Modelos de lealtad de clientes de una entidad financiera con el uso de datos heterogéneos

Loyalty models of customers of a financial institution with the use of heterogeneous data

María Murillo Olóriz

Trabajo fin de Máster

Universidad de Zaragoza

Modelos de lealtad de clientes de una entidad financiera con el uso de datos heterogéneos

María Murillo

14 de septiembre de 2018

Resumen

La cantidad de información que una empresa tiene sobre sus clientes es cada vez mayor. Además de la información que se genera de forma directa dada la relación comercial entre empresa y cliente, se consiguen datos de forma indirecta a partir de otros orígenes como las redes sociales o en el ámbito de las entidades financieras a través de aplicaciones para móviles y las páginas web. Por otro lado, el estudio de la lealtad del cliente a través de las experiencias de estos, ha ido cobrando cada vez más importancia y las entidades financieras han desarrollado un potencial interés en buscar patrones de comportamiento de estos consumidores con el objetivo de realizar políticas de captación y fidelización.

La importancia que se le asigna a la satisfacción del cliente en el ámbito empresarial actual es cada vez mayor y es por esto que una gran cantidad de empresas y entidades financieras están focalizando todos sus esfuerzos en conocer lo que piensa el cliente acerca de ellos y en conseguir una fidelización potente de estos.

El objetivo de este trabajo es crear una base de datos con orígenes de datos heterogéneos de una entidad financiera y a partir de ella ser capaz de diseñar un modelo de lealtad que permita estimar la probabilidad que tiene un cliente de ser leal dentro de la entidad financiera, para finalmente, clasificar a los clientes y establecer distintas políticas y campañas dirigidas a uno u otro segmento de población.

Esta memoria se ha organizado en cinco capítulos y un anexo. La lealtad, considerada como un concepto abstracto y difícil de modelizar. En el primer capítulo, tras describir el problema que la entidad financiera formuló inicialmente, se incluye una revisión de la literatura llevada a cabo acerca del concepto lealtad del cliente y los modelos de lealtad. Esta revisión se ha tenido en cuenta para establecer qué variables pueden ser interesantes considerar en los modelos de lealtad.

En el segundo capítulo se describe cómo se ha construido la base de datos propia que ha permitido estimar los modelos, a partir de la base de datos general de la entidad. Una labor en la que entra en juego el análisis de la base general y el estudio de las diferentes variables disponibles, así como la toma de decisiones a la hora de elegir que variables se van a tener en cuenta. En particular, se decide estimar la lealtad en dos ámbitos de interés: ámbito de negocio y ámbito de comunicaciones. La primera parte de la aplicación, una vez construida la base de datos propia, consiste en utilizar dos algoritmos de predicción y clasificación para estudiar este concepto de lealtad en el contexto de la entidad financiera bajo estudio: la regresión logística y los árboles de decisión. Con ambas metodologías se han estimado modelos de lealtad en cada uno de los dos ámbitos, y a continuación, se integran los ámbitos o submodelos de lealtad de negocio y comunicaciones para formar un modelo de lealtad global que permita estimarla de forma general. Se trata de un modelo para estimar la lealtad ampliable a tantos ámbitos como se requiera en un futuro. En el capítulo 3 se incluyen los

modelos estimados mediante el uso de regresión logística y en el capítulo 4, los resultados dados con los árboles de decisión.

Finalmente, en el capítulo 5 se desarrolla la segunda parte de la aplicación que consiste en la segmentación de los clientes obtenida a partir de los modelos estimados y las conclusiones extraídas de este análisis, así como las conclusiones finales del trabajo. En esta parte de segmentación o agrupación de los clientes, se trata de estudiar distintas variables demográficas que describan a los usuarios leales en cada ámbito, buscando similitudes y diferencias entre estos.

Abstract

The amount of information a company has about its customers is increasing. In addiction to the information that is generated directly by the customer-company relationship, data is indirectly obtained from other sources such as social networks or in the field of financial entities through apps and web pages. On the other hand, the field of study of customer loyalty conected with the customer experience has been gaining increasing importance and financial insititucions have developed a potential interest in seeking patterns of behavior of these customers with the aim of carrying out recruitment policies and loyalty.

The importance assigned to customer satisfaction in the current business environment is increasing and that is why a large number of companies and financial institutions are focusing all their efforts on knowing what the client thinks about them and in get a strong loyalty of these.

Having all this information available, the objective of this work is to create a common database with heterogeneous data sources from a financial institution and from it to desing a loyalty model that allows to estimate the probability that a client has to be loyal within the financial institution, to finally classify clients and establish different policies and campaigns aimed at one or another segment of the population.

This report has been organized into six chapters and an annex. Loyalty, considered as an abstract concept and difficult to model. In the first chapter, after describing the problem that the financial institution initially formulated, a review of the literature carried out on the concept of customer loyalty and loyalty models is included. This review has been taken into account to establish which variables can be interesting to consider in loyalty models.

In the second chapter, we describe how the own database has been constructed, which has made it possible to estimate the models, based on the general database of the entity. A work in which comes into play the analysis of the general base and the study of the different variables available, as well as the decision making when choosing which variables are going to be taken into account. In particular, it is decided to estimate loyalty in two areas of interest: business scope and communications field. The first part of the application, once the own database is built, consists of using two prediction and classification algorithms to study this concept of loyalty in the context of the financial institution under study: logistic regression and decision trees. With both methodologies, models of loyalty have been estimated in each of the two areas, and then the spheres or submodels of business loyalty and communications are integrated to form a model of global loyalty that allows estimating it in a general way. It is a model for estimating loyalty that can be extended to as many areas as required in the future. Chapter 3 includes the models estimated using logistic regression and in Chapter 4, the results given with the decision trees.

At the end, in chapter 5 the second part of the application is developed, consisting of the segmentation of the clients obtained from the estimated models and the conclusions drawn from this analysis, as well as the final conclusions . In this part of segmentation or clustering of clients, it consists in studying different demographic variables that describe the loyal users in each area, looking for similarities and differences between them.

Índice

Índice de figuras

Índice de tablas

1. Introducción

El trabajo que se presenta en esta memoria ha permitido aplicar parte de los conocimientos adquiridos durante la realización del Máster de Modelización, Estadística y Computación a la resolución de un caso real de gran interés y constituye el Trabajo Fin de Máster. El trabajo surge como resultado de la colaboración con una empresa de Zaragoza que lleva a cabo un proyecto con una entidad financiera cuyo objetivo es el de proporcionarle un modelo que mida la lealtad del cliente en términos de la probabilidad de ser leal. Este estudio se ha realizado a partir de los datos que ha proporcionado la entidad financiera, los cuales han sido anonimizados para su posterior tratamiento. Se han trabajado y estudiado profundamente para poder modelizarlos de forma adecuada y poder estimar de esta manera los modelos deseados. En este capítulo se va a describir el problema formulado por la entidad financiera y se va a realizar una revisión de la literatura sobre el concepto de lealtad y algunos modelos de lealtad.

1.1. Descripción del problema

El contexto de partida del trabajo es el de la solicitud por parte de la entidad financiera a una empresa de la estimación de un modelo de lealtad sobre sus clientes, aprovechando la información disponible en su base de datos.

La empresa que lleva a cabo el proyecto, es una empresa que ofrece servicios de consultoría, especializada en tecnologías de la información y comunicaciones; digital business; ingeniería y tranformación digital. El cliente, que es la entidad financiera, pone a disposición de la empresa sus equipos informáticos y determinados datos comerciales de los que dispone, que permiten llevar a cabo el proyecto.

Hasta ahora, los modelos de los que dispone la entidad se basan más en producto, es decir, en la probabilidad que tiene un cliente de contratar un producto. El modelo de lealtad que se plantea, está más ligado al cliente. Se trata de identificar clientes que tienen una relación emocional con la entidad, identificar que ha motivado esta relación y analizar cómo el banco podría gestionar el resto de clientes para conseguir esta fidelidad.

El esquema que se pretende seguir es el siguiente:

1. Crear un modelo de los datos proporcionados que facilite su posterior análisis.
2. Perfilar, categorizar y segmentar a los clientes.
3. Buscar patrones y evidencias que proporcionen valor para el cálculo del modelo.
4. Proponer y definir el modelo.

Tras el planteamiento del problema por parte de la empresa y tratándose de un concepto tan abstracto como es la lealtad, se realizó una exhaustiva revisión de la literatura sobre modelos y técnicas que estudian la lealtad y que tratan de segmentar o agrupar a los clientes.

1.2. Concepto de lealtad

Uno de los objetivos principales de cualquier empresa o entidad financiera radica en la fidelización y retención de sus clientes. A partir de este objetivo surgen distintas técnicas y teorías que van cambiando y actualizándose con el paso de los años sobre la lealtad del cliente.

El concepto de lealtad ha sido uno de los temas más investigados y ha ido evolucionando a lo largo de los años, siendo abordado desde dos corrientes distintas: como un comportamiento

efectivo y como una actitud. Primeramente se atribuye la lealtad a un comportamiento de repetición de compra, es decir a un comportamiento efectivo. Se trata de una reacción sistemática que viene dada por el mecanismo estímulo-respuesta. Este enfoque se denomina enfoque conductual y hace referencia a que la probabilidad de volver a comprar un producto viene dada por la frecuencia de compra y los resultados satisfactorios que proporciona. Es decir, cuanto mayor sea esta probabilidad más leal es el consumidor bajo este enfoque (Delgado, 2004; Park y Bai,2014). Esta perspectiva ha recibido numerosas críticas puesto que puede considerarse ineficiente predecir la lealtad del cliente únicamente por la frecuencia de compra del producto, además de que deja fuera de medición la ”disposición” y ”emoción” del cliente que hacen que este tome la decisión de compra, las cuales según algunos autores, son las únicas que pueden permitir catalogar al cliente como leal.

De acuerdo con la segunda corriente, la lealtad es concebida como una actitud que relaciona las emociones y los sentimientos entre el cliente y la marca. Este enfoque pretende poner mayor hincapié en la comprensión de las estructuras emocionales del cliente, ofreciendo así una explicación teórica y deductiva del fenómeno lealtad (Delgado, 2004). La perspectiva actitudinal se basa en el comportamiento del cliente y debe expresar todos los elementos que la integran (cognitivos, afectivos y conativos), puesto que entiende la lealtad como un concepto multivariable (Dick y Basu, 1994; Oliver, 1999; *The Journal of Marketing*). Un consumidor leal es aquel que manifiesta sentimientos positivos por su marca, de forma que la actitud de este es un antecedente que condiciona la lealtad. Por tanto, este enfoque defiende que la lealtad no puede ser solamente resultado de una alta frecuencia de compra, ya que esta frecuencia puede ser el resultado de una inercia en el comportamiento y que no representan compromiso.

Posteriormente surgió una corriente que ya no creía en la medición exclusiva de la lealtad desde un ámbito o desde otro, sino que materializaron el estudio de la lealtad a través de la repetición de compra unido al compromiso que surge entre el cliente y la marca. Esta corriente defiende que la frecuencia de compra de un producto por sí sola no es capaz de medir la lealtad, es necesario ampliar la definición puesto que puede venir sesgada por determinadas campañas vigentes en el momento de la compra. Entiende la lealtad como un proceso dinámico y bidimensional que debe abarcar el constructo actudinal y el de comportamiento. Jacoby y Chestnut (1978) coinciden con Day en abarcar un concepto de lealtad más amplio, enfocando la definición a percibir procesos cognitivos asociados al comportamiento de elección, puesto que la evaluación de la lealtad por la frecuencia de compra puede quedar sesgada por diversos factores, cómo por ejemplo, la presencia de determinadas ofertas en momentos puntuales como descuentos o subidas o bajadas de precios.

Resumiento, la lealtad del cliente se define como el esfuerzo que realiza una empresa para mantener la conexión con el cliente; siendo de vital importancia su trabajo diario, debido a la fuerte competencia a la que se enfrenta.

Una estrategia de marketing exitosa orientada al cliente es de vital importancia para una entidad financiera puesto que puede ayudar a fortalecer las relaciones entre los clientes y la entidad. Comprender las características del cliente y satisfacer sus necesidades contribuye a mejorar la lealtad del cliente y además permite generar grandes beneficios al poder conocer (o al menos estimar) el resultado de una campaña comunicativa antes de su lanzamiento.

1.3. Modelos de lealtad

1.3.1. Modelo de Oliver

El modelo de Oliver (1999) es conocido mundialmente dentro de la literatura de lealtad del consumidor. Este autor, comienza sus estudios basándose en el concepto de satisfacción por un servicio y/o producto y en la fuerte relación que comparte con la lealtad. Según Oliver exiten cuatro fases o niveles de lealtad las cuales se basan en el modelo clásico de actitud de Fishbein y Ajzen,

de 1975 y que se conoce como Teoría de la Acción Razonada. La lealtad es un proceso gradual que va evolucionando según el grado de proximidad de los clientes con la entidad. Estas cuatro fases son: cognitiva, afectiva, conativa y de acción, las cuales se explican en las siguientes líneas:

- Fase cognitiva: En esta primera fase de lealtad la información disponible acerca de los atributos de la marca determina cuál de ellas es preferible entre las distintas alternativas. Está ligada a los atributos del producto o servicio en los cuales se incluyen: características del producto, precio, beneficios asociados al uso de este producto, etc. Esta etapa, se basa en la creencia en la marca por experiencias anteriores o conocimiento previo de información. La principal dificultad existente en esta etapa de la lealtad es provocada cuando las marcas de la competencia presenten precios más competitivos o mejores características de producto. Entre las estrategias para superar dicha dificultad juegan un papel importante las campañas publicitarias en los medios de comunicación.

- Fase afectiva: En la segunda etapa, se generan lazos de unión o actitudes entre el consumidor y la marca, producto o servicio que se van desarrollando de acuerdo con la acumulación de situaciones satisfactorias, por lo que uno de los obstáculos que pueden surgir es que en la fase anterior el cliente haya experimentado insatisfacción. Además, el cliente reparte su lealtad, es decir, no la genera únicamente hacia un producto o marca determinada, sino que puede aumentar la preferencia por marcas de la competencia. Se trata de una lealtad por simpatía, es decir, que surge porque al cliente le gusta la marca o producto.

- Fase conativa: Esta fase está asociada a la motivación e intención de compra que proviene de la conexión emocional que ha surgido en etapas anteriores. El desarrollo de lealtad en esta etapa viene influenciado por situaciones satisfactorias repetidas anteriormente. Llegar a esta fase implica un compromiso con la marca, producto o servicio más ligado a la motivación. La principal dificultad a la que se puede enfrentar esta fase que surjan argumentos de persuasión provenientes de empresas de la compentencia.

- Fase de acción: En esta etapa la lealtad se materializa en fidelidad del cliente, en la cual, el cliente está en sintonía perfecta con la marca, producto o servicio y se encuentra profundamente involucrado. En esta etapa, la intención y motivación de etapas anteriores se traducen en diposición a actuar porque se produce un deseo adicional para sortear los obstáculos.

  Resumiendo, la fase cognitiva se centra en las características de la marca, producto o servicio; la fase afectiva hace referencia a empatizar con la marca, la fase conativa se produce en el momento en que el consumidor orienta sus intereses para volver a comprar por una experiencia satisfecha anterior, por último, la fase de acción se refiere al compromiso de volver a comprar y al hecho de manifestar orgullo por la marca que se materializa en divulgaciones positivas acerca de ella.

1.3.2. Modelo de Dick y Basu

Dick y Basu (1994) defienden que la lealtad viene determinada por el grado de relación existente entre la actitud relativa del consumidor y la repetición de compra y además toma en consideración elementos actitudinales y conductuales. Esto se debe a que la disposición del individuo para volver a adquirir nuevos productos se basa en gran medida en la actitud que este tiene hacia la compañía. Por tanto, es necesario observar el comportamiento y la disposición que tiene el consumidor para conceptualizar la lealtad y poder medirla. Dick y Basu distinguen cuatro tipos de lealtad, en función de si esa actitud positiva que tiene el cliente hacia la entidad se materializa finalmente en adquisición de productos o servicios que esta ofrece.

