANÁLISIS VARIABLES CATEGORICAS POR TIPO DE RESPUESTA

Figura 46. Niv_Educación por Respuesta

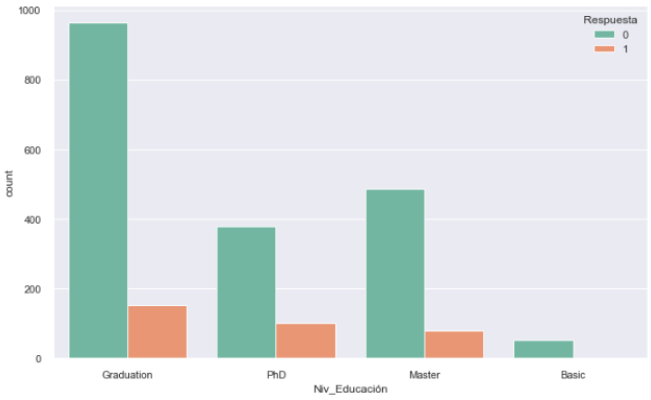


Figura 47. Estado_Civil por respuesta

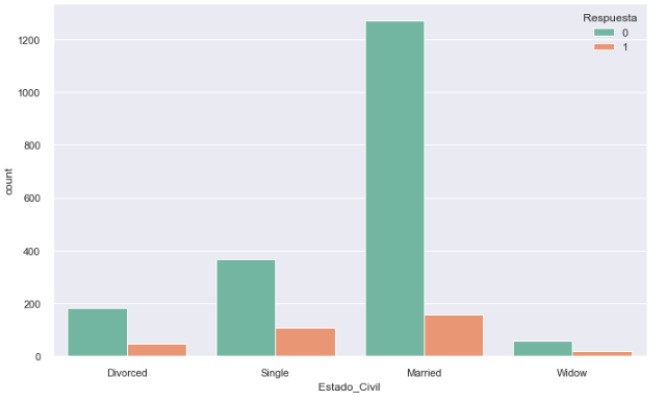


Figura 48. N_Niños por Respuesta

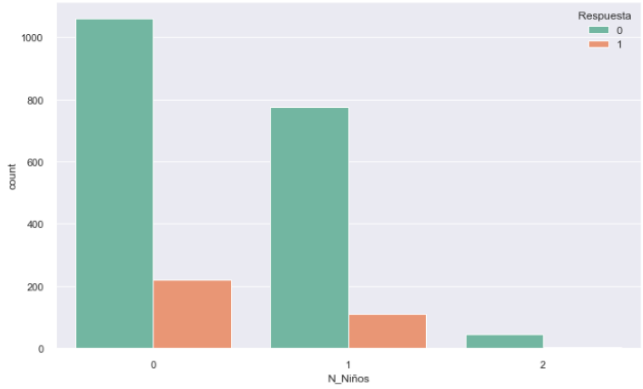


Figura 49. N_Adolescentes por respuesta

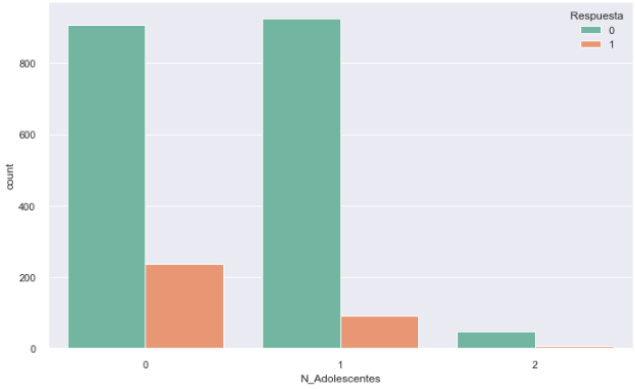


Figura 50. N_CompPromos por respuesta

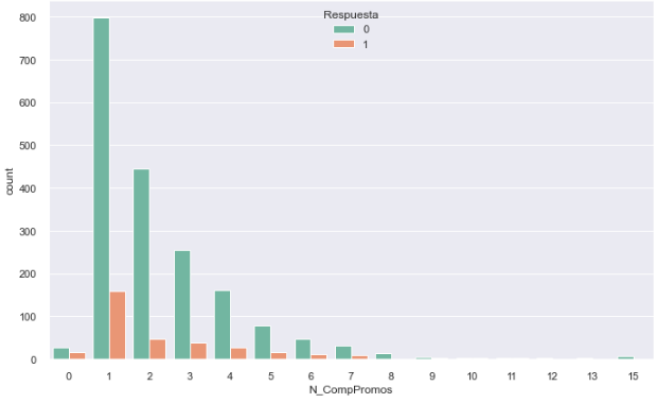


Figura 51. N_CompWeb por respuesta

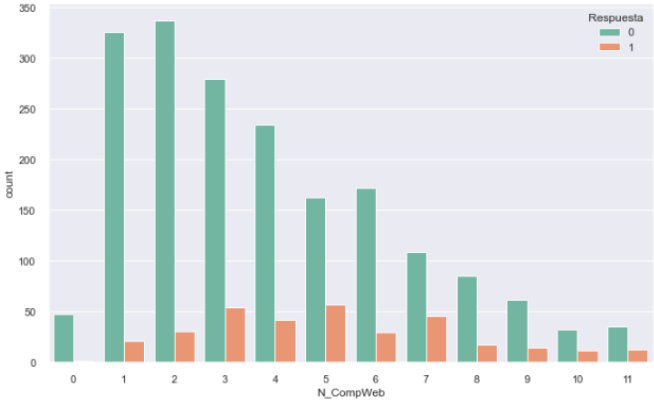


Figura 52. N_CompCatalogo por respuesta

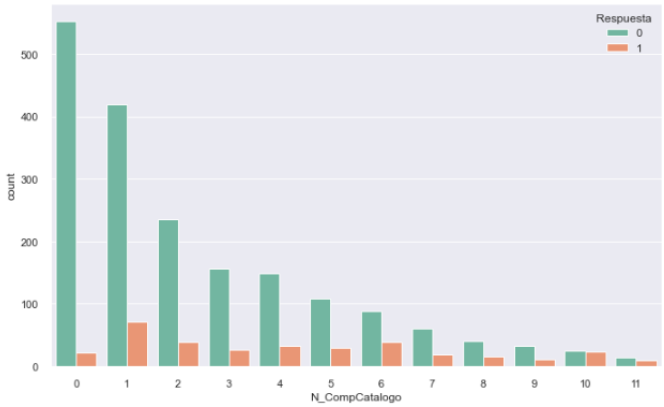


Figura 53. N_CompTiendas por respuesta

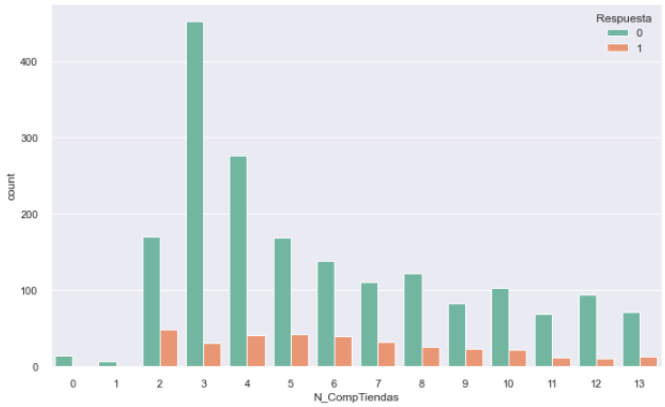


Figura 54. N_VisitasWebMes por respuesta

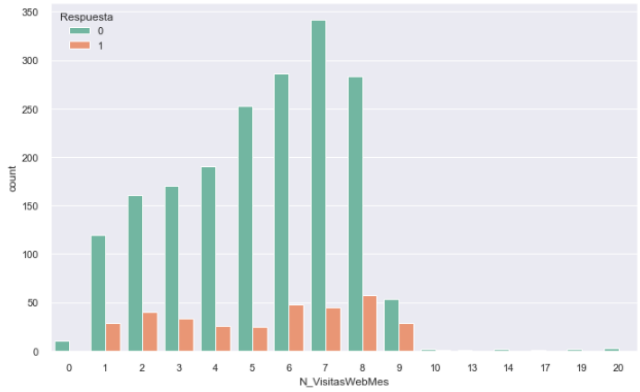


Figura 55. Reclamos por respuesta

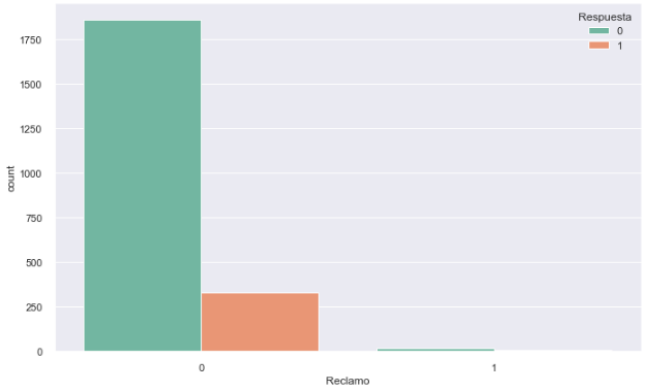
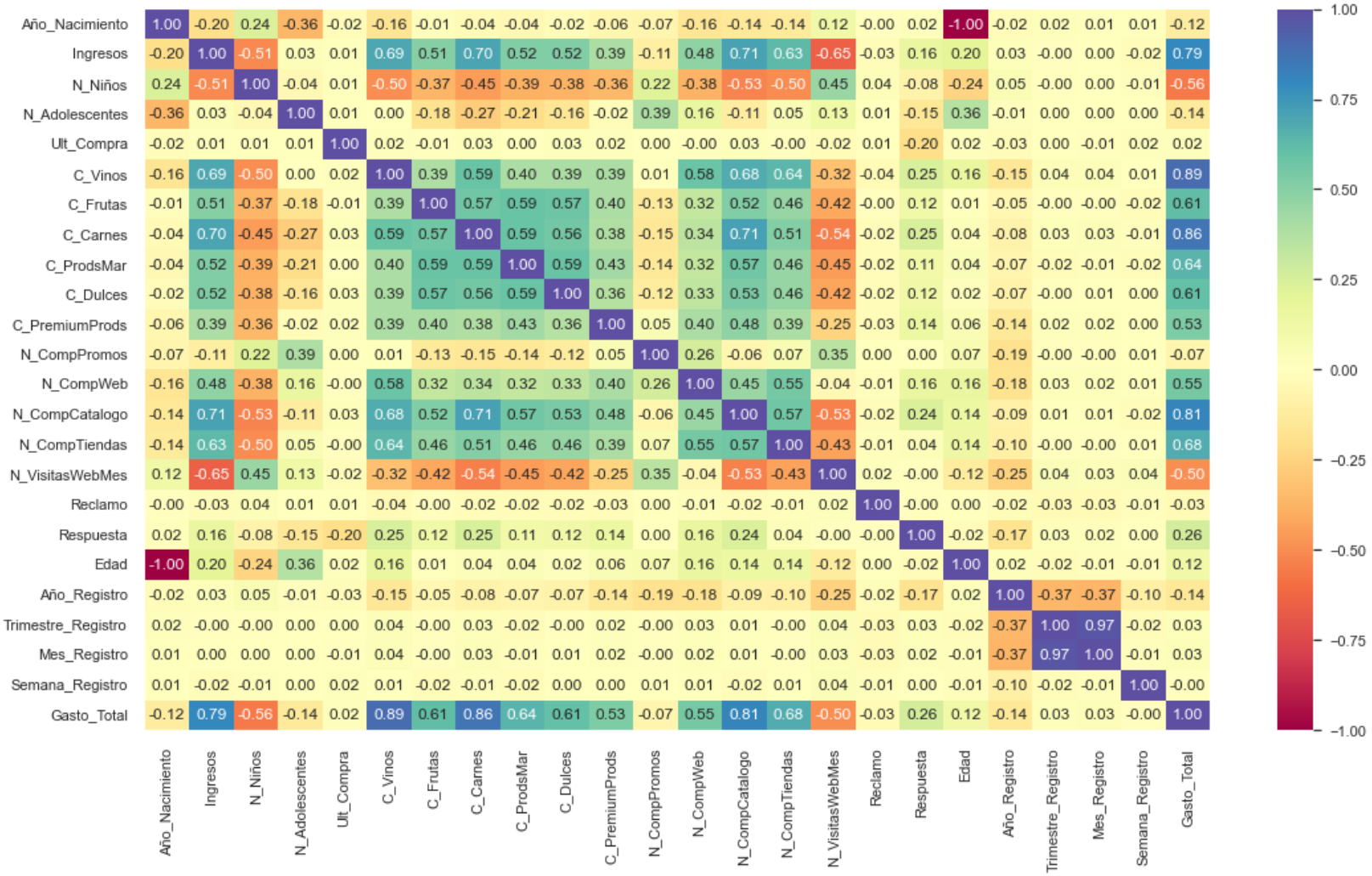


Figura 56. Correlación entre variables del estudio



Fuente: *Elaboración propia*

8. RESULTADOS

Después de haber detallado las variables que se toman en cuenta en el estudio con respecto a sus características demográficas, se procede al modelado de los mismos utilizando los modelos de regresión logística, árbol de decisión, random forest y extra trees. La estrategia consiste en poder contrastar los resultados de cada uno de estos modelos y seleccionar al que muestre el mejor indicador y su eficiencia.

8.1 MATRICES DE CONFUSION

Las matrices de confusión son una herramienta que permite analizar los resultados de la implementación de un algoritmo de aprendizaje supervisado. En forma de matriz, se presentan columnas donde aparecen valores del número de predicciones por cada clase, mientras que cada fila muestra el número real de las instancias de cada clase. Esto permite observar el desempeño que ha tenido cada modelo en base a los errores y aciertos en las corridas respectivas de cada caso.

8.2 MÉTRICAS DE LA MATRIZ DE CONFUSIÓN

Exactitud. Representa el porcentaje de predicciones correctas frente a la totalidad de registros. Es útil cuando en la variable de respuesta se necesita mirar si las predicciones tanto positivas como negativas han sido acertadas correctamente.

Precisión. Denota cuan cerca está el resultado del valor verdadero. También se puede mencionar como la división entre los casos positivos bien identificados y el total de predicciones positivas.

Sensibilidad. Representa la tasa de verdaderos positivos. Es decir, es la proporción de casos positivos bien clasificados por el modelo respecto al total de positivos.

Figura 57. Interpretación de matrices de confusión

Matriz de confusión		Estimado por el modelo			
		Negativo (N)	*Positivo (P)*		
Real	*Negativo*	a: (TN)	b: (FP)		
	Positivo	c: (FN)	d: (TP)	**Precisión** ***("precision")*** Porcentaje predicciones positivas correctas:	d/(b+d)
		Sensibilidad, exhaustividad ***("Recall")*** Porcentaje casos positivos detectados	**Especifidad** ***(Specifity)*** Porcentaje casos negativos detectados	**Exactitud ("accuracy")** Porcentaje de predicciones correctas *(No sirve en datasets poco equilibrados)*	
		d/(d+c)	a/(a+b)	(a+d)/(a+b+c+d)	

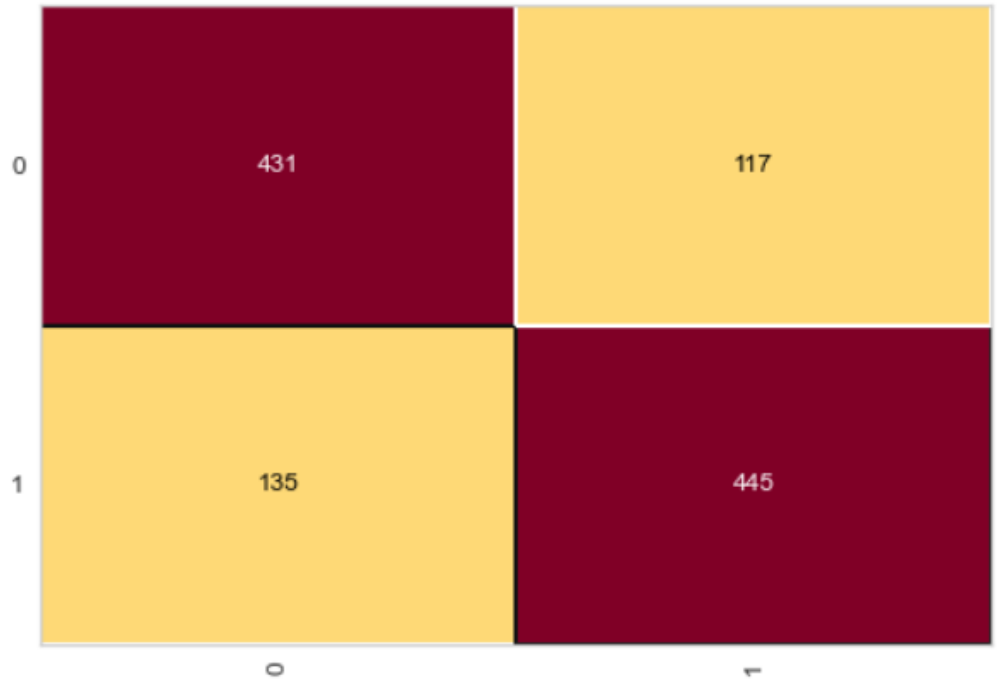
Fuente: *Telefónica Tech / Elaboración propia*

8.3 RESULTADOS DE PREDICCIÓN BINARIA

8.3.1 Regresión Logística

Los resultados del modelo de regresión logística resultan en la siguiente matriz de confusión y los siguientes valores.

Figura 58. Matriz de confusión Modelo de regresión Logística



Fuente: Elaboración propia / Python

Tabla 2. Resultados del Modelo de Regresión Logística

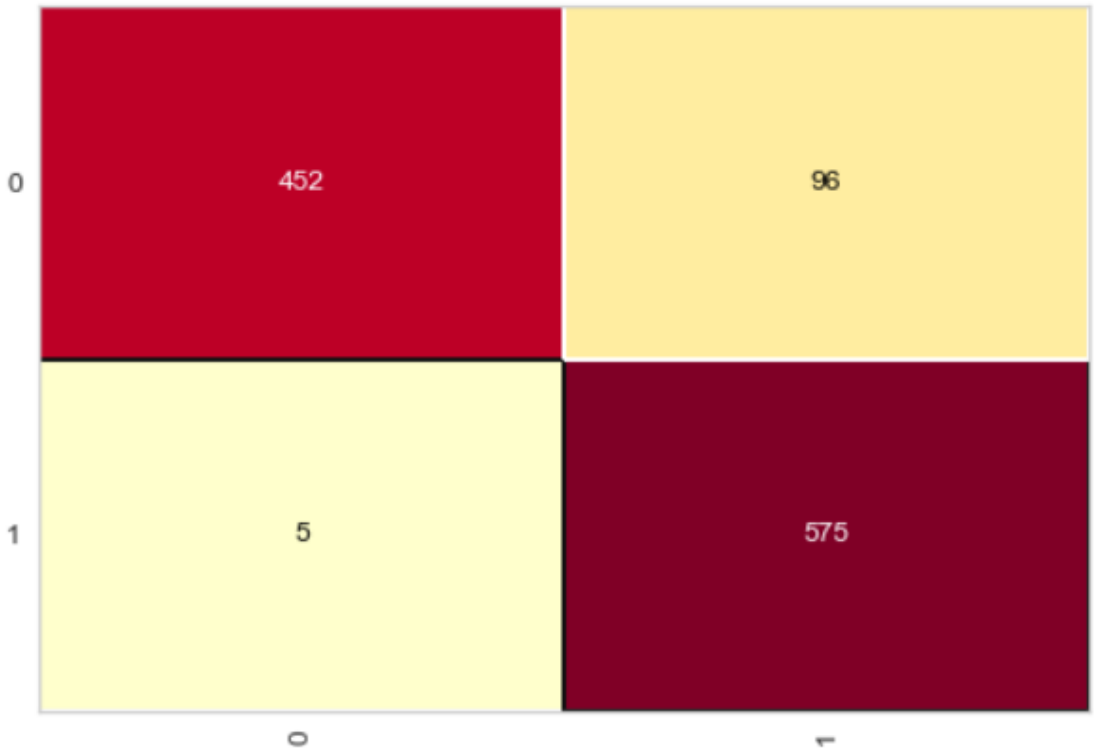
	precision	recall	f1-score	support
0	0.76	0.79	0.77	548
1	0.79	0.77	0.78	580
accuracy			0.78	1128
macro avg	0.78	0.78	0.78	1128
weighted avg	0.78	0.78	0.78	1128

Fuente: *Elaboración propia / Phyton*

8.3.2 Árbol de decisión

Los resultados del modelo de árbol de decisión resultan en la siguiente matriz de confusión y los siguientes valores.