- Lealtad verdadera: Hace referencia a la existencia de una correlación fuerte entre actitud relativa del consumidor y la repetición de compra o comportamiento efectivo. Viene dado por la voluntariedad y motivación del consumidor para establecer lazos más fuertes con la marca a pesar de los problemas que puedan surgir. El consumidor siente un alto compromiso con la marca influenciado por la satisfacción que le produce.

- Lealtad latente: Está asociada con alta actitud relativa, pero bajo patrón de repetición de compra. Es decir, se da en aquellas situaciones en las que el cliente se muestra altamente comprometido pero este compromiso no se materializa en alta frecuencia de compra y a su vez la van alternando con opciones de la competencia.

- Lealtad falsa o espuria: Asociada a una baja actitud relativa y alto patrón de repetición de compra. Es una relación más de dependencia que voluntaria motivada principalmente porque el consumidor no tiene más alternativas o no cuenta con recursos suficientes para poder elegir. Conceptualmente podría ser similar a la inercia, en la que un consumidor percibe poca diferencia entre marcas. La influencia social también puede desencadenar en lealtad espuria.

- Ausencia de lealtad: En ella se entrelazan una baja actitud relativa y un bajo comportamiento de repetición de compra. Agrupa al segmento de clientes que resulta más desleal, el cual se caracteriza por falta de compromiso y de conexión afectiva así como de comportamiento hacia la marca. La baja actitud relativa en algunos mercados específicos puede deberse en parte a que la competencia ofrece condiciones muy similares. Este podría ser el caso de las entidades bancarias.

1.3.3. Modelo de Vilares y Coelho

Como explica Araújo en su tesis, Vilares y Coelho creen en la existencia de una estrecha relación entre satisfacción y lealtad, pero ocupando distintos ámbitos. La satisfacción habitualmente se dirige de forma específica al producto o servicio y resultad dinámica puesto que se adapta en mayor o menor medida al consumidor. Por el contrario, la lealtad es un concepto más amplio y estático que se dirige de forma más directa a la empresa en conjunto. De acuerdo con Vilares y Coelho, la lealtad se puede categorizar en 3 clases:

- Lealtad afectiva: relacionada directamente con conexiones emocionales del cliente con la empresa. Un ejemplo de esto son las opiniones y comentarios que estos clientes realizan.

- Lealtad racional o cognitiva: estrechamente relacionada con las evaluaciones sobre la relación comercial entre las que se incluyen los juicios de valor sobre los productos, los precios, el valor percibido o los costes.

- Lealtad conductual: la cual es un resultado de las dos anteriores, hace referencia a la intención de continuar la relación empresa-cliente y de recomendar la empresa a terceros. En conclusión, Vilares y Coehlo, insisten en que la lealtad puede ser afectiva, cognitiva o conductual, siempre influenciada por la satisfacción del cliente.

1.3.4. Modelo RFM

La denominación del modelo procede de sus siglas en inglés: Recency, Frequency y Monetary. El modelo RFM se basa en el comportamiento pasado que ha tenido un cliente. Fue propuesto por Hughes en 1994 y trata de diferenciar por grupos a los clientes cuando se tienen grandes bases de datos en función de tres variables, que reflejan este comportamiento.

- Recency (Recencia): Periodo de tiempo pasado desde la última compra. Cuanto más pequeños sean los valores de esta variable, mayor es la probabilidad del cliente de ser leal.
- Frecuency (Frecuencia de compra): Número de compras que se han realizado en un determinado periodo de tiempo.
- Monetary (Valor monetario): Cantidad de dinero total que se ha gastado en un periodo de tiempo determinado.

Los clientes que compraron recientemente, asiduamente y gastaron grandes cantidades en productos de la compañía tienen más probabilidad de ser leales. Este modelo se basa en el famoso principio de Pareto 80/20, ya que el 80 % de las compras se realizan por el 20 % de los clientes, que son los considerados más leales.

El modelo RFM tiene una larga trayectoria en campañas de marketing, aplicándose con la intención de encontrar segmentos o grupos de clientes que tienen alta probabilidad de responder a una determinada campaña para después ver que caraterísticas demográficas tienen. Esto se materializa, por ejemplo, en si abren o no los correos publicitarios que se les envían, acuden a las oficinas, están en contacto con la compañía, etc.

2. Construcción de la base de datos

2.1. Software utilizado

Tanto la extracción como el diseño de la base de datos se llevó a cabo a través del programa SQL Server 2016 13.0., que debe su nombre a sus siglas en inglés: Structured Query Language. Se trata de un sistema de manejo de bases de datos relacionales desarrollado por Microsoft, que utiliza un lenguaje de desarrollo por líneas de comandos denominado Transact-SQL(T-SQL).

T-SQL permite realizar operaciones en SQL Server entre las que destacan la creación y modificación de esquemas de bases de datos, inserción y modificación de datos en la base de datos, así como la administración de los servidores. T-SQL requiere del manejo del álgebra y el cálculo relacional para efectuar consultas para obtener información de la base de datos. En resumen,T- SQL tiene las siguientes funciones:

- Lenguaje de definición de datos (LDD): proporciona comandos para la definición de esquemas de relación, borrado de relaciones y modificaciones de los esquemas de relación.
- Lenguaje de manipulación de datos (LMD): iincluye comandos para realizar consultas basadas en álgebra relacional.
- Integridad: A través del LDD se permite especificar restricciones que deben cumplir los datos almacenados en las bases de datos.
- Control de transacciones: permite especificar el comienzo y el fin de una transacción.
- Dinamismo: permite incorporar instrucciones de SQL en otros lenguajes de programación como C++, Java, Cobol, Pascal y Fortran.
- Autorización: La parte de LDD permite especificar los derechos de accceso a las relaciones establecidas.

2.2. Datos disponibles

Tratándose de que el caso de estudio es una entidad financiera que almacena una gran cantidad de datos sobre sus clientes, la recolección y extracción de los datos conllevó cierta dificultad. Para economías domésticas y pequeñas empresas las bases de datos pequeñas son adecuadas, pero para grandes organizaciones como es el caso de estudio es necesaria la implantación de sistemas de bases de datos construídas sobre servidores.

El sistema de almacenamiento de datos SQL server nombrado anteriormente, organiza la información en base a un conjunto de servidores que a su vez contienen una serie de bases de datos. Un servidor es una unidad de hardware que gestiona múltiples funciones y que conecta red y clientes.

En el caso de la entidad financiera su sistema de base de datos se estructura en torno a 16 servidores, los cuales se clasifican en diferentes entornos. Existen 4 servidores que se encuentran en el entorno de desarrollo, 4 que se sitúan en el de pruebas, 4 que están en el entorno de preproducción y por último 4 que se encuentran en producción. De esta manera, los entornos de desarrollo y prueba pertenecen a la etapa 1 y los de preproducción y producción a la etapa 2. En los entornos de etapa 1 se encuentran los servidores cuyas bases de datos almacenan tablas que no están terminadas y procesos todavía en creación. Los entornos de la segunda etapa contienen las bases de datos cuyas tablas están listas para el uso y que se cargan y actualizan bien sea diariamente, semanalmente o mensualmente. Son del entorno de producción del cual se extraerán finalmente los datos.

El entorno de producción, cuenta con 80 bases de datos, entre las que destacan el uso de la base de datos comercial y la base de datos CRM para la composición de la base de datos de uso en el caso real. Esta segunda base, CRM debe su nombre a sus siglas en inglés: Customer Relationship Management y permite a la entidad gestionar de forma ordenada las relaciones con sus clientes. La existencia de una base de datos CRM va ligada a la transformación digital iniciada por las compañias en las que se desea conseguir una única base de datos que integre las áreas de ventas, marketing, publicidad, etc y gestione toda la información asociada a los clientes.

A partir de la información disponible se extraen muestras de 20.000 clientes y distintas variables para tratar de explicar la probabilidad de que cada cliente sea fiel a la entidad.

- Fueran clientes primeros titulares.
- Fueran titulares con algún contrato vigente.
- Fueran personas físicas (no jurídicas).
- Tuvieran una edad comprendida entre 18 y 90 años.
- No constaran como fallecidos o incapacitados.
- Hubieran contestado al cuestionario de satisfacción. El cual se utiliza como medida de valoración de la entidad y se lleva a cabo en las oficinas.

Las muestras utilizadas para estimar cada uno de los modelos, en el ámbito de negocio, en el ámbito de comunicaciones y en el modelo global son diferentes. En cada uno de los modelos se utiliza el 80 % de los datos para la estimación, es decir, 16.000 clientes y el 20 % para la validación. El motivo de utilizar tres muestras distintas se debe a que no se sabía hasta donde se iba a llegar con el modelo. En primer lugar se hizo el estimó el modelo de negocio, después el de comunicaciones y finalmente el total. Además, el utilizar tres muestras distintas fue una forma de asegurar que las variables elegidas eran las correctas.

2.3. Selección del conjunto de variables

En primer lugar, atendiendo a la información disponible de cada uno de los clientes se valora la creación de un modelo que proporcione una medida de lealtad en tres vertientes distintas: Negocio, Comunicaciones y Redes sociales/web (ámbito tecnológico). El sistema de base de datos de la entidad contiene variables para poder estudiar estos segmentos, especialmente, los ámbitos de negocio y comunicaciones, para los cuales se disponen de datos suficientes.

Sin embargo, las variables disponibles para cuantificar el ámbito tecnológico son escasas así como la disponibilidad de los datos. Por tanto, se decide prescindir del tercer ámbito e intentar incluirlo en parte con el ámbito de comunicaciones, especialmente en las variables permanencia e involucración, que parecen ser las que más se aproximan al ámbito.

El hecho de que el primer objetivo de este trabajo sea crear un modelo que mida la lealtad del cliente, y pueda ser utilizado para campañas posteriores de captación y fidelización del departamento de marketing de la entidad, el diseño de los modelos está basado en parte en el modelo de lealtad de segmentación del cliente RFM (ver Capítulo 1). Por ser un modelo de lealtad muy utilizado en la práctica en campañas de marketing y que permite después segmentar a los clientes, es decir, clasificarlos según estos ámbitos de lealtad y ver como son. Los indicadores que tieneninterés incluir en el modelo son principalmente ***Recencia*** y ***Frecuencia***.Además, un indicador de ***Permanencia*** fue introducida porque se consideró importante para explicar la lealtad y son numerosos los modelos de

lealtad que la consideran, haciendo referencia a ella como la duración de la relación cliente-empresa (Dick & Basu,1994 ; Buckinx et Al.2007). Por último, un indicador ***Involucración*** se considera por incluir en la medida de lealtad una variable más tecnológica que recoja, en particular, la transformación digital. Además de que numerosos estudios aseguran que el nivel de involucración influye en la lealtad (Dick & Basu 1994) haciendo referencia a la "perspectiva actitudinal" del cliente, a la relacionada con los sentimientos del cliente y a los elementos afectivos que la caracterizan.

En la selección de las variables a introducir en el modelo, se llevó a cabo un profundo estudio de las variables disponibles susceptibles de medir los indicadores anteriores: Recencia, Frecuencia, Permanencia e Involucración. Se obtuvieron diferentes variables de cada indicador hasta encontrar la variable adecuada para que encajara en el modelo.

En primer lugar, respecto al indicador ***Frecuencia*** se disponían de dos variables, frecuencia de operaciones en los últimos 3 meses y en los últimos 12 meses. Tras estudiar las dos en el modelo, se obtuvo que al introducir las variables de frecuencia en los ultimos 3 meses en cada ámbito junto con la variable recencia una de las dos no se obtenía como significativa en alguno de los modelos. Esto puede deberse a la existencia de cierto solapamiento entre ambas, en el sentido de que la variable recencia, hace incapié en la importancia del tiempo más reciente. Se opta por utilizar la frecuencia en el periodo de doce meses puesto que permite eliminar ese efecto.

En segundo lugar, para medir la ***Permanencia*** también se barajan diferentes opciones. Se toman las variables: permanencia en años y permanencia en meses, no obstante para interpretar los resultados de forma más coherente se decide optar por la permanencia medida en años que lleva el cliente respecto al ámbito de negocio. En el ámbito de comunicaciones, este indicador trata de reflejar el ámbito digital puesto que la recepción de comunicaciones se realiza por correo electrónico.Se trata de acotar en el tiempo de alguna manera, no obstante los datos de por sí ya están acotados, puesto que si pasan más de 24 meses no se tienen en cuenta.

El indicador ***Involucración*** en primer lugar se concibe más orientada a producto, puesto que en el ámbito de negocio hace referencia al número de productos contratados por el cliente. No obstante, para relacionarlo con la parte emocional entre el cliente y la entidad unido al ámbito digital se consider la variable de si el cliente expresa consentimiento a la recepción de comunicaciones.

La descripción de las variables finalmente utilizadas para incorporar en el modelo cada uno de los indicadores viene detallada a continuación y se resume en la Figura 1:

- ***Recencia***: El concepto *Recencia* hace referencia a una de las formas en las que tiende a operar la memoria. Es un término utilizado en psicología para explicar que la información facilitada al final es la que mejor se recuerda. Este término ha sido extrapolado al ámbito empresarial definido de la siguiente forma: "es la medida de tiempo que ha pasado desde la última transacción que realizó un cliente". Es decir, se entiende que un cliente que recientemente ha tenido contacto con la entidad, está más comprometido y por tanto será más leal a la entidad.

  - *Ámbito de Negocio*: En el ámbito de negocio, este indicador se incluye con el uso de la variable, número de meses que han pasado desde que el cliente realiza la última operación.
  - *Ámbito de Comunicaciones*: Meses que han pasado desde que el cliente ha recepcionado una comunicación, es decir, ha abierto una comunicación recibida por alguno de los canales por los que opera como por ejemplo el email.

- ***Frecuencia***: Viene definido por la RAE como el número de veces que se repite un proceso periódico por unidad de tiempo. En el caso de estudio se trasladaría a la pregunta:

  " ¿Utiliza habitualmente el cliente los servicios que ofrece la entidad? "

Es de esperar que un cliente que tiene mayores niveles de frecuencia tiene una mayor probabilidad de incrementar su fidelidad con la entidad.

  - *Ámbito de Negocio*: Introducido con la variable, número de operaciones que ha realizado el cliente en los últimos 12 meses
  - *Ámbito de Comunicaciones*: Porcentaje de comunicaciones recepcionadas por el cliente sobre el total de comunicaciones enviadas.

- ***Permanencia***: La permanencia se refiere a la duración o estancia en un lugar o sitio. En el caso de estudio se trasladaría a tratar de medir la duración de la relación del cliente con la entidad. Entendiendo que un cliente permanente es más leal con la entidad.

  - *Ámbito de Negocio*: Se introduce con la variable número de años que el cliente lleva en la entidad.
  - *Ámbito de Comunicaciones*: Permanencia en el tiempo en la recepción de comunicaciones medido por la variable " diferencia en el tiempo entre la primera fecha de apertura de comunicaciones y la última"

- ***Involucración***: Se refiere a la acción y efecto de involucrar, al nivel de compromiso que el cliente adquiere con la entidad. Se espera una correlación positiva con la probabilidad de ser leal del cliente.

  - *Ámbito de Negocio*: Número de productos diferentes contratados por el cliente.
  - *Ámbito de Comunicaciones*: Manifiesto expreso del cliente para la recepción de comunicaciones. Se trata de una variable binaria que toma valor 0 si el cliente no manifiesta expresamente la recepción de comunicaciones y 1 en caso de que si lo exprese.