Figura 59. Matriz de confusión Modelo de árbol de decisión



Fuente: *Elaboración propia / Python*

Tabla 3. Resultados del Modelo de árbol de decisión

	precision	recall	f1-score	support
0	0.99	0.82	0.90	548
1	0.86	0.99	0.92	580
accuracy			0.91	1128
macro avg	0.92	0.91	0.91	1128
weighted avg	0.92	0.91	0.91	1128

Fuente: *Elaboración propia / Phyton*

Tabla 4. Variables de más incidencia del Modelo de árbol de decisión

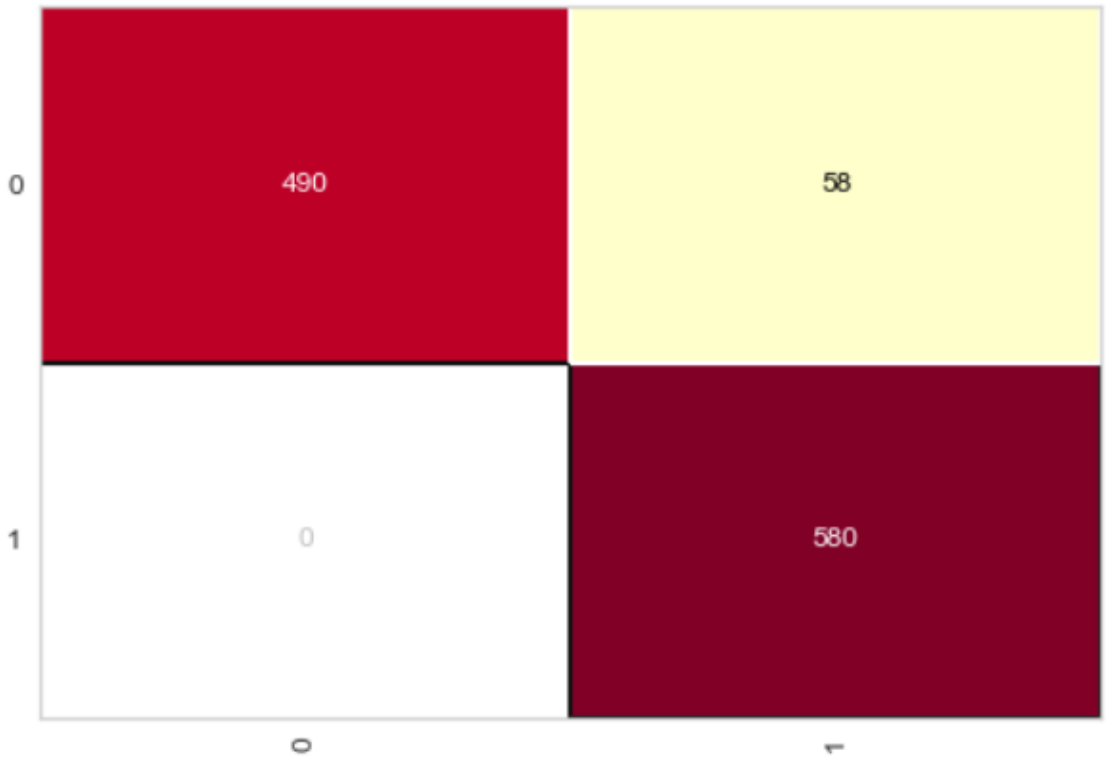
Gasto_Total	0.122121
Ult_Compra	0.110246
Ingresos	0.085185
N_CompCatalogo	0.081668
N_VisitasWebMes	0.074686
Año_Registro	0.056051
C_Carnes	0.046741
C_PremiumProds	0.046604
Mes_Registro	0.046265
C_ProdsMar	0.044952
C_Vinos	0.042751
N_CompTiendas	0.040069
Estado_Civil	0.036827
C_Frutas	0.033482
N_CompPromos	0.029870
Año_Nacimiento	0.027995
Edad	0.019933
C_Dulces	0.019093
N_CompWeb	0.008894
N_Adolescentes	0.007320
Niv_Educación	0.007162
Trimestre_Registro	0.004203
N_Niños	0.004065
Semana_Registro	0.003814
Reclamo	0.000000

Fuente: *Elaboración propia / Phyton*

8.3.3 Random Forest

Los resultados del modelo de random forest resultan en la siguiente matriz de confusión y los siguientes valores.

Figura 60. Matriz de confusión Modelo de random forest



Fuente: *Elaboración propia / Python*

Tabla 5. Resultados del Modelo de random forest

	precision	recall	f1-score	support
0	1.00	0.89	0.94	548
1	0.91	1.00	0.95	580
accuracy			0.95	1128
macro avg	0.95	0.95	0.95	1128
weighted avg	0.95	0.95	0.95	1128

Fuente: *Elaboración propia / Phyton*

Tabla 6. Variables de más incidencia del Modelo de random forest

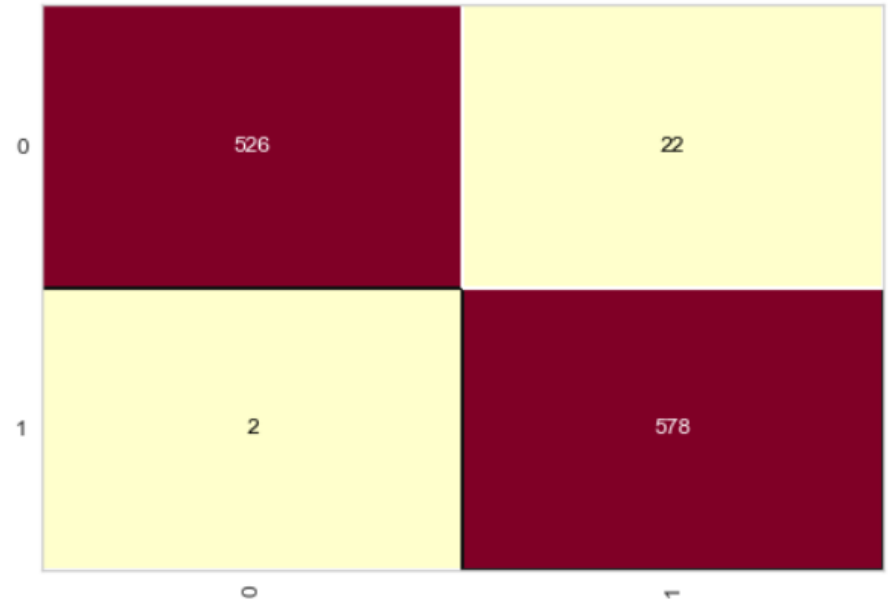
Ult_Compra	0.121871
Gasto_Total	0.077065
Ingresos	0.073850
N_CompCatalogo	0.070520
C_Carnes	0.064339
C_Vinos	0.064317
C_PremiumProds	0.055420
N_CompTiendas	0.048866
N_VisitasWebMes	0.042897
Año_Registro	0.039514
C_ProdsMar	0.037021
Edad	0.035631
C_Dulces	0.034915
Año_Nacimiento	0.032873
N_CompWeb	0.032366
C_Frutas	0.032242
Mes_Registro	0.025127
N_CompPromos	0.024508
Niv_Educación	0.018804
Semana_Registro	0.017885
Estado_Civil	0.016927
N_Adolescentes	0.015441
Trimestre_Registro	0.009621
N_Niños	0.007688
Reclamo	0.000291

Fuente: Elaboración propia / Phyton

8.3.4 Extra trees

Los resultados del modelo de extra trees resultan en la siguiente matriz de confusión y los siguientes valores.