ÁMBITO	VARIABLES			
	RECENCIA	FRECUENCIA	PERMANENCIA	INVOLUCRACIÓN
ÁMBITO NEGOCIO	Meses que han pasado desde que el cliente realiza la última operación	Número de operaciones que ha realizado el cliente en los últimos 12 meses	Número de años que lleva en la entidad el cliente	Número de productos diferentes contratados por el cliente
ÁMBITO COMUNICACIONES	Meses que han pasado desde que ha recepcionado una comunicación	Porcentaje de comunicaciones recepcionadas sobre el total de enviadas	Permanencia en el tiempo en la recepción de comunicaciones	Cliente manifiesta expresamente la recepción de comunicaciones

Figura 1: Variables por ámbito de estudio

De ahora en adelante, nos referimos a las variables elegidas por el nombre del indicaador sobre el que proporcionan información, ya que se ha elegido una variable por indicador. Se elige una variable por indicador en lugar de varias, puesto que se ve más claro el efecto y es más fácil de llevar a la práctica en la entidad, no obstante, podría ampliarse en el futuro.

2.4. Definición de lealtad en la entidad financiera

La variable respuesta se trata de una variable binaria que puede tomar valores 0 o 1. Siendo, 0 si el cliente no es leal y 1 si el cliente es leal. En cada cliente, para conocer el valor de la variable respuesta se tuvieron en cuenta una serie de criterios establecidos por la entidad financiera, los cuales se enumeran a continuación:

- Cumplimiento de la variable antigüedad comercial, es decir, que no hayan salido y entrado de la entidad.
- Cumplimiento en domiciliación de recibos básicos (luz, agua, gas, teléfono), para clientes con más de 18 años.
- Cumplimiento en cuanto a la no realización de transferencias unipersonales o traspasos a otras entidades.
- Valoración de más de nueve puntos en la encuesta de calidad.

La entidad financiera considera que el cliente es leal si cumple todos los puntos anteriores, si alguno de estos criterios es incumplido por un cliente, dicho cliente es considerado por la entidad como no leal.

2.5. Análisis de las variables seleccionadas

En este apartado se va a realizar un análisis descriptivo de las variables seleccionadas en una muestra de 20.000 clientes.

	LEALTAD	MEDIA	DESVIACIÓN TÍPICA	MÍNIMO	MÁXIMO	COEFICIENTE DE ASIMETRÍA	CUARTIL 1º	CUARTIL 3º
RECENCIA	0	1,557	1,898	1	26	8,321	1	1
	1	1,223	1,146	1	25	5,765	1	1
	Total	1,392	1,580	1	26	6,248	1	1
FRECUENCIA	0	62,856	59,613	0	596	1,91	21	87
	1	77,491	66,782	0	542	1,689	30	106
	Total	70,087	63,676	0	596	1,796	25	96
PERMANENCIA	0	19,201	12,135	0	71	0,687	10	26
	1	23,586	12,599	0	77	0,476	13	34
	Total	21,367	12,559	0	77	0,571	12	30
INVOLUCRACIÓN	0	5,262	1,981	1	15	0,468	4	7
	1	6,068	2,157	1	15	0,258	5	7
	Total	5,660	2,108	1	15	0,380	4	7

Figura 2: Medidas descriptivas de las variables en el ámbito de negocio

En la figura anterior se muestran los estadísticos principales de las variables utilizadas en el ámbito de negocio segmentadas por la variable lealtad para los clientes no leales (0 en lealtad) y leales (1 en lealtad) y para el total de clientes sin distinción. Se observa como los clientes leales hace menos tiempo que operaron en promedio que los clientes no leales, así como que realizan más operaciones en promedio. También los clientes leales permanencen aproximadamente 4 años más en promedio que los clientes no leales en la entidad y tienen más productos contratados. Los

clientes no leales permanecen como máximo 71 años en la entidad, mientras que en los leales el máximo se sitúa en los 77.

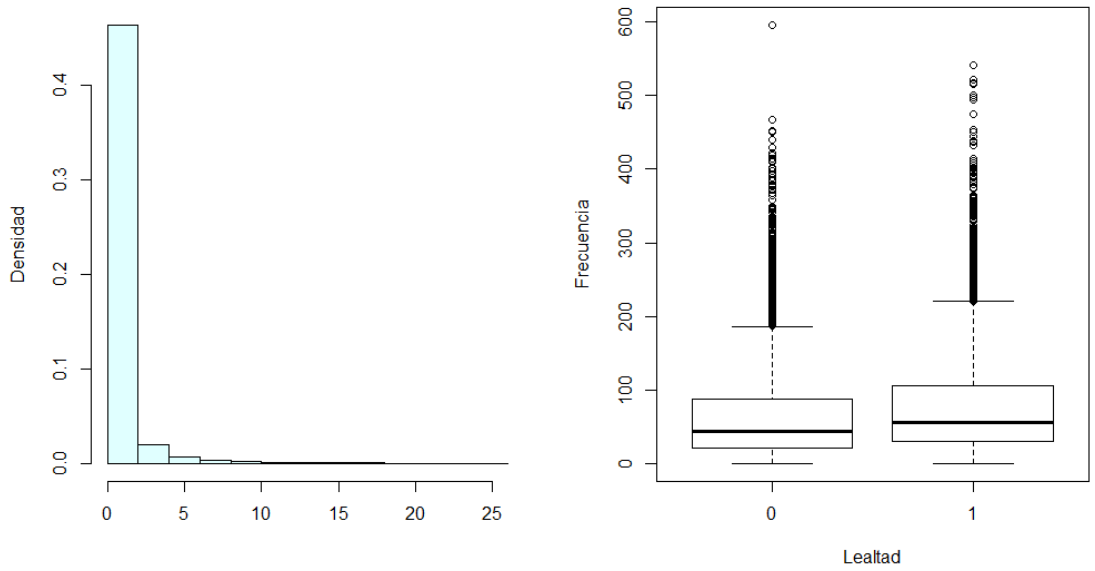


Figura 3: Histograma de recencia y gráfico de caja frecuencia

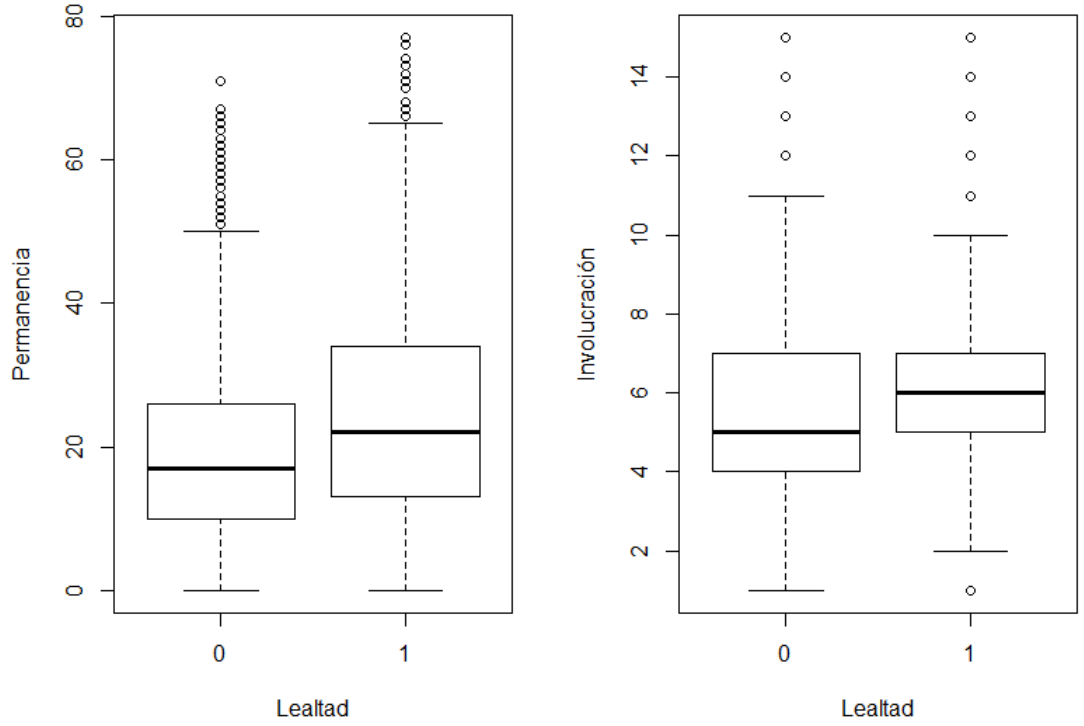


Figura 4: Gráficos de caja permanencia e involucración

A través de los diagramas de caja por grupos se puede identificar de forma visual los valores de las variables y la distribución de las mismas. El primer diagrama de caja muestra que la variable recencia en negocio debe sus valores máximos a valores atípicos puesto que la mediana de los datos se sitúa en 1 mes aproximadamente y el primer y tercer cuartil están muy próximos. Respecto a la frecuencia con la que los clientes realizan operaciones, el comportamiento en ambos grupos es similar, aunque operan con mayor frecuencia los clientes leales. Respecto a la variable permanencia se observan algunos valores extremos en los valores altos. El primer cuartil se situa en los 10 años para los clientes no leales y en 13 para los leales. A si mismo, el tercer cuartil tiene un valor de 26 años para los no leales y de 34 para los leales. Por último, en la variable involucración se observan de nuevo algunos valores atípicos, destaca el valor mínimo de los clientes leales (1 producto contratado), el cual hace que se obtenga el mismo valor mínimo en clientes leales y no leales. Existe un mayor rango intercuartílico en los clientes no leales, por lo que se deduce que los valores de los clientes leales están más concentrados, presentan menos dispersión. Eso también se observa en el coeficiente de asimetría, que presenta un valor menor para los clientes leales.

	LEALTAD	MEDIA	DESVIACIÓN TÍPICA	MÍNIMO	MÁXIMO	COEFICIENTE DE ASIMETRÍA	CUARTIL 1º	CUARTIL 3º
RECENCIA_C	0	17,01	44,37	0	100	-0,1367	11	100
	1	3,881	3,26	0	22	0,6996	1	7
	Total	10,856	42,28	0	100	0,935	3	100
FRECUENCIA_C	0	34,91	42,29	0	100	0,6220	0	80
	1	73,51	22,12	4,34	100	-0,6202	58,065	92,105
	Total	53,01	39,37	0	100	-0,241	0	91,11
PERMANENCIA_C	0	0,458	1,76	0	16	4,439	0	0
	1	9,491	7,89	0	22	0,035	0	17
	Total	4,69	7,15	0	22	1,179	0	10
INVOLUCRACION_C	0	0,51	0,49	0	1	-0,075	0	1
	1	0,88	0,31	0	1	-2,406	1	1
	Total	0,69	0,46	0	1	1,051	0	1

Figura 5: Medidas descriptivas de las variables en el ámbito de comunicaciones

En la figura 5 al igual que para el ámbito de negocio, se muestran los estadísticos principales de las variables utilizadas en el ámbito de comunicaciones para el cliente leal y no leal y para el total de clientes. A simple vista se aprecian importantes diferencias entre un tipo de clientes y otro. Las diferencias más llamativas se encuentran en los valores promedio de las variables, como por ejemplo en la frecuencia de operaciones. Los clientes leales en media recepcionan 73 % de las comunicaciones que se les envían, mientras que en los clientes no leales solo se abren el 34 % de las comunicaciones. A su vez, los clientes leales permanecen mucho más tiempo abriendo comunicaciones que los no leales. Obteniendo estos resultados, a priori, tiene sentido explicar la lealtad por esta colección de variables.

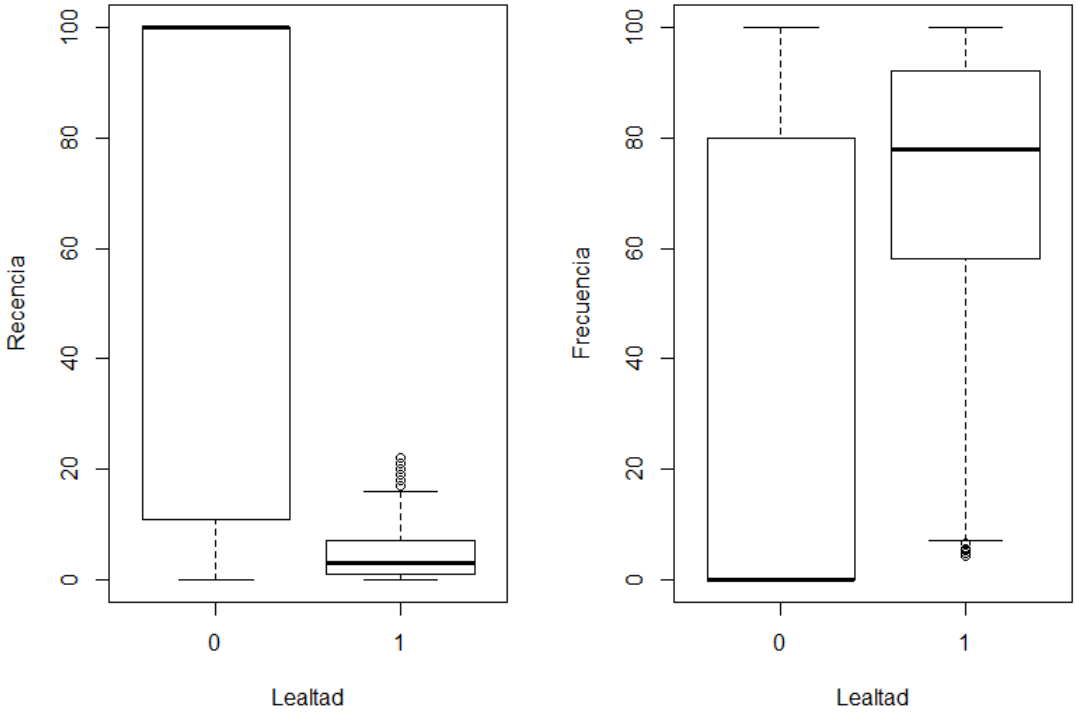


Figura 6: Gráficos de caja recencia y frecuencia

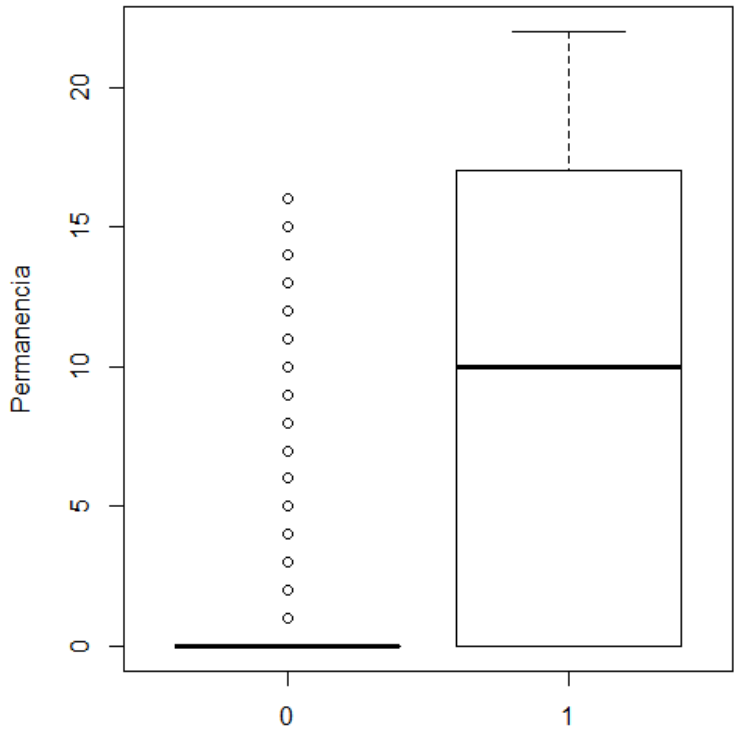


Figura 7: Gráficos de caja permanencia

Observando los diagramas de caja de las variables en comunicaciones se tiene que la variable recencia en los clientes no leales presenta mucha mayor dispersión que en los clientes leales. Se comprobó que ocurría porque cuando un cliente llevaba más de 24 meses sin recepcionar comunicaciones la entidad le ponía valor 100. Respecto a la frecuencia de recepción de comunicaciones, los clientes leales en media recepcionan más del 70 % de las comunicaciones que reciben, mientras que los clientes no leales no llegan al 40 %. Presenta un rango intercuartílico mayor en clientes no leales y por tanto mayor dispersión en los datos. La variable permanencia muestra muy poco rango intercuartílico en los clientes desleales, en media permanecen menos de un año recepcionando comunicaciones y esto es así en la mayoría de la muestra, los datos están muy concentrados. No obstante, se observan valores atípicos altos, lo que repercute en el coeficiente de asimetría.

3. Método de clasificación basado en regresión logística

La clasificación de individuos en grupos es uno de los principales objetivos en minería de datos y es por este motivo que se han desarrollado diferentes técnicas para conseguir este objetivo.

Entre estas técnicas figuran los modelos de regresión logística y los árboles de decisión. El primero se enmarca dentro de los discriminantes lineales, mientras que los árboles de decisión son un tipo de discriminante basado en reglas. El texto de Hastie (2009) incluye ambas metodologías y se describen con detalle en los siguientes apartados. Todo el análisis estadístico se va a realizar con el programa R-3.4.2.Se han incluído en el anexo los códigos utilizados en el análisis. En el primer apartado se va a describir el método y a continuación, se aplica en el caso bajo estudio.

3.1. Descripción del método

El análisis de regresión es una técnica estadística para analizar y modelar la influencia de un conjunto de variables en una variable de interés o variable respuesta. Los modelos de regresión persiguen predecir el valor de la variable respuesta a partir de los valores de un conjunto de variables explicativas.

Los modelos lineales generalizados (GLM) son una unificación de modelos de regresión lineales y no lineales. En un modelo GLM, la variable respuesta debe tener una distribución que pertenezca a la familia exponencial, la cual incluye la normal, Poisson, binomial, exponencial y gamma. Así pues, la regresión logística es un caso particular de estos modelos GLM en la que la variable respuesta es discreta (binaria o con más de dos categorías).

El uso de modelos de regresión logística se ha disparado durante la última década. Desde su aceptación original en la investigación epidemiológica, el método ahora se emplea comúnmente en muchos campos que incluyen negocios y finanzas, criminología, ecología, ingeniería y política de salud.

En este trabajo se van a ajustar modelos de regresión, en los que la variable respuesta tiene solo dos posibles valores, generalmente denominados éxito y fracaso y que vienen denotados por 1 y 0. En el caso real que se va a considerar en este trabajo, el valor 1 identificará al cliente leal. El modelo se formula:

$$g(x) = y = x'\beta + \varepsilon \tag{1}$$

donde, $x^{'} = [1, x_1, x_2, ..., x_k]$, es el vector asociado a las variables explicativas $\beta^{'} = [\beta_0, \beta_1, \beta_2, ..., \beta_k]$, y la variable respuesta y toma valor 0 o 1. Asumimos que la variable y es una variable Bernuilli con distribución de probabilidad:

y	Probabilidad
1	$P(y = 1) = \pi$
0	$P(y = 0) = 1 - \pi$

Tabla 1: Distribución de probabilidad.

Siendo, $E(\varepsilon) = 0$, el valor esperado de la variable respuesta es:

$$E(y) = 1(\pi) + 0(1 - \pi) = \pi \tag{2}$$

Esto implica que:

$$E(y) = x^{'}\beta = \pi \tag{3}$$

La variabler ε es conocida como el error y hace referencia a lo que se desvía la observación de la media condicional. En los modelos de regresión se asume que el error sigue una distribución normal.

En general, cuando la variable de respuesta es binaria, se emplea una función en forma de S monótonamente creciente (o decreciente). Esta función se llama función de respuesta logística y tiene la forma:

$$P(y=1) = E(y) = \pi = \frac{exp(x^{'}\beta)}{1+exp(x'\beta)} = \frac{e^{g(x)}}{1+e^{g(x)}} = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}} = \frac{1}{1+exp(-x'\beta)} \tag{4}$$

La función de respuesta logística es facilmente linealizable. Siendo: $\eta = x'\beta$ el predictor linear cuando η está definido en la transformación:

$$\eta = ln\frac{\pi}{1-\pi} = x'\beta \tag{5}$$

$$g(x) = ln\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{6}$$

Esta transformación se conoce como la transformación logit de la probabilidad π, y el ratio $\frac{\pi}{1-\pi}$ en la transformación se conoce como el odds o índice de probabilidad . La importancia de la transformación es que η posee muchas de las propiedades de un modelo de regresión lineal: que es lineal en sus parámetros, es continuo y tiene rango entre $-\infty$ y $+\infty$ dependiendo del rango de x.

3.1.1. Estimación de los parámetros en los modelos de regresión logística

El método de máxima verosimilitud se utiliza para la estimación de los coeficientes de un modelo de regresión logística. Cada observación de una muestra de n observaciones independientes sigue la distribución de Bernoulli, por lo que la distribución de probabilidad de cada observación de la muestra es:

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i},\ i = 1, 2, ..., n \tag{7}$$

Dado que las observaciones son independientes, la función de verosimilitud es:

$$L(y_1, y_2, ..., y_n, \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{8}$$

Los software libres de estadística utilizan en su mayoría el método de estimación máximo verosímil para el cálculo de los parámetros de los modelos. Es el caso del programa utilizado R.

Dada una colección de variables explicativas es importante determinar si el efecto de todas ellas es significativo o se puede prescindir de alguna. Una prueba de razón de verosimilitud es un constraste que permite comparar un modelo que incluya todas las variables al que se le denomina modelo completo (MC) con un modelo reducido (MR) que contiene un subconjunto de las variables. El estadístico del contraste se define:

$$LR = 2ln\frac{L(MC)}{L(MR)} = 2[lnL(MC) - lnL(MR)] \tag{9}$$

Para muestras grandes, cuando el modelo reducido es correcto, el estadístico de prueba LR sigue una distribución de chi-cuadrado con grados de libertad igual a la diferencia en el número de parámetros entre los modelos completo y reducido. Por lo tanto, si el estadístico de prueba LR (likelihood ratio) excede el punto porcentual α superior de esta distribución de chi-cuadrado, rechazaríamos la afirmación de que el modelo reducido es apropiado. Este contraste se utiliza en el método por pasos que se explica en el siguiente apartado.

3.1.2. Modelización de la regresión logística por pasos

Los criterios para incluir una variable en un modelo pueden variar de una corriente científica a otra. El enfoque tradicional para la construcción de modelos estadísticos implica buscar el modelo más parsimonioso que explique los datos lo suficientemente bien. El motivo por el que se minimizan el número de variables en el modelo es debido a que es más probable que el modelo resultante sea numéricamente estable y se generalice más fácilmente. Cuantas más variables se incluyen en un modelo, mayores son los errores estándar estimados.

Otra corriente más epidemiológica aboga por incluir todas las variables intuitivamente relevantes en el modelo, independientemente de su significación estadística. La razón de este enfoque es proporcionar un control más completo de la confusión posible dentro del conjunto de datos dado. El problema principal con este enfoque es que el modelo puede estar sobre ajustado, produciendo estimaciones numéricamente inestables. El sobreajuste suele caracterizarse por errores estándar estimados poco realistas, lo que puede ser especialmente problemático en los problemas donde el número de variables en el modelo es grande en relación con el número de sujetos y / o cuando la proporción de individuos que presenta la característica respecto al global es cercana a 0.

La selección gradual de variables se usa ampliamente en los modelos de regresión. Actualmente los principales paquetes de software ofrecen una opción para la regresión logística por pasos, que es una herramienta útil y efectiva para el análisis de datos. Cualquier procedimiento paso a paso para la selección o eliminación de variables de un modelo se basa en un algoritmo estadístico que verifica la inclusión de las variables y las incluye o excluye aplicando una regla de decisión fija. La importancia de una variable se define en términos de una medida de la significación estadística para la variable. En la regresión lineal por pasos se usa el test de la F ya que se supone que los errores se distribuyen normalmente. En la regresión logística, se asume que los errores siguen una distribución binomial, y la significatividad se evalúa mediante la prueba de la razón de verosimilitud de la medida de χ^2. Por lo tanto, en cualquier paso del procedimiento, la variable más importante, en términos estadísticos, es la que aporta el mayor cambio en el logaritmo de la función de verosimilitud con respecto a un modelo que no contiene la variable.

El método se describe considerando los cálculos estadísticos que se deben realizar en cada paso del procedimiento. (Hosmer y Lemeshow,2000) A continuación se describe el método de regresión por pasos (Hosmer y Lemeshow, 2000) en el que se consideran p variables independientes, las cuales se considera revisten interés para explicar la variable respuesta:

Paso 0: Este paso comienza con un ajuste del modelo sin la introducción de ninguna variable y una evaluación del logaritmo de la función de verosimilitud, L_0.

A continuación, se ajusta cada uno de los posibles modelos de regresión logística univariable y se comparan sus respectivas funciones logarítmicas de verosimilitud. La variable de mayor importancia es aquella que tiene el p valor más pequeño. Se introduce esta variable.

Paso 1: Comienza con la introducción de la variable en el modelo . Para determinar cuales de las p-1 variables son susceptibles de entrar en el modelo, se calculan los p-1 modelos regresión de

logística introduciendo la variable del paso 0 y cada una de las demás variables y se calculan para ellos los test de máxima verosimilitud. Tras la estimación de estos test se comparan los p valores de estos con el test obtenido en el paso anterior. Si el p valor del test obtenido en el paso 1 es menor que el obtenido en el paso 0 se pasa al paso 2 y sino se detiene le proceso.

Paso 2: es idéntico al paso 1. El programa ajusta los modelos incluyendo la variable seleccionada en el paso anterior y continúa con este proceso hasta que termina.

Paso final: El proceso termina cuando todas las variables están incluídas en el modelo.

3.1.3. Evaluación del ajuste del modelo

Se denotan los valores de muestra observados de la variable respuesta en forma de vector y, donde $y^{'} = (y_1, y_2, y_3, ..., y_n)$. Se denotan los valores predichos por el modelo o valores ajustados como $\hat{y}$, donde $\hat{y}^{'} = (\hat{y}_1, \hat{y}_2, \hat{y}_3, ..., \hat{y}_n)$. Se concluye que el modelo ajusta si: Las medidas de resumen de la distancia entre y y $\hat{y}$ son pequeñas.

Una forma de evaluar el ajuste de un modelo de regresión logística es la construcción de una tabla de clasificación. El modelo ajustado proporciona el valor de la probabilidad estimada para cada individuo de la muestra de pertenecer al grupo de interés. Para obtener el valor de la variable respuesta de interés se ha de proporcionar un punto c, de forma que si la probabilidad estimada en el individuo es superior a c se considera que el individuo pertenece al grupo de interés y el valor de la respuesta y es igual a 1. En otro caso, el valor de y en el individuo es 0. Una forma intuitiva de resumir los resultados de un modelo de regresión logística ajustado es a través de una tabla de clasificación. El valor más comúnmente utilizado para c es 0.5. El atractivo de este tipo de enfoque para la evaluación de modelos proviene de la estrecha relación de la regresión logística con el análisis discriminante cuando la distribución de las variables explicativas es normal multivariante dentro de los dos grupos en los que se clasifican los individuos. Si el modelo predice los miembros de cada grupo de manera precisa de acuerdo con algún criterio, entonces quiere decir que existen evidencias de que el modelo se ajusta correctamente a los datos. Desafortunadamente, este puede ser el caso o no. Puede darse la situación en la que el modelo de regresión logística sea el correcto y sin embargo, la clasificación sea deficiente.

Una clasificación precisa o inexacta no aborda los criterios de bondad de ajuste: que las distancias entre los valores observados y los esperados no sean sistemáticas, y dentro de la variación del modelo. Sin embargo, la tabla de clasificación puede ser útil junto con otras medidas basadas más directamente en los residuos.

Como se menciona anteriormente, el valor más habitual para c suele ser 0,5, no obstante, es conveniente adaptarlo a la distribución de los datos. Para determinar este punto de corte c se llevan a cabo curvas ROC (curvas características de operación). Las curvas ROC proporcionan una descripción más completa de la precisión con la que clasifica cada punto de corte, la cual viene dada por el área bajo la curva. Esta curva traza la probabilidad de clasificar corretamente los casos de éxito (sensibilidad) y la probabilidad de error en los casos de fracaso (1-especificidad).

En resumen, la sensibilidad caracteriza la capacidad del modelo para detectar la presencia de la característica de estudio en sujetos que realmente la tienen. Es decir, clientes leales bien clasificados. Por su parte, la especificidad, mide la proporción de clientes que no son leales a la entidad y que han sido identificados como no leales; es decir la proporcion de no leales correctamente clasificados. Por último, el complementario de especificidad se refiere a los clientes no leales que han sido clasificados como leales.

3.2. Aplicación en el caso real

El procedimiento que se siguió fue el siguiente: el primer modelo estimado fue el modelo de negocio, considerando las variables indicadas en el Capitulo 2. Tras estudiarlo y conseguir su estimación, se pasó al siguiente paso, el modelo de lealtad en el ámbito de comunicaciones. Por último se decidió integrar todas las variables en un modelo único de lealtad y ver como encajarían todas en conjunto.

3.2.1. Ajuste del modelo en el ámbito de negocio

El log odds ratio o logit del modelo sería:

$\hat{g}(x)$= -1.2163-0.1121068***Recencia_N***+0.001685***Frecuencia_N***+0.0243168***Permanencia_N***

+0.1250987***Involucracion_N***

	ESTIMATE	STD.ERROR	Z VALUE	PR(>|Z |)
INTERCEPT	-1.2163679	0.0557508	-21.818	< 2e-16***
RECENCIA_N	-0.1121068	0.0133333	-8.408	< 2e-16***
FRECUENCIA_N	0.0016850	0.0002713	6.210	5.29e-10***
PERMANENCIA_N	0.0243168	0.0012727	19.106	< 2e-16***
INVOLUCRACION_N	0.1250987	0.0082656	15.135	< 2e-16***

Los resultados que se obtienen son alentadores, puesto que todas las variables del modelo aparecen como significativas. El modelo de negocio lo que trata de estimar es la probabilidad que tiene el cliente de ser leal en el ámbito de negocio.

Todas las variables tienen signo positivo excepto ***recencia_N***, lo cual es lógico puesto que indica los meses que han pasado desde que el cliente realiza la última operación. Esto indica que a más nivel de recencia menos leal es el cliente. La variable que más peso tiene en el modelo es ***involucracion_N***. Es decir, el número de productos contratados que tiene el cliente.

3.2.2. Ajuste del modelo de comunicaciones

$\hat{g}(x)$= 0.791905-0.336265***Recencia_C***+0.004232***Frecuencia_C***+0.210888***Permanencia_C***

+1.566412***Involucracion_C***

A continuación se muestra el modelo de comunicaciones estimado por R:

	ESTIMATE	STD.ERROR	Z VALUE	PR(>|Z |)
INTERCEPT	0.791905	0.133358	5.938	2.88e-09 ***
RECENCIA_C	-0.336265	0.008860	-37.954	< 2e-16 ***
FRECUENCIA_C	0.004232	0.001136	3.727	0.000194 ***
PERMANENCIA_C	0.210888	0.007936	26.574	< 2e-16 ***
INVOLUCRACION_C	1.566412	0.074626	20.990	< 2e-16 ***

Este modelo trata de estimar la probabilidad que tiene el cliente de ser leal en el ámbito de comunicaciones.