Figura 61. Matriz de confusión Modelo de extra trees



Fuente: Elaboración propia / Python

Tabla 7. Resultados del Modelo de extra trees

	precision	recall	f1-score	support
0	1.00	0.96	0.98	548
1	0.96	1.00	0.98	580
accuracy			0.98	1128
macro avg	0.98	0.98	0.98	1128
weighted avg	0.98	0.98	0.98	1128

Fuente: *Elaboración propia / Phyton*

Tabla 8. Variables de más incidencia del Modelo de extra trees

Ult_Compra	0.092289
N_CompCatalogo	0.069126
Año_Registro	0.060676
Gasto_Total	0.059969
C_Carnes	0.052329
C_Vinos	0.050040
N_CompTiendas	0.048012
N_CompWeb	0.042402
N_VisitasWebMes	0.041493
Ingresos	0.040188
C_PremiumProds	0.039998
Niv_Educación	0.036891
N_CompPromos	0.036278
Estado_Civil	0.033640
N_Adolescentes	0.032826
Semana_Registro	0.032735
Edad	0.032355
C_Dulces	0.032351
C_Frutas	0.031334
Año_Nacimiento	0.030785
C_ProdsMar	0.030296
Mes_Registro	0.028957
Trimestre_Registro	0.025141
N_Niños	0.018890
Reclamo	0.000997

Fuente: *Elaboración propia / Phyton*

Tabla 9. Resumen de resultados por modelo de predicción

	Decision Tree	Random Forest	Extra Trees	Logistic Regression
Model	Decision Tree	Random Forest	Extra Trees	Logistic Regression
Scaling	Normal Data	Normal Data	Normal Data	Normal Data
Type	Gini	Gini	Gini	-
Accuracy	0.9104	0.9485	0.9787	0.7765

Fuente: *Elaboración propia / Phyton*

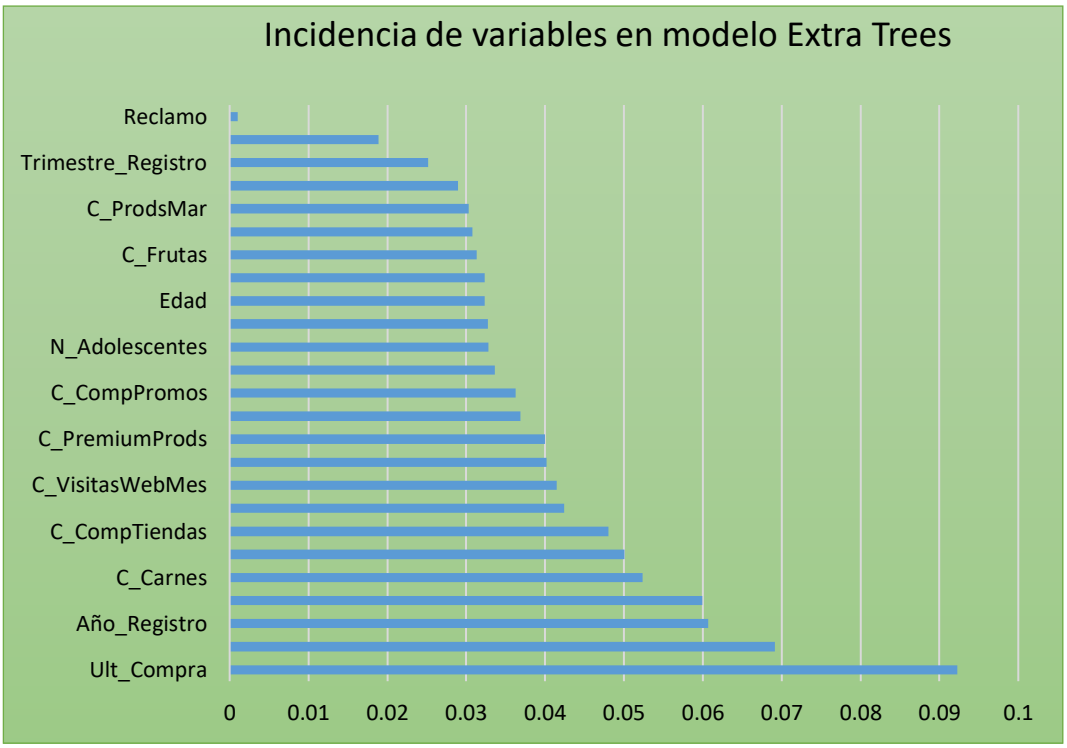
8.4 DISCUSIÓN DE LOS RESULTADOS

Tabla 10. Importancia de las variables según el modelo más exacto (Extra trees)

Variable	Importancia	% Importancia individual	% Importancia acumluada
Ult_Compra	0.092289	9%	9%
N_CompCatalogo	0.069126	7%	16%
Año_Registro	0.060676	6%	22%
Gasto_Total	0.059969	6%	28%
C_Carnes	0.052329	5%	33%
C_Vinos	0.05004	5%	38%
C_CompTiendas	0.048012	5%	43%
C_CompWeb	0.042402	4%	47%
C_VisitasWebMes	0.041493	4%	52%
Ingresos	0.040188	4%	56%
C_PremiumProds	0.039998	4%	60%
Niv_Educación	0.036891	4%	63%
C_CompPromos	0.036278	4%	67%
Estado_Civil	0.03364	3%	70%
N_Adolescentes	0.032826	3%	74%
Semana_Registro	0.032735	3%	77%
Edad	0.032355	3%	80%
C_Dulces	0.032351	3%	83%
C_Frutas	0.031334	3%	86%
Año_Nacimiento	0.030785	3%	90%
C_ProdsMar	0.030296	3%	93%
Mes_Registro	0.028957	3%	95%
Trimestre_Registro	0.025141	3%	98%
N_Niños	0.01889	2%	100%
Reclamo	0.000997	0%	100%

Fuente: *Elaboración propia*

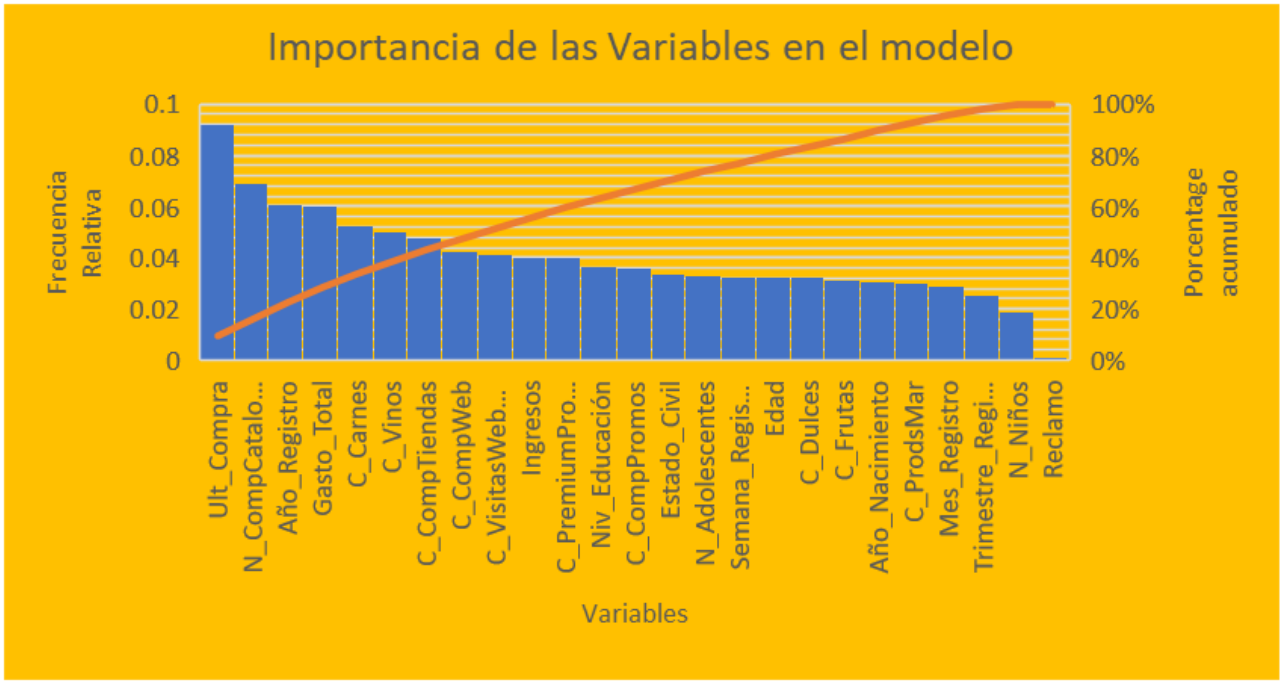
Figura 62. Incidencia de variables en modelo Extra Trees