El modelo presenta todas las variables positivas excepto recencia, al igual que ocurría en el modelo de negocio, puesto que a su vez se interpreta como a más nivel de la variable, menor nivel de lealtad. Se observa que la variable que más importancia tiene en el modelo es de nuevo ***involucracion_C***

Finalmente, para estimar la probabilidad de ser leal en los dos ámbitos conjuntamente, se estimó un modelo de lealtad en el que se incluyen todas las variables utilizadas anteriormente. El modelo que se obtiene es el siguiente:

3.2.3. Ajuste de un modelo global

$\hat{g}(x)$= -3.0537118 -0.1173446***Recencia_N***+ 0.0021471***Frecuencia_N***+0.0256522***Permanencia_N***

+0.334924***Involucracion_N***-0.0034296***Recencia_C***+ +0.0026390***Frecuencia_C***+ 0.0594169***Permanencia_C***

+0.5606071***Involucracion_C***

	ESTIMATE	STD.ERROR	Z VALUE	PR(>|Z |)
INTERCEPT	-3.0537118	0.1110162	-27.507	< 2e-16 ***
RECENCIA_N	-0.1173446	0.0123424	-9.507	< 2e-16 ***
FRECUENCIA_N	0.0021471	0.0003974	5.403	6.54e-08 ***
PERMANENCIA_N	0.0256522	0.0015663	16.378	< 2e-16 ***
INVOLUCRACION_N	0.3671072	0.0117352	31.282	< 2e-16 ***
RECENCIA_C	-0.0034296	0.0009858	-3.479	0.000503 ***
FRECUENCIA_C	0.0026390	0.0007934	3.326	0.000881 ***
PERMANENCIA_C	0.0594169	0.0031587	18.810	< 2e-16 ***
INVOLUCRACION_C	0.5606071	0.0460623	12.171	< 2e-16 ***

Se observa que al introducir todas las variables juntas para estimar el modelo de lealtad total todas las variables aparecen como significativas.

Para comprobar si realmente todas las variables del modelo deben ser incluídas llevo a cabo la construcción del modelo por pasos. El criterio que R tiene para introducir cada variable en el modelo es el AIC o Akaike pero existen otros como el criterio del R^2 ajustado, el de Mallows's o el criterio SBIC.

```
> step(modelo1prueba,scope=list(upper=LEALTAD_TOTAL),
direction="forward",data=datosTrain3, k=2)
Start:  AIC=22183.73
LEALTAD ~ 1

                  Df Deviance   AIC
+ INVOLUCRACION_N  1    18132 18136
+ PERMANENCIA_C    1    19839 19843
+ RECENCIA_C       1    20231 20235
+ RECENCIA_N       1    20614 20618
+ FRECUENCIA_N     1    20639 20643
+ FRECUENCIA_C     1    21372 21376
+ PERMANENCIA_N    1    21522 21526
+ INVOLUCRACION_C  1    21548 21552
<none>                  22182 22184

Step:  AIC=18135.5
LEALTAD ~ INVOLUCRACION_N

                  Df Deviance   AIC
+ PERMANENCIA_C    1    17440 17446
```

```
+ RECENCIA_C       1    17694 17700
+ RECENCIA_N       1    17724 17730
+ INVOLUCRACION_C  1    17847 17853
+ PERMANENCIA_N    1    17882 17888
+ FRECUENCIA_C     1    17932 17938
+ FRECUENCIA_N     1    18008 18014
<none>                  18132 18136

Step:  AIC=17445.86
LEALTAD ~ INVOLUCRACION_N + PERMANENCIA_C

                  Df Deviance   AIC
+ RECENCIA_N       1    17097 17105
+ PERMANENCIA_N    1    17164 17172
+ RECENCIA_C       1    17227 17235
+ INVOLUCRACION_C  1    17283 17291
+ FRECUENCIA_C     1    17289 17297
+ FRECUENCIA_N     1    17407 17415
<none>                  17440 17446

Step:  AIC=17105.3
LEALTAD ~ INVOLUCRACION_N + PERMANENCIA_C + RECENCIA_N

                  Df Deviance   AIC
+ PERMANENCIA_N    1    16843 16853
+ RECENCIA_C       1    16977 16987
+ INVOLUCRACION_C  1    16979 16989
+ FRECUENCIA_C     1    16998 17008
+ FRECUENCIA_N     1    17093 17103
<none>                  17097 17105

Step:  AIC=16852.64
LEALTAD ~ INVOLUCRACION_N + PERMANENCIA_C + RECENCIA_N
+ PERMANENCIA_N

                  Df Deviance   AIC
+ INVOLUCRACION_C  1    16662 16674
+ RECENCIA_C       1    16726 16738
+ FRECUENCIA_C     1    16778 16790
+ FRECUENCIA_N     1    16819 16831
<none>                  16843 16853

Step:  AIC=16673.96
LEALTAD ~ INVOLUCRACION_N + PERMANENCIA_C + RECENCIA_N
+ PERMANENCIA_N + INVOLUCRACION_C

               Df Deviance   AIC
+ RECENCIA_C    1    16581 16595
```

```
+ FRECUENCIA_C  1    16600 16614
+ FRECUENCIA_N  1    16636 16650
<none>               16662 16674

Step:  AIC=16595.43
LEALTAD ~ INVOLUCRACION_N + PERMANENCIA_C + RECENCIA_N
 + PERMANENCIA_N + INVOLUCRACION_C + RECENCIA_C

               Df Deviance   AIC
+ FRECUENCIA_N  1    16563 16579
+ FRECUENCIA_C  1    16576 16592
<none>               16581 16595

Step:  AIC=16578.62
LEALTAD ~ INVOLUCRACION_N + PERMANENCIA_C + RECENCIA_N + PERMANENCIA_N +
    INVOLUCRACION_C + RECENCIA_C + FRECUENCIA_N

               Df Deviance   AIC
+ FRECUENCIA_C  1    16554 16572
<none>               16563 16579

Step:  AIC=16571.76
LEALTAD ~ INVOLUCRACION_N + PERMANENCIA_C + RECENCIA_N
+ PERMANENCIA_N + INVOLUCRACION_C + RECENCIA_C
+ FRECUENCIA_N + FRECUENCIA_C


Call:  glm(formula = LEALTAD ~ INVOLUCRACION_N
+ PERMANENCIA_C + RECENCIA_N + PERMANENCIA_N
+ INVOLUCRACION_C + RECENCIA_C + FRECUENCIA_N
+  FRECUENCIA_C, family = binomial, data = datosTrain3)

Coefficients:
    (Intercept)  INVOLUCRACION_N    PERMANENCIA_C       RECENCIA_N    PERMA
      -3.044881         0.363274         0.060325        -0.106448
INVOLUCRACION_C       RECENCIA_C     FRECUENCIA_N     FRECUENCIA_C
       0.571488        -0.003997         0.001836         0.002371

Degrees of Freedom: 16000 Total (i.e. Null);  15992 Residual
Null Deviance:     22180
Residual Deviance: 16550  AIC: 16570
```

Al aplicar el método de Stepwise logistic regression hacia adelante, es decir, partiendo del modelo sin variables y añadiendo una por una se llega al modelo final. Este método se basa en elegir el modelo que tiene el mínimo AIC. El orden en que se han introducido las variables es: $involucracion\_N$, $permanencia\_C$, $recencia\_N$, $permanencia\_N$, seguidas por $involucracion\_C$, $recencia\_C$, $frecuencia\_N$ y en último lugar $frecuencia\_C$. Se concluye que, entran en primer lugar las variables del ámbito de negocio.

3.2.4. Validación de los modelos

Para comprobar si las variables utilizadas en cada modelo son útilies, es decir, si los modelos mejoran con la introducción de estas, se ha hecho el test de máxima verosimilitud para cada modelo, en el que la hipótesis nula hace referencia a que el modelo es adecuado solamente con la constante frente a la hipótesis alternativa que dice que el modelo mejora con la introducción de las variables. Los resultados obtenidos son los siguientes:

- Test de máxima verosimilitud modelo de negocio.

```
Model 1: LEALTAD ~ RECENCIA_N + FRECUENCIA_N + PERMANENCIA_N
+ INVOLUCRACION_N
Model 2: LEALTAD ~ 1
  #Df LogLik Df Chisq Pr(>Chisq)
1   5 -11994
2   1 -12570 -4  1152  < 2.2e-16 ***
```

- Test de máxima verosimilitud modelo de comunicaciones.

```
Model 1: LEALTAD_C ~ RECENCIA_C + FRECUENCIA_C
 + PERMANENCIA_C + INVOLUCRACION_C
Model 2: LEALTAD_C ~ 1
  #Df   LogLik Df Chisq Pr(>Chisq)
1   5  -3421.1
2   1 -10403.6 -4 13965  < 2.2e-16 ***
```

- Test de máxima verosimilitud modelo de lealtad total.

```
Model 1: LEALTAD ~ RECENCIA_N + FRECUENCIA_N
 + PERMANENCIA_N + INVOLUCRACION_N +   RECENCIA_C
 + FRECUENCIA_C + PERMANENCIA_C + INVOLUCRACION_C
Model 2: LEALTAD ~ 1
  #Df   LogLik Df Chisq Pr(>Chisq)
1   9  -8276.9
2   1 -11090.9 -8  5628  < 2.2e-16 ***
```

Las conclusiones que se obtienen para los tres modelos son las mismas. Al obtenerse un p valor $<$0.5 se rechaza la hipótesis nula que dice que el modelo es adecuado solo con la constante. Los modelos mejoran al introducir las variables.

Una manera de validar los resultados de un modelo de regresión logística es a través de una tabla de clasificación o confusión. Esta tabla es resultado de una clasificación cruzada de la variable explicada con la variable binaria cuyos valores se derivan de las probabilidades estimadas por el modelo determinado. Para obtener la variable binaria estimada, debemos definir un punto de corte c y comparar cada probabilidad estimada con este punto c. Si la probabilidad estimada por el modelo es superior a c, la variable binaria estimada toma valor 1; de lo contrario es igual a 0. El valor más habitual para c suele ser 0,5, no obstante, es conveniente adaptarlo a la distribución de los datos.

Para determinar este punto de corte c se llevaron a cabo curvas ROC (curvas características de operación). Las curvas ROC proporcionaron una descripción más completa de la precisión con la que clasifica cada punto de corte, la cual viene dada por el área bajo la curva. Esta curva traza la probabilidad de clasificar corretamente los casos de éxito (sensibilidad) y la probabilidad de error en los casos de fracaso (1-especificidad).

En resumen, la sensibilidad caracteriza la capacidad del modelo para detectar la presencia de la característica de estudio en sujetos que realmente la tienen. Es decir, clientes leales bien clasificados. Por su parte, la especificidad, mide la proporción de clientes que no son leales a la entidad y que han sido identificados como no leales; es decir la proporcion de no leales correctamente clasificados.

Por último el complementario de especificidad se refiere a los clientes no leales que han sido clasificados como leales.

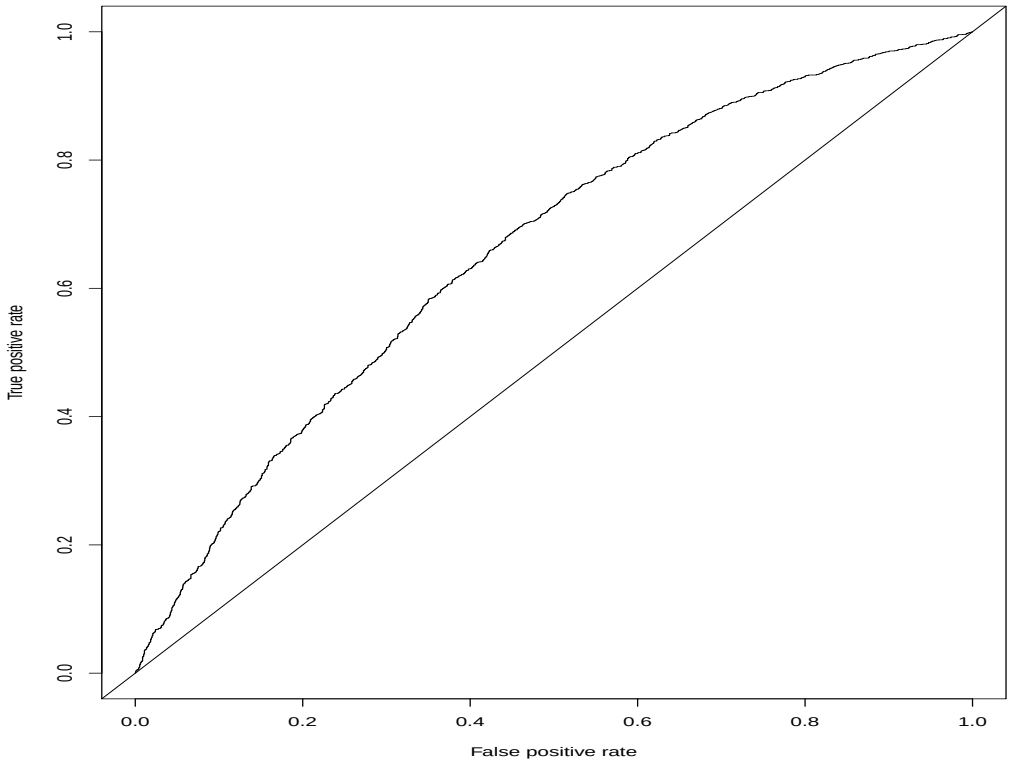


Figura 8: Curva ROC Modelo Negocio

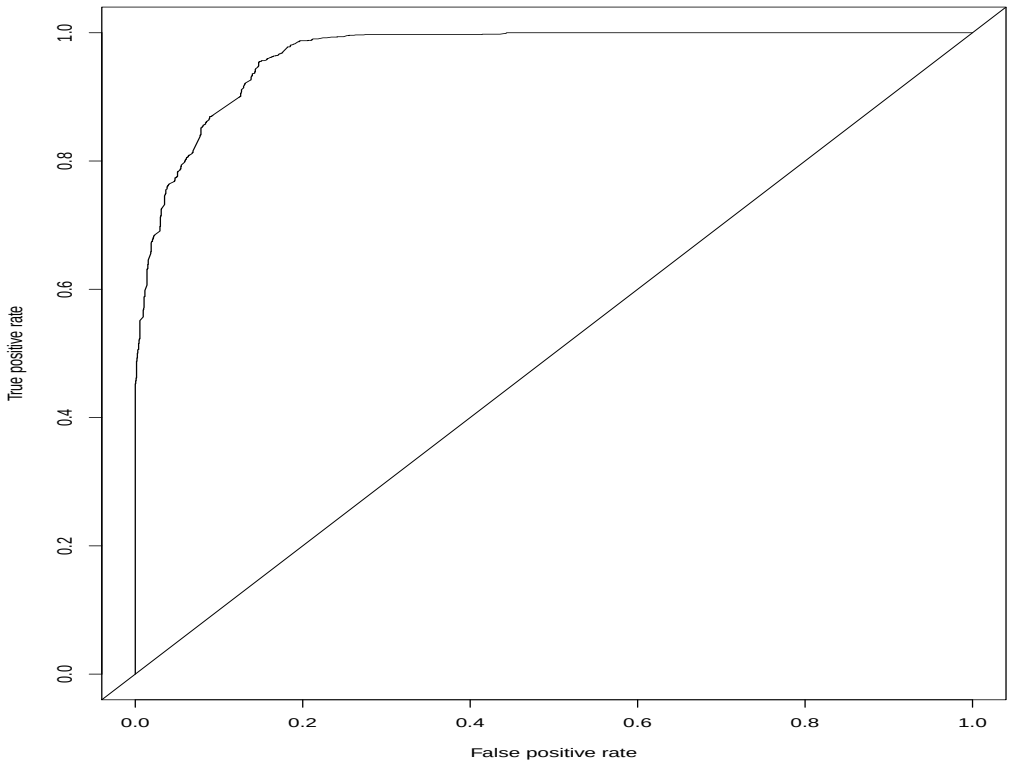


Figura 9: Curva ROC Modelo Comunicaciones

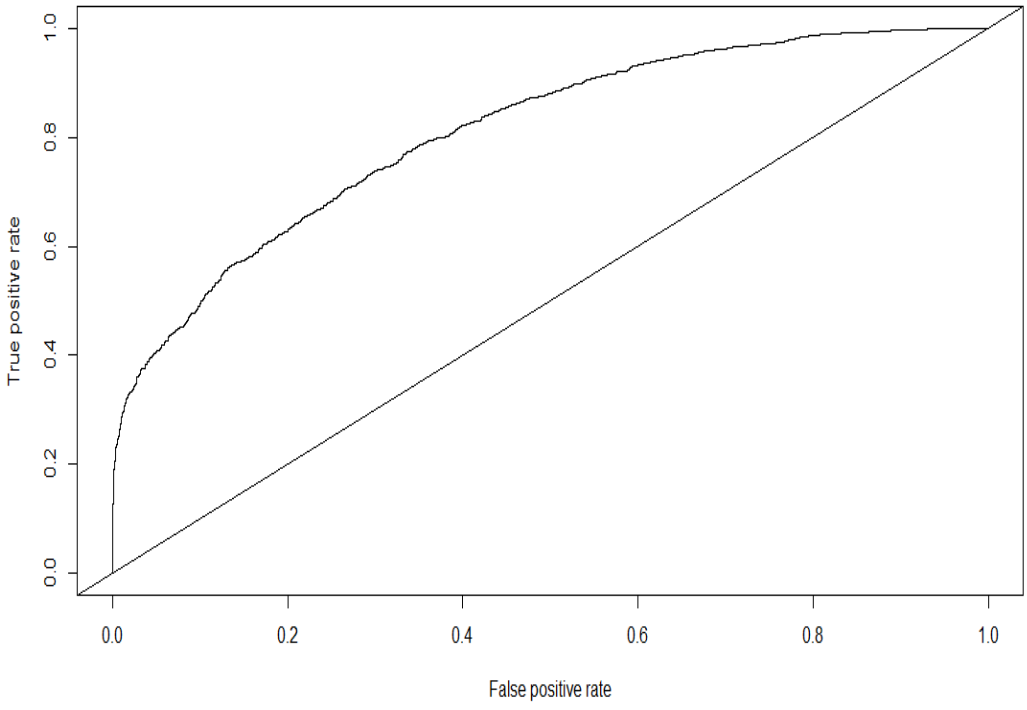


Figura 10: Curva ROC Modelo lealtad total

El área de debajo la curva ROC, proporciona una medida de la capacidad de los modelos para distinguir entre clientes leales y no leales. Si el área bajo la curva valiese 0,5 (50 %, área bajo la diagonal) se consideraría que el modelo no es capaz de distinguir entre leales y no leales, existiría la misma probabilidad de clasificar a un cliente leal como leal que como no leal. Como se puede observar en los gráficos de las curvas estimadas, ambos modelos son capaces de discriminar entre clientes leales y no leales puesto que las cuvas ROC se encuentran por encima de la diagonal. Se observa que en este caso, el modelo que estima la lealtad en el ámbito de comunicaciones tiene una mayor capacidad discriminativa, distingue mejor entre clientes leales y no leales. La curva ROC se formó a partir de distintos puntos de corte que se muestran a continuación:

Punto de corte	Sensibilidad	Especificidad	1-Especificidad
0,40	0,9663073	0,6456212	0,3543788
0,43	0,9636119	0,6558045	0,3441955
0,45	0,9575472	0,6639511	0,3360489
0,47	0,9521563	0,6643177	0,3156823
0,50	**0,9454178**	**0,6965377**	**0,3034623**
0,52	0,9393531	0,7067121	0,293279
0,54	0,930593	0,7169043	0,2830957
0,55	0,9272237	0,7189409	0,2810591
0,57	0,9123989	0,7617108	0,2382892
0,60	0,9043127	0,7780041	0,2219959

Figura 11: Curva ROC Modelo Negocio

Punto de corte	Sensibilidad	Especificidad	1-Especificidad
0,40	0,865625	0,322848	0,677151
0,43	0,788839	0,428134	0,571865
0,45	0,732589	0,494976	0,505024
0,47	0,665178	0,567933	0,432066
0,50	**0,567857**	**0,656618**	**0,343381**
0,52	0,499107	0,703363	0,296636
0,54	0,441071	0,753604	0.246395
0,55	0,412501	0,774574	0,225426
0,57	0,362558	0,814329	0,185670
0,60	0,275892	0,868501	0,131498

Figura 12: Curva ROC Modelo Comunicaciones

El punto de corte c que se eligió para determinar si un cliente es leal o no fue 0,5 en ambos modelos, puesto que en el sensibilidad y especificidad donde se equilibran y es lo que interesa para poder clasificar correctamente al máximo número de clientes. A partir del punto de corte definido en cada modelo los resultados que se obtuvieron en la tabla de clasificación fueron los siguientes:

Modelo de Negocio

	Predicción		Nº de clientes reales
Real	No leal	Leal	
No leal	1.503	786	2.289
Leal	968	1.272	2.240
Total	4.595		3.771

Tabla 2: Clasificación total de clientes modelo de negocio.

	Predicción	
Real	No leal	Leal
No leal	65 %	34 %
Leal	43 %	57 %
Total	60,3 %	

Tabla 3: Bondad del ajuste en el modelo de negocio

Se observa que con el punto de corte fijado en el 50 %, el modelo de negocio clasifica con una mayor precisión a los clientes no leales que a los leales. Es preferible en este caso que el modelo

clasifique mejor a los clientes no leales puesto que aunque se deje el 43 % de los leales sin clasificar a lo que nos lleva el análisis es a tener unos clientes leales más puros y así poder determinar sus características.

Modelo de Comunicaciones

	Predicción		Nº de clientes reales
Real	No leal	Leal	
No leal	1.748	254	2.002
Leal	157	1.612	1769
Total	3.771		3.771

Tabla 4: Clasificación de clientes modelo de Comunicaciones.

	Predicción	
Real	No leal	Leal
No leal	87 %	12 %
Leal	8 %	91 %
Total	89,10 %	

Tabla 5: Bondad del ajuste en el modelo de Comunicaciones

Respecto al modelo de comunicaciones, se observa que estableciendo el punto de corte en el 0.5 % de probabilidad el modelo clasifica a los clientes muy bien puesto que están más separados que en el ámbito de negocio. Estima un 12 % de falsos positivos (clientes no leales clasificados como leales) y un 8 % de falsos negativos (clientes leales estimados como no leales), no obstante el 91 % de los clientes leales los estima como leales.

Modelo de lealtad total

	Predicción	
Real	No leal	Leal
No leal	1.418	572
Leal	555	1.454
Total	4.000	

Tabla 6: Clasificación de clientes modelo de lealtad total.

	Predicción	
Real	No leal	Leal
No leal	71 %	29 %
Leal	28 %	72 %
Total	72,3 %	

Tabla 7: Bondad del ajuste en el modelo de lealtad total

Se observa a través de la tabla de clasificación que el modelo estima un 29 % de falsos positivos y un 28 % de falsos negativos. El poder de clasificación total del modelo es del 72 %.

El problema de utilizar la bondad del ajuste a través de las tablas de clasificación para evaluar los modelos es que esta clasificación es sensible a los tamaños relativos de los dos grupos componentes, resultando favorecida la clasificación al grupo más grande. No obstante, este problema no afecta en el análisis puesto que los tamaños de los grupos resultan muy similares. Además, el objetivo final del modelo es determinar los clientes más leales a la entidad y poder extraer sus características principales por lo que se considera una parte fundamental del análisis.

4. Métodos de clasificación basados en árboles

En este trabajo se va a utilizar un segundo segmento de clasificación sobre la base de datos basado en árboles de decisión. Estos métodos permiten segmentar a los clientes. La referencia de Aluja, proporciona información sobre estos métodos. En particular, dos de los primeros métodos son los denominados AID (Automatic Interaction Detection) y CHAID (Chi Square Automatic Interaction Detection).

- AID: Tiene por objetivo dividir el conjunto de datos en subgrupos cuyos miembros sean lo más homogéneos posibles y a su vez se maximice la diferencia entre los subgrupos. Este algoritmo de clasificación utiliza la variable de respuesta continua y produce árboles binarios. Como criterio de partición utiliza la F y como criterio de parada el umbral de la significación.
- CHAID: Muy similar al AID en su objetivo, el CHAID utiliza como variable respuesta una variable categórica y también produce árboles binarios. Su criterio de partición se basa en la Chi cuadrado y el criterio de parada en el umbral de la significación

4.1. Descripción del método

Conocidos como CART, los árboles de decisión binarios son un método no paramétrico de modelización y segmentación binaria. Un esquema del procedimiento que se sigue para generar el árbol es el siguiente (Aluja,T)

1. Se sitúan todos los datos en el nodo padre o nodo raíz, que es el nodo inicial.
2. Se busca la partición óptima que de lugar a los nodos hijos o subnodos en los que se divide el nodo inicial. El algoritmo trata así de encontrar la variable independiente entre las variables disponibles en el conjunto de datos que mejor separe a estos en grupos. La separación se produce por una regla y en cada división se producen dos grupos excluyentes. En los árboles cuya variable respuesta sea de tipo binaria, se produce una partición por nodo. Esta partición requiere un criterio que determina la impureza del nodo y que indica el grado de homogeneidad entre grupos.
3. En cada nodo hijo obtenido, se decide si se detiene el proceso o se repite el paso anterior. Una vez hecha la separación en dos grupos a partir de la regla se repite el mismo proceso para cada uno de los grupos que han resultado en el proceso anterior. Este proceso se realiza de forma recursiva hasta que no es posible obtener una mejor separación o hasta que los subgrupos alcancen un tamaño mínimo. Estos últimos subgrupos reciben el nombre de nodos terminales u hojas.

En el procedimiento CART aparecen tres pasos que se describen a continuación en detalle (Timofeev, 2004):

1. Construcción del árbol máximo

   El arbol máximo se construye mediante un procedimiento de partición binario, partiendo del nodo raiz, grupo en el que se concentra el conjunto de datos de entrenamiento, como se ha mencionado ya. El arbol que se obtiene suele ser sobreajustado, lo que indica que contiene demasiados niveles y nodos y que hace que resulte demasiado complejo, es por ello que el siguiente paso es la poda del árbol. Los grupos se caracterizan por la distribución de la

variable respuesta en el caso de que esta sea categórica (como es el caso real de estudio) o por la media (variable respuesta continua).

El criterio de partición es la función de impureza, la cual permite determinar la calidad del nodo y se denomina $i(t)$:

$$i(t) = \begin{cases} \Theta(P_j/t) j = 1, ..., m \\ \Psi(Y_i/t) i = 1, ..., n_t \end{cases} \tag{10}$$

$$i(t) = max\, si\, p_j/t = cte$$
$$i(t) = 0\, si\, p_j/t = 0\, \forall_j \neq k\, p_k/t = 1$$
$$i(t) \geq \alpha i(t_r) + (1-\alpha) t(i_j) 0 \leq \alpha \leq 1$$

La selección de la partición óptima se determina maximizando el decremento de impureza:

$$\triangle i(t) = i(t) - \frac{n_{tl}}{n_t} i(t_l) - \frac{n_{tr}}{n_t} i(t_r)$$

Donde, n hace referencia al total de individuos, t a un nodo del árbol T; t_r es el nodo hijo derecho de t y t_l es el nodo hijo izquierdo de t; n_t son los individuos del nodo t; n_{jt} son los individuos del nodo t pertenecientes a la clase j; $p(j/t)$ es la probabilidad de la clase j en el nodo t; T_t es el subárbol descendiente de $\tilde{T}$; $\mid \tilde{T} \mid$ es el número de nodos terminales (tamaño del árbol).

Existen diferentes medidas de impureza, que vienen diferenciadas según la variable respuesta sea continua o categórica. El índice de Gini es el más utilizado cuando la variable es categórica. Este índice trata de separar en cada división la categoría más grande en un grupo a parte, tiene la siguiente expresión:

$$i(t) = \sum_{i \neq j} p(j \mid t) p(i \mid t) \tag{11}$$

Otros índices muy utilizados son el índice de entropia o el índice de Towing.

2. Poda del árbol

La poda del árbol consiste en cortar ramas sucesivas y nodos terminales para encontrar un tamaño de árbol adecuado, puesto que el obtenido primeramente es generalmente sobreajustado. La poda del árbol constituye una alternativa al criterio de parada. Una forma de llevar a cabo el proceso de poda consiste en buscar árboles anidados de tamaños decrecientes, cada uno de los cuales es el mejor de todos los árboles de su tamaño. Estos subárboles son comparados para determinar el óptimo y la comparación se lleva a cabo a través de una función que evalua la complejidad del árbol. Esta técnica se conoce como penalización según la complejidad.

El coste del árbol está basado en la probabilidad de mala clasificación. Siendo $P(j,t) = \frac{n_{tj}}{n_t}$, la probabilidad de la clase j en el nodo t, el coste de un nodo se define como:

$$r(t) = -max_j\, p(j/t)$$

y el coste de un árbol :

$$R(t) = \frac{\sum_{t \in \tilde{T}} p(t) r(t)}{r(raiz)} * 100$$

El árbol se construye de forma que en cada partición se minimiza $R(t)$

Así, la penalización por complejidad consiste en:

$$Min(R(T) + \alpha \mid T_t \mid)$$

donde $R(T)$ es el promedio de la suma de cuadrados entre los nodos, la cual puede ser la tasa de mala clasificación o la suma total del cuadrado de los residuos, dependiendo del tipo de árbol. $\mid T_t \mid$ es la complejidad del árbol, es decir, el número total de nodos del sub-árbol y α es el parámetro de complejidad, el cual indica para cada nodo cuanto se debe penalizar el tamaño del árbol para que no merezca la pena continuar partiendo a partir de ese nodo. Es un número mayor o igual que 0, de tal forma que cuando $\alpha = 0$ se tiene el árbol más grande y conforme α se incrementa, se reduce el tamaño del árbol.

3. Selección del árbol óptimo mediante procedimiento de validación cruzada

   La validación cruzada es una técnica para estimar con precisión el error de predicción. Tiene por objetivo encontrar la proporción entre la tasa de mal clasificados (cociente entre los individuos mál clasificados y el total de individuos)y la complejidad del árbol. El método de validación cruzada se puede aplicar de dos formas:

   - Una de ellas es reservando una parte de la muestra total de los datos para validar y el resto se deja como muestra de aprendizaje. Se construye una secuencia de árboles utilizando la muestra de aprendizaje para cada árbol. Después se predice la respuesta de los datos que se sacaron al inicio del proceso. Por último se estima el error de las predicciones y se selecciona el árbol con el menor error de predicción. No obstante esta aplicación solo se realiza cuando se cuenta con suficientes datos.

   - Normalmente no se cuentan con suficientes datos y se recurre a la validación cruzada con partición en V, (v-fold cross validation). Consiste en dividir la muestra en diez grupos (se suele utilizar V=10) al azar, sacar un grupo de los diez y formar los árboles máximos con los $n, n - n_1, n - n_2, ...$ Después se calcula el error estimado para cada subconjunto . Por último se selecciona el árbol con la menor tasa de mala clasificación.