Fuente: Elaboración propia

GRÁFICO DE PARETO: IMPORTANCIA DE VARIABLES EN EL MODELO

Figura 63. Importancia de las variables en el modelo (Extra Trees)



Fuente: *Elaboración propia*

El gráfico de Pareto anterior permite conocer cuáles son las variables que más incidencia causan en el modelo analizado. El gráfico de Pareto tiene como fundamento el concepto del 80-20. Esto es, básicamente que el 20% de las causas ocasionan el 80% de los problemas. Para efectos de nuestro estudio, esto se entendería como que el 20% de los factores o variables independientes inciden en casi el 80% del modelo.

Pero en el presente estudio se busca adaptar esto tomando en cuenta el número de variables y el contexto de comercial de la base de datos para extender el análisis más allá de estos porcentajes tradicionales. Para lo cual se toma en enfoque del 50-50. Es decir que el 50% de las variables individuales influyen directamente o son las más importantes para el modelo de predicción.

Bajo este enfoque las variables que más incidencia tienen en el modelo y requerirían análisis posteriores son:

Tabla 11. Variables con más incidencia en el modelo de predicción (Extra Trees)

Variable	Importancia	% Importancia individual	% Importancia acumluada
Ult_Compra	0.092289	9%	9%
N_CompCatalogo	0.069126	7%	16%
Año_Registro	0.060676	6%	22%
Gasto_Total	0.059969	6%	28%
C_Carnes	0.052329	5%	33%
C_Vinos	0.05004	5%	38%
C_CompTiendas	0.048012	5%	43%
C_CompWeb	0.042402	4%	47%
C_VisitasWebMes	0.041493	4%	52%
Ingresos	0.040188	4%	56%

Fuente: Elaboración propia

8.5 VARIABLES CON MÁS INCIDENCIA

Ult_Compra: Definida como el periodo de tiempo transcurrido (días) de la última compra del cliente. Esto indica que si un cliente realiza compras más frecuentemente este es más proclive a dar una respuesta positiva en una determinada campaña de promoción. A esta variable se le atribuye un nivel de importancia global del 9%, siendo el más alto de entre todas las variables.

N_CompCatálogo: Definida como el número de compras realizadas por medio de la utilización de catálogos. Esto indica que un porcentaje alto 7%, tomando en cuenta el aporte de las otras variables del modelo, explica que un cliente que realiza varias compras a partir de catálogos tiene más probabilidad de poder indicar una respuesta positiva en la campaña.

Año_Registro: Definida como el año de registro del cliente en la base de datos de sus consumidores. Esto indica que, al parecer para el presente estudio, el año particular en el cuál una persona se ha inscrito en la lista de consumidores de la empresa influye en su respuesta a la campaña. Con un valor del 6%, se puede inferir que hubo un periodo de tiempo en donde clientes se registraron y estos son más susceptibles a responder favorablemente en la campaña.

Gasto_Total: Definida como el valor del gasto total que una persona ha incurrido en la empresa a lo largo del tiempo. Indica que el valor del gasto histórico tiene influencia en la variable de respuesta con un valor del 6%.

C_Carnes: Definida como el valor total del consumo de cliente en productos cárnicos. Indica que para el modelo esta variable aporta con un 5% en términos de incidencia. Cuanto más una persona presenta consumos en esta categoría de producto, su respuesta tendería a ser también positiva en la campaña.

C_Vinos: Definida como el valor de gasto histórico del cliente en productos detallado como vinos. Indica que, si una persona tiene un historial más marcado del consumo de vinos, esta persona es más proclive a tener una respuesta positiva en la variable de respuesta.

C_CompTiendas: Definida como el valor del consumo que una persona registra en compras físicas en la tienda del comercio. Indica que, si una persona tiene un historial más alto de consumo en las tiendas del comercio, este también podría tener una respuesta positiva en la respuesta frente a la campaña.

C_CompWeb: Definida como el valor del consumo en compras realizadas a través de páginas web del comercio. Indica que, si una persona tiene un historial marcado en compras web, pues esta persona también tiene posibilidades de dar una respuesta positiva a la campaña de marketing. Esta variable influye con el 4% de la incidencia en el modelo.

C_VisitasWebMes: Definida como la cantidad de visitas por parte del consumidor a la página web del comercio. Indica que la cantidad de visitas Web que un consumidor hace al negocio, pues esto también se traduce en una mayor probabilidad de que una respuesta positiva a la campaña se realice.

Ingresos: Definida como el ingreso que un cliente tiene en el periodo de tiempo de un año calendario. Indica que el nivel de ingresos de una persona también tiene incidencia en el modelo y posible aceptación a la campaña de marketing ofertada.

9. PROPUESTA DE SOLUCIÓN

9.1 IMPLICACIONES ORGANIZACIONALES

La analítica de datos conjuntamente con las herramientas proporcionadas por las técnicas de Machine Learning en el presente estudio permiten tener distintas perspectivas de la problemática organizacional del caso. Tomando en cuenta que el problema organizacional principal es el de la poca efectividad que ha tenido la campaña de marketing para una población de clientes de una empresa y con esto todos los recursos que han sido invertidos y no retribuidos a la compañía. Con el fin de poder hacer que este tipo de campañas se realicen con mejor dirección y mayor impacto en la gente se toman en cuenta las siguientes propuestas de solución.

- ✓ **Entender los factores que afectan a la variable de respuesta:**

Tomando en cuenta el análisis exploratorio de datos, así como también la identificación de las características más influyentes en el modelo se observa cuáles de estas son las más importantes. En el estudio por ejemplo se puede mencionar que factores como la frecuencia de compra, el acceso a medios digitales de compra, el gasto total, nivel de ingresos y consumo de ciertos tipos de productos (vinos) tienen un mayor aporte al modelo, lo que quiere decir que influyen en la decisión final de un consumidor frente a un tipo de campaña. Es necesario poder analizar cada uno de estos para poder entender que hace que estos factores incidan más en el público. Al poder identificar estos factores de mayor incidencia, también logramos tener la base para que a nivel empresarial se puede tener un mayor enfoque de mejora en estas características específicas del modelo de negocio.