4.2. Aplicación en el caso real

Además de los modelos de regresión logística se han llevado a cabo árboles de decisión.En este caso se han estimado a través de una técnica de aprendizaje supervisado que se conoce como CART (Classification And Regregsion Trees). Para estimarlos se llevó acabo una división de los datos reservando el 20 % de las muestras para validarlos y el 80 % para la estimación- La librería en R que se utiliza para estimar los árboles es RPART (Recursive Partitioning and Regression Trees)

4.2.1. Árbol de decisión en el ámbito de negocio

Del entrenamiento de los datos para el modelo se obtiene lo siguiente:

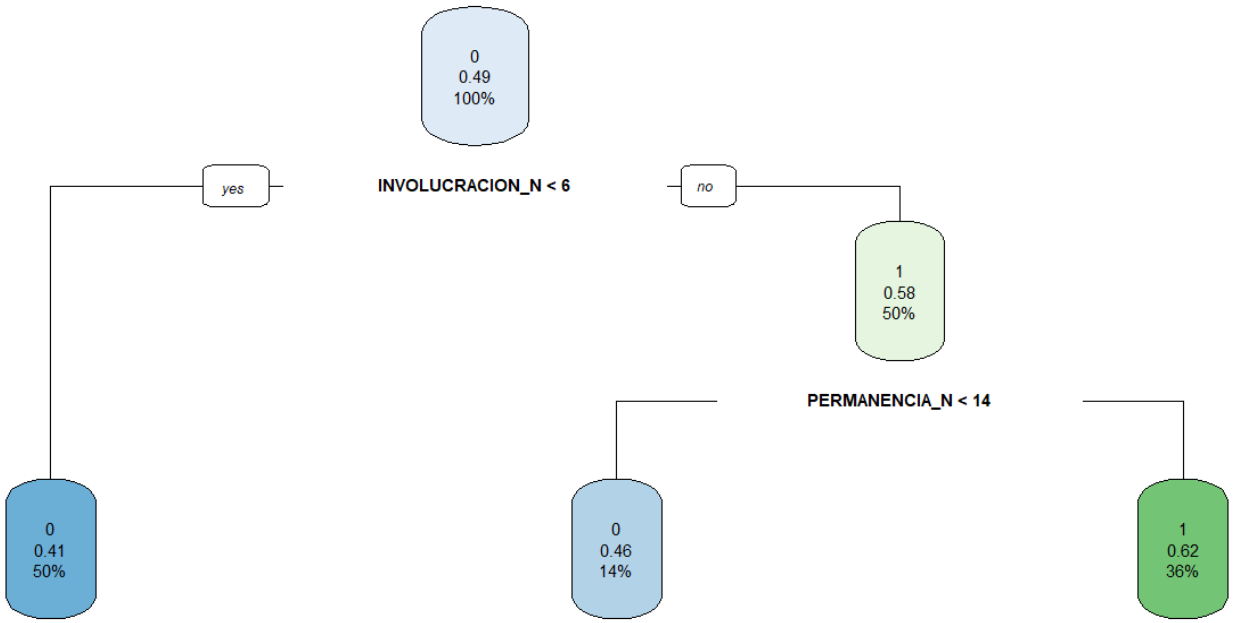


Figura 13: Árbol de decisión modelo de negocio

En el gráfico cada rectángulo representa un nodo y su regla de clasificación correspondiente.

A continuación se muestra el esquema del árbol de clasificación. Cada punto hace referencia a un nodo y a la regla que se aplica. Siguiendo los nodos se llega a las hojas finales del árbol.

```
n= 18145

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 18145 8965 0 (0.5059245 0.4940755)
  2) INVOLUCRACION_N< 5.5 9106 3746 0 (0.5886229 0.4113771) *
  3) INVOLUCRACION_N>=5.5 9039 3820 1 (0.4226131 0.5773869)
    6) PERMANENCIA_N< 13.5 2572 1190 0 (0.5373250 0.4626750) *
    7) PERMANENCIA_N>=13.5 6467 2438 1 (0.3769909 0.6230091) *
```

Partiendo del nodo inidical que contiene todos los datos del conjunto de entrenamiento, la primera variable que utiliza el algoritmo para separar los datos, por ser la que mejor los separa es ***Involucración_N***. La regla que sigue es que si el cliente tiene menos de seis productos contratados lo clasifica como no leal, por contra si tiene más, lo clasifica como leal (clase 1). Este nodo contiene el 50 % de todos los datos. Esta regla clasifica al 41 % como clientes de tipo no leal y 58 % de tipo leal. Esta regla ya produce un nodo terminal par los clientes no leales. Pasando al segundo nivel, se observa que la siguiente variable que utiliza el algoritmo para definir la regla es ***permanencia_N***. Un cliente que lleva en la entidad menos de 14 años lo clasifica como no leal y si lleva más como leal. Del 58 % de los clientes que anteriormente habian sido clasificados como leales, el 46 % pasan a ser no leales según esta regla.

4.2.2. Árbol de decisión en el ámbito de comunicaciones

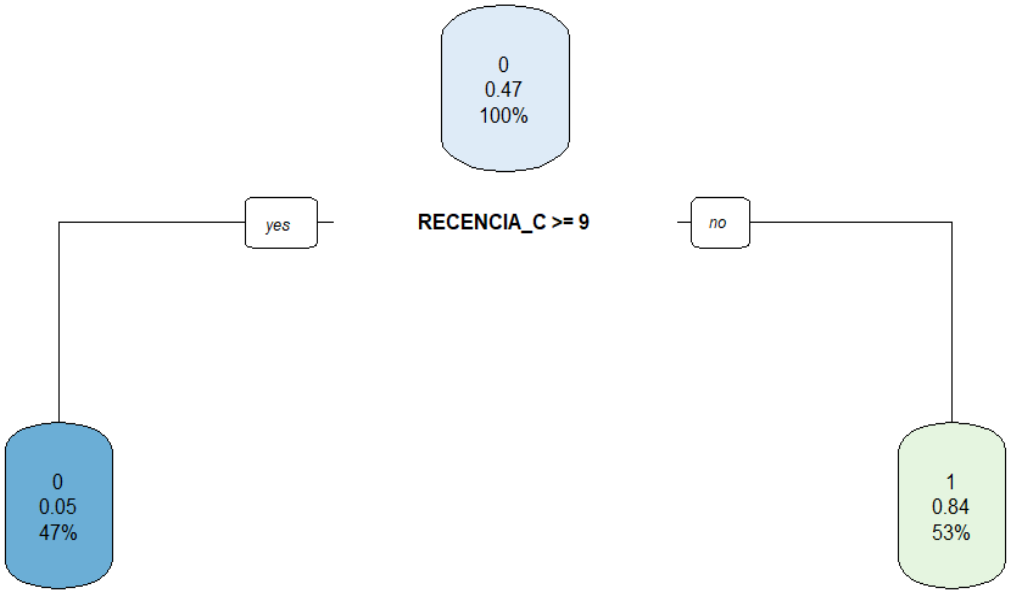


Figura 14: Árbol de decisión modelo de comunicaciones

```
n= 15059

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 15059 7080 0 (0.52984926 0.47015074)
  2) RECENCIA_C>=8.5 7065  370 0 (0.94762916 0.05237084) *
  3) RECENCIA_C< 8.5 7994 1284 1 (0.16062047 0.83937953) *
```

El árbol para el modelo de comunicaciones se observa que clasifica todos los datos bajo una única regla, para la cual utiliza la variable ***Recencia_C***. Esta clasifica a los clientes que llevan un número de meses superior o igual a 9 sin recepcionar comunicaciones como no leales y los que llevan menos de 9 meses como leales. Esto hace que se tenga clasificada al 47 % de la muestra de entrenamiento como no leal y al 53 % como leal.

4.2.3. Árbol de decisión global

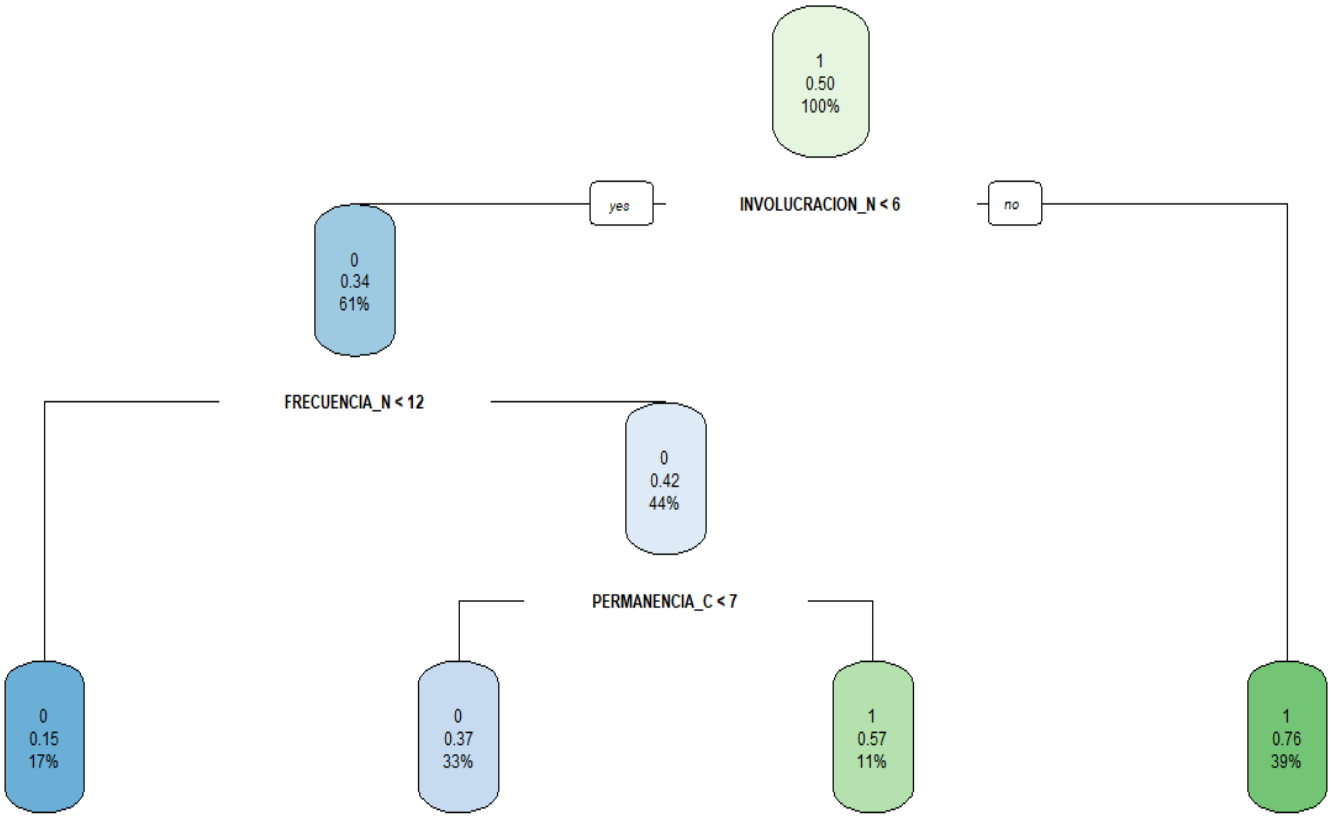


Figura 15: Árbol de decisión modelo de lealtad total

```
n= 16001

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 16001 7962 1 (0.4975939 0.5024061)
2) INVOLUCRACION_N< 5.5 9831 3366 0 (0.6576137 0.3423863)
4) FRECUENCIA_N< 11.5 2715  402 0 (0.8519337 0.1480663) *
5) FRECUENCIA_N>=11.5 7116 2964 0 (0.5834739 0.4165261)
10) PERMANENCIA_C< 6.5 5348 1964 0 (0.6327599 0.3672401) *
11) PERMANENCIA_C>=6.5 1768  768 1 (0.4343891 0.5656109) *
3) INVOLUCRACION_N>=5.5 6170 1497 1 (0.2426256 0.7573744) *
```

Al incluir todas las variales para obtener un árbol que represente el modelo de lealtad total se obtiene que la primera regla que establece a partir del nodo raíz o padre y que da lugar a dos nodos hijos es ***Involucración_N***. Clasifica como no leales a los clientes que tienen menos de seis productos contratados y como leales a los que tienen más de seis, quedando el 61 % de la muestra a un lado y el 39 % al otro, siendo el nodo de clientes leales un nodo terminal. El segundo nivel de

clasificación es la variable ***Frecuencia_N***, que establece como no leales a los clientes que operan menos de 12 veces al mes (17 %) siendo este nodo terminal. A los clientes que operan más de doce veces en el último año y llevan más de 7 meses recepcionando comunicacioens los clasifica como leales. Por último los que llevan operan más de doce veces al año, pero no cumplen con más de siete meses abriendo comunicaciones los clasifica como no leales.

4.2.4. Validación de los modelos

4.2.5. Modelo de Negocio

	Predicción	
Real	No leal	Leal
No leal	1.709	585
Leal	1.228	1.013
Total	4.535	

Tabla 8: Clasificación de clientes modelo de Negocio.

	Predicción	
Real	No leal	Leal
No leal	74 %	26 %
Leal	55 %	45 %
Total	61 %	

Tabla 9: Bondad del ajuste en el modelo de Negocio

Se obtiene a través del análisis realizado que el modelo de negocio clasifica con menor tasa de error a los clientes no leales que a los clientes no leales, es decir se equivoca menos con los clientes no leales. Ocurre algo que ya se esperaba y muy similar ala clasificación obtenida a través del algoritmo de regresión logística.

4.2.6. Modelo de Comunicaciones

	Predicción	
Real	No leal	Leal
No leal	1.695	326
Leal	79	1.664
Total	3.764	

Tabla 10: Clasificación de clientes modelo de Comunicaciones.

	Predicción	
Real	No leal	Leal
No leal	83 %	17 %
Leal	1 %	99 %
Total	89,2 %	

Tabla 11: Bondad del ajuste en el modelo de Comunicaciones

Los resultados para la clasificación de los clientes en el árbol de comunicaciones son similares a los obtenidos para regresión logística. No obstante, la clasificación de clientes en el árbol acierta más en los clientes leales, mientras que comete mayor error en los clientes no leales puesto que se obtiene un 12 % de clientes no leales clasificados como leales en regresión logística contra un 17 % en el árbol de clasificación.

4.2.7. Modelo de lealtad Global

	Predicción	
Real	No leal	Leal
No leal	1402	586
Leal	588	1423
Total	3.764	

Tabla 12: Clasificación de clientes modelo de lealtad total.

	Predicción	
Real	No leal	Leal
No leal	70 %	30 %
Leal	29 %	71 %
Total	70,6 %	

Tabla 13: Bondad del ajuste en el modelo de lealtad total

Finalmente se obtienen los resultados de la clasificación de los clientes según el árbol global. En él se obtiene una clasificacion muy similar de ambos grupos y el poder de clasificación global es prácticamente el mismo que en el método de clasificación de regresión logística.

5. Segmentación del cliente a partir de los modelos de lealtad

Tras la estimación de los modelos y la clasificación de los distintos usuarios de las muestras, el modelo se extrapoló a la base de datos general de la entidad con el objetivo de poder clasificar a todos los clientes y poder llevar a cabo determinadas políticas de captación y fidelización de clientes. Se aligieron aquellas variables de interés sobre características descriptivas del cliente leal con el objetivo de poder determinar su perfil. El análisis de los datos se llevó a cabo con los programas R, excel y tableau.

5.1. Descripción de las variables de segmentación

A continuación se definen las variables utilizadas en la segmentación del cliente, las cuales fueron definidas para los clientes leales en negocio, leales en comunicaciones y leales en lealtad total/global. Son variables descriptivas del cliente en cuanto a su edad, zona geográfica de procedencia, beneficios que deja en la entidad...

- Puntos de vinculación: Se trata de una variable que toma valores entre 0 y 28 en función de los puntos de vinculación que puede tener un cliente (0-28).
- Margen de beneficio Variable categórica dividida en 6 categorías. Estas categorías tienen los siguientes tramos: <0; 0-60; 61-300; 301-600; 601-1500;>1500

  Esta variable indica el margen de beneficio que deja cada cliente a la entidad.
- Patrocinio: Se trata de una variable binaria que toma valores 0 o 1 en función de si el cliente ha participado en eventos (fundamentalmente carreras populares de la entidad) o no. Toma valor 0 si no participa y 1 si participa.
- Profesión: Variable que puede tomar diferentes categorías en función de la profesión del cliente, algunas de las que recoge son: trabajador cualificado; agricultor, médico, profesor, funcionario...
- Uso oficina: Es una variable categórica según la frecuencia de uso del cliente de la oficina. Puede tomar las siguientes categorías: inoperante, anual, semestral, trimestral, mensual, semanal y diaria.
- Territorial: Variable que determina la zona geográfica de pertenencia del cliente. Se divide en las siguientes categorías: Aragón; Arco Mediterráneo; Extremadura y zona sur; Madrid y norte; Rioja-Burgos-Guadalajara.
- Provincia: Variable que contiene por categorías las 52 provincias de España.
- Carterización: Variable binaria que toma valor 0 si el cliente no es carterizado (su gestor no es banca personal) y 1 si es carterizado (si su gestor es del grupo banca personal).
- Renta: Variable que indica el nivel de renta de cada cliente. Es una variable categórica que viene segmentada por niveles de renta. Se distribuye según las siguientes categorías: Bajo, medio-bajo; medio; medio-alto; alto; muy alto.
- Edad: Indica la edad del cliente de la entidad. Puede tomar valores entre 18 y 90 años.