- ✓ **Predecir el éxito o la negativa de la campaña para un individuo:**

Los modelos de predicción del estudio han permitido conocer en un alto porcentaje si un cliente cae en el caso de tener posibilidad de aceptar una campaña de marketing ofrecida. Esto es de mucha ayuda para la empresa, pues se logra poder identificar mejor a la población objetivo que reúne estas características de dar una respuesta positiva y de diferenciar al grupo de personas que es proclive a no aceptar estas campañas. Al poder diferenciar estos grupos, se da la posibilidad a la administración que optimice mejor sus recursos y se ponga más atención que posiblemente acepte una campaña. Por otro lado, el enfoque también puede ser el de tratar de tomar en cuenta a la población que no acepta la campaña e investigar que es lo que se puede hacer para que las necesidades de esta población estén mejor cubiertas.

- ✓ **Clasificar a los clientes por características aportantes al modelo:**

Según las características de la base de datos y tomando en cuenta los valores de importancia y aporte de las variables hacia el modelo la empresa puede proceder a agrupar estas características y también agrupar a los clientes. Este agrupamiento vendría a realizarse en función de características demográficas similares o de consumo de los clientes con el objetivo de poder llegar con incentivos a los mismos basados en su probabilidad de aceptar la campaña. Poder segmentar a los clientes puede ser un aporte para las intenciones del negocio puesto el estudio de predicción de respuesta frente a campañas no solo se extendería a nivel de individuos específicos, lo cual tomaría más tiempo y recursos, sino más bien tomarlos en cuenta como grupos a lo que se puede llegar de manera más ordenada y eficiente.

- ✓ **Mejorar la toma de decisiones basadas en datos:**

Todo el proceso de analítica en el presente proyecto se puede tomar en cuenta para diferentes proyectos. Desde ejemplos como el del presente estudio referentes a campañas de marketing hasta considerar proyectos enfocados a la fidelización de clientes y el incremento de ventas. Todos estos procesos al estar basados en datos brindan la oportunidad analizar el campo de acción y reducir la probabilidad de fracaso en su implementación. Toda la información basad en datos se asocia con la finalidad de poder

tener procesos más eficiencia y apegados a la realidad de la empresa. En las organizaciones, la posibilidad de poder tomar decisiones basadas en datos hace que se pueden estructurar planes de trabajo en donde todos los actores puedan tener los objetivos claros de cumplimiento y dirección.

9.2 ESTRATEGIA ORGANIZACIONAL

IMPLICACIONES SOBRE INNOVACIÓN EMPRESARIAL

En el caso de la empresa de consumo masivo referente al presente estudio, las siguientes estrategias puede ser usadas con la finalidad de mejorar la analítica de datos de la empresa, así como también tener una mejorar en el proceso de optimización de las campañas de marketing y de fidelización para sus clientes.

1. Establecer objetivos claros:

La empresa tomando en cuenta el alcance del estudio debe proponerse realizar una ruta de trabajo para tener y establecer los objetivos para el proyecto de ciencia de datos que sean específicos y medibles. Se necesita saber también cuáles son las métricas que se van a usar en la implementación del proyecto y cuáles de estas se deben mejorar. Como instancia final en esta etapa también se debe tomar en cuenta cuáles son los resultados que se espera en términos de retorno de inversión, así como también qué se busca en términos de fidelización de clientes.

2. Grupos de clientes:

Utilizar análisis de datos para hacer segmentaciones de los clientes y dividir a los mismos en grupos más específicos con características específicas y similares. Esto permitirá la creación de campañas dirigidas y adaptadas a cada grupo buscando tomar en cuenta las características definidas de cada grupo y cuáles serían las estrategias para abarcar cada uno de estos de la mejor manera.

3. Recopilación y gestión de información:

La recopilación y gestión de información es un factor fundamental para mejorar la eficiencia organizacional de una empresa que se plantea adentrarse en un proyecto de ciencia de datos. Implementar sistemas para recopilar y administrar de manera efectiva los datos de los clientes es una buena manera para poder mejorar el flujo de los datos de una empresa. Esto podría incluir datos sobre compras anteriores, comportamiento en línea y preferencias, entre otras cosas u cualquier otra información que pudiese aportar al estudio.

4. Análisis para la predicción:

Para predecir el comportamiento futuro de los clientes, se puede utilizar técnicas de análisis predictivo. Esto ayudará a modificar las estrategias de fidelización y marketing de acuerdo. Tomando en cuenta que se puede ajustar los aspectos negativos en ciertas campañas y tomar las mejores herramientas que de campañas exitosas.

5.Configuración de campañas:

La configuración y personalización de las campañas pueden ser el resultado de utilizar los datos recopilados y haber realizado la identificación de las variables más importantes para crear campañas extremadamente personalizadas. Esto incluye el contenido del mensaje, el momento de entrega y la forma preferida de comunicación de cada cliente.

6. El aprendizaje autónomo:

Implemente modelos de aprendizaje automático para encontrar patrones y tendencias en la información. Esto puede ayudar a comprender mejor el comportamiento del cliente y predecir sus reacciones a una variedad de estrategias de marketing.

7. Optimización de los canales en línea:

Analizar qué canales digitales, como redes sociales, correo electrónico, aplicaciones móviles, etc., funcionan mejor para llegar a segmentos de clientes específicos. Asignar recursos para estos canales de manera estratégica.

8. Vigilancia y medición constante:

Para evaluar el desempeño de las campañas, establezca un sistema de seguimiento continuo. Los resultados y las opiniones de los clientes determinarán la estrategia.

9. Trabajo colaborativo entre equipos:

Fomentar la colaboración entre los equipos de tecnología, análisis de datos y marketing. Ayudará a garantizar que las estrategias estén alineadas con los objetivos mediante una comunicación fluida.

10. Educación y capacitación:

ayudar a los empleados a entender los datos y interpretar los resultados. Esto ayudará a los equipos a tomar decisiones basadas en la analítica.

11. Respecto a la Privacidad:

asegurarse de que las leyes de privacidad y protección de datos se cumplan al recopilar y utilizar los datos de los clientes.

12. Experimentos y pruebas A/B:

Realizar experimentos y pruebas A/B para evaluar varios enfoques de marketing y determinar cuáles generan los mejores resultados en términos de fidelización y participación.

13. Mantenimiento de la tecnología:

Mantener actualizadas las herramientas y plataformas tecnológicas para garantizar la eficacia y precisión en la recopilación y análisis de datos.

14. Los comentarios de los clientes:

Recopile con frecuencia las opiniones de los clientes sobre las campañas y su experiencia general. Use esta información para mejorar continuamente las estrategias.

15. La adaptación continua:

El entorno digital y las preferencias de los clientes cambian constantemente. La táctica debe ser adaptable y estar lista para cambiar según sea necesario.

Estas pueden ser algunas de las estrategias básicas que se pueden tomar en cuenta para que las empresas de consumo, como en el presente caso puedan aumentar la eficiencia de los proyectos de analítica a los cuales deseen adentrarse. Tanto para casos de fidelización o aumentar algún índice de aceptación de sus clientes, estas herramientas pueden ayudar a la creación de un plan de trabajo ordenado y que tenga muchas características y métricas de desempeño bien marcadas.

10. CONCLUSIONES

La implementación de proyectos de analítica, como modelos de predicción, para mejorar la eficiencia y el éxito de las campañas de promoción y marketing en una empresa de consumo masivo puede ofrecer una serie de beneficios significativos. Al aprovechar los datos disponibles y aplicar técnicas de ciencia de datos, la empresa puede obtener información valiosa sobre el comportamiento y las preferencias de sus clientes. En base a esto se puede mencionar las siguientes conclusiones:

Personalización Mejorada: Los modelos de predicción permiten una personalización más profunda en las estrategias de marketing. Al comprender las preferencias individuales de los clientes, las campañas pueden adaptarse para satisfacer sus necesidades específicas.

Mayor Eficiencia: La segmentación precisa y las predicciones ayudan a dirigir los recursos y esfuerzos de marketing de manera más efectiva hacia los clientes más propensos a responder positivamente. Esto ahorra tiempo y dinero al evitar la dispersión de esfuerzos en audiencias menos relevantes.

Mejor Retorno de Inversión (ROI): Al concentrarse en los clientes que tienen más probabilidades de participar en las campañas, la empresa puede lograr un ROI más alto al generar ventas y lealtad entre aquellos que tienen un mayor potencial de conversión.

Adaptación Continua: Los modelos de predicción no son estáticos. Con el tiempo, la empresa puede refinar y mejorar sus modelos a medida que recopila más datos y retroalimentación de las campañas anteriores.

Competitividad: Las empresas que utilizan la analítica de datos de manera efectiva tienen una ventaja competitiva al comprender mejor su base de clientes y responder rápidamente a las tendencias del mercado.

11. RECOMENDACIONES

Inversión en Capacidades Analíticas: Invertir en tecnología y talento especializado en análisis de datos y ciencia de datos es esencial para el éxito de los proyectos de predicción y analítica.

Calidad de Datos: Asegurarse de que los datos recopilados sean precisos y de alta calidad. Los modelos se basan en datos confiables, por lo que es crucial tener sistemas sólidos de recopilación y gestión de datos.

Validación Constante: Validar y ajustar los modelos de predicción de manera continua. La realidad puede diferir de las predicciones iniciales, por lo que es importante hacer ajustes según sea necesario.

Pruebas Rigurosas: Antes de implementar en grande, realizar pruebas piloto de las estrategias de marketing basadas en los modelos de predicción. Evaluar los resultados y hacer mejoras antes de lanzar a gran escala.

Colaboración entre Equipos: Fomentar la colaboración entre los equipos de marketing, análisis de datos y tecnología para garantizar una implementación fluida y efectiva de los proyectos de analítica.

Cumplimiento Normativo: Asegurarse de que la recopilación y el uso de datos cumplan con las regulaciones de privacidad y protección de datos.

Educación Interna: Capacitar a los equipos de marketing en la comprensión de cómo funcionan los modelos de predicción y cómo utilizar los resultados para diseñar campañas efectivas.

Flexibilidad y Adaptación: Estar dispuesto a adaptar las estrategias según los resultados y el feedback de los clientes. La analítica es un proceso iterativo.

Medición y Evaluación: Establecer métricas claras para medir el éxito de las campañas basadas en los modelos de predicción. Evaluar el impacto en el ROI, la participación y la fidelización de los clientes.

Aprendizaje Continuo: La analítica de datos es un campo en constante evolución. Mantenerse al tanto de las nuevas tendencias y técnicas para seguir mejorando las estrategias en el tiempo.

Al seguir estas recomendaciones, la empresa estará bien posicionada para aprovechar el potencial de los proyectos de analítica y modelos de predicción para aumentar la eficiencia y el éxito de sus campañas de promoción y marketing, logrando una mayor fidelización y satisfacción de los clientes.

REFERENCIAS

Amat, J. (2020). Arboles de decision Python. Ciencia de datos, teoría y ejemplos prácticos en R y Python. https://cienciadedatos.net/documentos/py07_arboles_decision_python

Calva, Karen. (2021). Modelo de predicción del rendimiento académico para el curso de nivelación de la Escuela Politécnica Nacional a partir de un modelo de aprendizaje supervisado. https://lajc.epn.edu.ec/index.php/LAJC/article/download/264/159/

Cárdenas, J. (2022). Qué es la regresión logística binaria Y Como analizarla. Networkianos. Blog de Sociología. https://networkianos.com/regresion-logistica-binaria/

Carrasco Ortega, M. (2017). Herramientas del marketing digital que permiten desarrollar presencia online, analizar la web, conocer a la audiencia y mejorar los resultados de búsqueda. Scielo(45). Obtenido de http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S1994-37332020000100003

Delgado, M. (2017). Gestión de la relación con los clientes y segmentación. Sistema de Información Científica Redalyc. Red de Revistas Científicas. https://www.redalyc.org/pdf/4259/425942412008.pdf

Espino, C. (2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso. https://openaccess.uoc.edu/bitstream/10609/59565/6/caresptimTFG0117mem%C3%

Grewal, D., & Kopalle, P. (2019). The future of technology and marketing: A multidisciplinary perspective. SpringerLink. https://doi.org/10.1007/s11747-019-00711-4

Huertas, A. (2020). Algoritmos de aprendizaje automático supervisado utilizando datos de monitoreo de condiciones: Un estudio para el pronóstico de fallas en máquinas.

https://repository.usta.edu.co/bitstream/handle/11634/29886/2020alexanderhuertas.pdf?sequence=1&isAllowed=y

Giraldo, L. (2018). Los desafíos del marketing en la era del big data. Sistema de Información Científica Redalyc, Red de Revistas Científicas. https://www.redalyc.org/journal/4768/476852090003/

Gonzalez, L. (2018). Aprendizaje Supervisado: Random forest classification. https://aprendeia.com/aprendizaje-supervisado-random-forest-classification/

Marín, J. (2019). Análisis de datos para el marketing digital emprendedor: Caso de estudio del Parque de Innovación Empresarial de Manizales. http://www.scielo.org.co/pdf/unem/v22n38/2145-4558-unem-22-38-65.pdf

Namdhini, S. (2021). Behavioral aspects: Search engine optimization strategy focusing for new entrants. Turkish Journal of Computer and Mathematics Education. https://doi.org/10.17762/turcomat.v12i11.6087

Roland, T. (2020). The future of marketing. International Journal of Research in Marketing. https://doi.org/10.1016/j.ijresmar.2019.08.002

Rubio, A. (2019). Estrategia de marketing digital para fidelizar a nuevos clientes a traves de redes sociales Y estrategias de SEO Y SEM: DJ Klaus Hidalgo. Pirhua - Universidad de Piura. https://pirhua.udep.edu.pe/handle/11042/3957

Salazar, A. (2019). MPORTANCIA DE UNA INVESTIGACIÓN DE MERCADO. https://www.itson.mx/publicaciones/pacioli/documents/no71/49a.-_importancia_de_la_investigacion_de_mercado_nx.pdf

Saura, J. (2021). Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics. Journal of Innovation & Knowledge. https://www.sciencedirect.com/science/article/pii/S2444569X20300329

Shobhana C. (2022). Personalization in personalized marketing: Trends and ways forward. Psychology & MarketingVolume 39, Issue 8. https://doi.org/10.1002/mar.21670

Zamorano, Juan. (2018). Comparativa y análisis de algoritmos de aprendizaje automático para la predicción del tipo predominante de cubierta arbórea. https://docta.ucm.es/entities/publication/7f2287a4-7122-454d-803e-0a8b47786649

Zúñiga, Freddy & Poveda, Diego & Llerena, William. (2023). El Big data y su implicación en el marketing. Revista de Comunicación de la SEECI. 56. 302-321. 10.15198/seeci.2023.56.e83