- Grupo Estratégico: Variable que indica la pertenencia de cada cliente a un determinado grupo que en la entidad determinan como estratégico. Se divide en tres grupos: Banca personal; pontenciales y promesas y resto de familias y particulares.
- Segmento Estratégico: Variable categórica que indica la pertenencia de cada cliente a un determinado segmento estratégico. Se divide en las siguientes categorías: Gestión comercial; economías domésticas 50-64; economías domésticas ≥ 65

5.2. Estadística descriptiva variables segmentación

A partir de un análisis descriptivo de estas variables se pueden determinar las características principales del cliente leal en cada ámbito. El análisis se llevo a cabo en la base total de clientes de la entidad que supera los 3 millones.

- PUNTOS DE VINCULACIÓN

Figura 16: Puntos de vinculación total de clientes

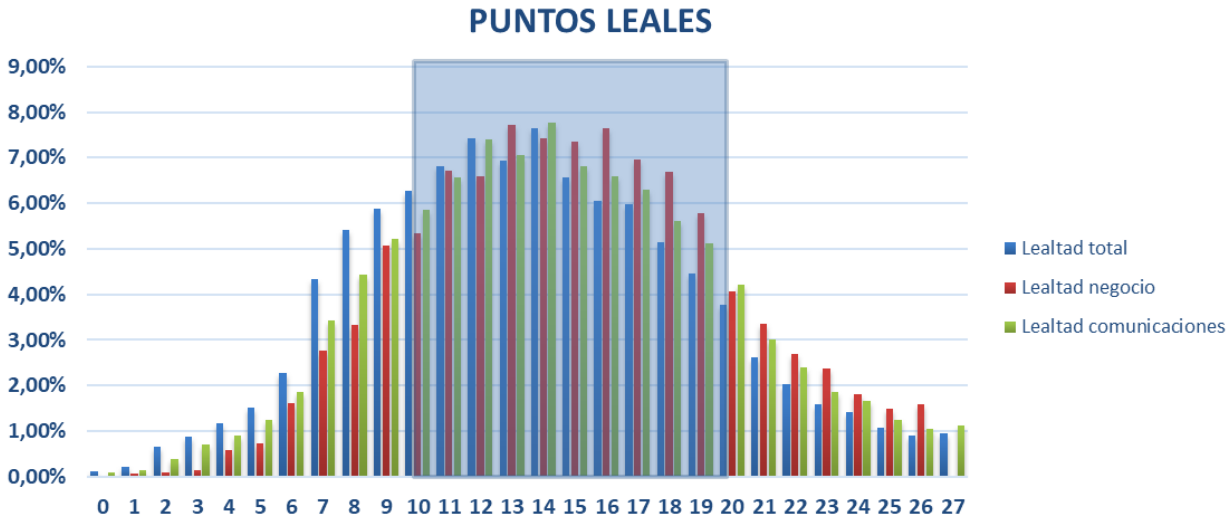


Figura 17: Puntos de vinculación clientes leales

Los gráficos anteriores muestran la distribución de los puntos de vinculación, por un lado en los clientes totales de la entidad y por otro en los clientes leales. Se observa a través de ellos que los clientes leales se concentran en puntos más altos que el resto de clientes de la entidad, es decir, tienen un mayor grado de vinculación con la entidad.

- MARGEN DE BENEFICIO

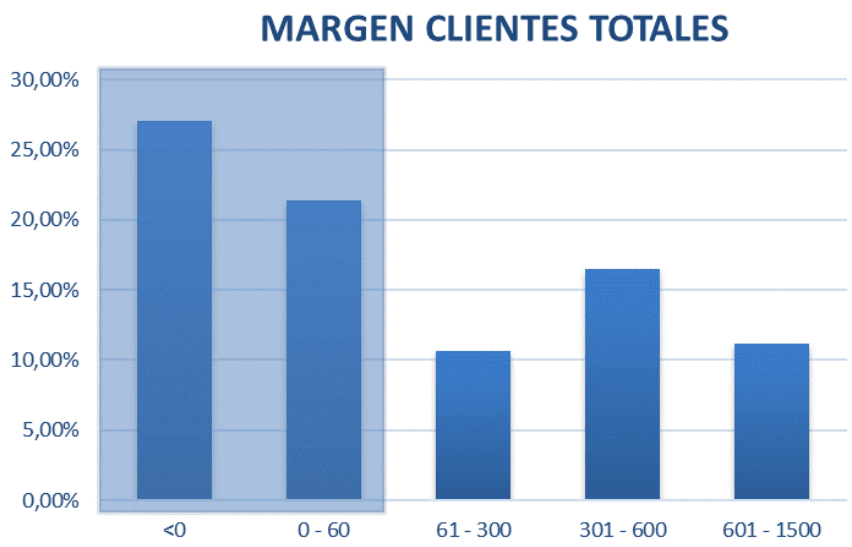
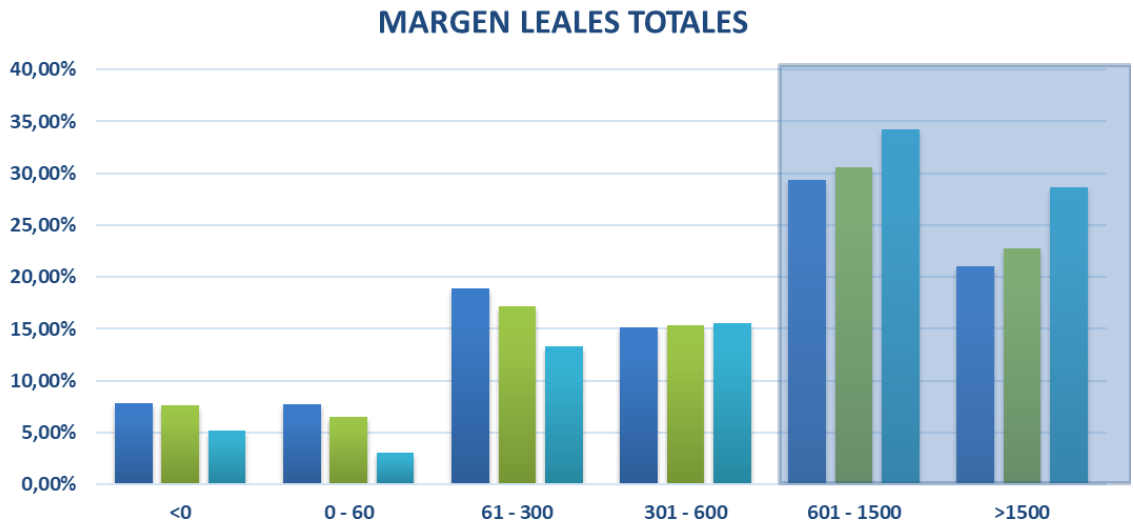


Figura 18: Margen clientes totales



heightheight

Figura 19: Margen clientes leales

A través de estas gráficas de márgenes de beneficio de todos los clientes y de los clientes más leales se concluye que los clientes leales dejan mayores márgenes de beneficio bruto, puesto que se sitúan en tramos altos, principalmente en el tramo de 601-1500 y $\geq$ 1500 €

- PATROCINIO

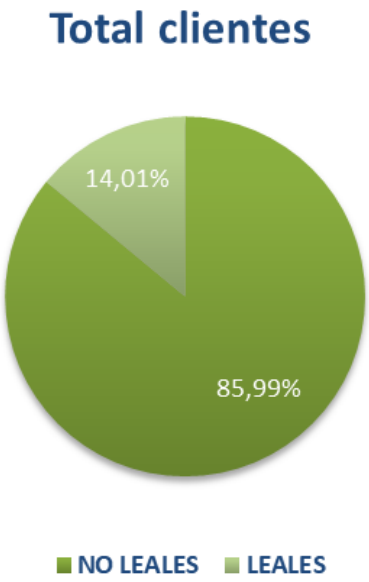


Figura 20: Distribución lealtad total de clientes

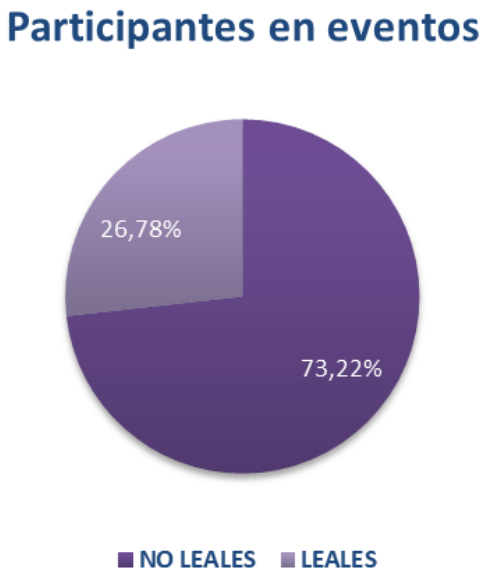


Figura 21: Distribución lealtad participantes en eventos

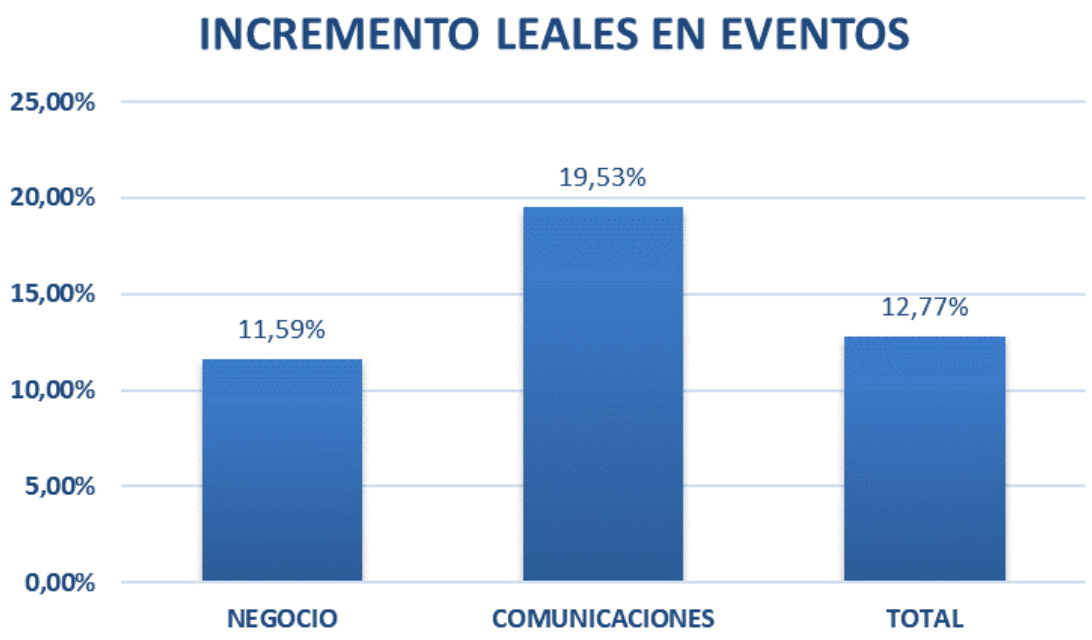


Figura 22: Incremento leales participantes en eventos

Si se compara el porcentaje de leales dentro del total de clientes y el porcentaje de leales dentro del total de clientes participativos en eventos (principalmente carreras populares), se observa que el porcentaje de leales es mayor en el segundo caso. Por lo tanto, se puede determinar que los clientes leales son más participativos en eventos que organiza la entidad.

La figura 8 muestra el incremento en porcentaje de leales en cada ámbito en la participación de eventos. El incremento más notable se obtiene por los clientes leales en comunicaciones, hecho que era de esperar puesto que estas actividades son altamente patrocinadas por los distintos medios de los que dispone la entidad.

- PROFESIÓN

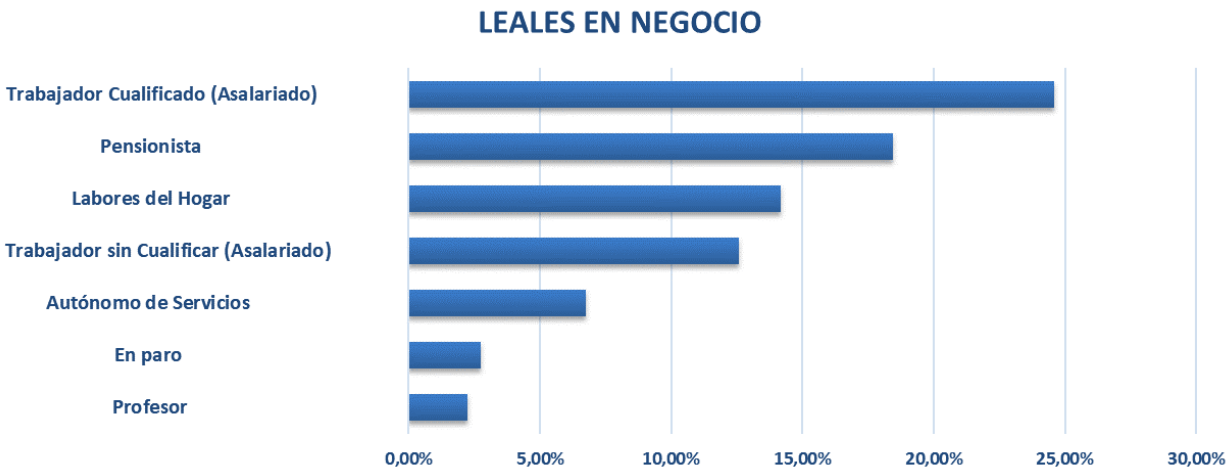


Figura 23: Profesiones clientes leales en negocio

Figura 24: Profesiones clientes leales en comuniceaciones

Las figuras 23 y 24 muestran las profesiones que más se repiten entre los clientes leales en el ámbito de negocio y en el ámbito de comunicaciones. Las profesiones más comunes son similares en los dos ámbitos de lealtad: en primer lugar se encuentran los trabajadores cualificados, seguido por los pensionistas. También están entre los clientes leales los agricultores y cargos directivos. Estos resultados parecen indicar que el cliente leal tiene un perfil de edad madura.

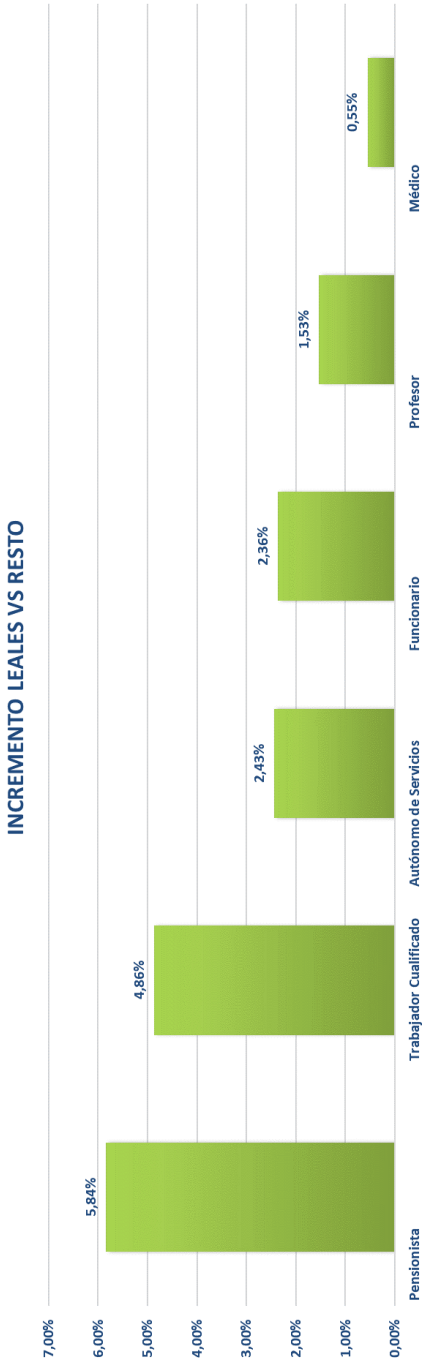


Figura 25: Incremento leales por profesiones

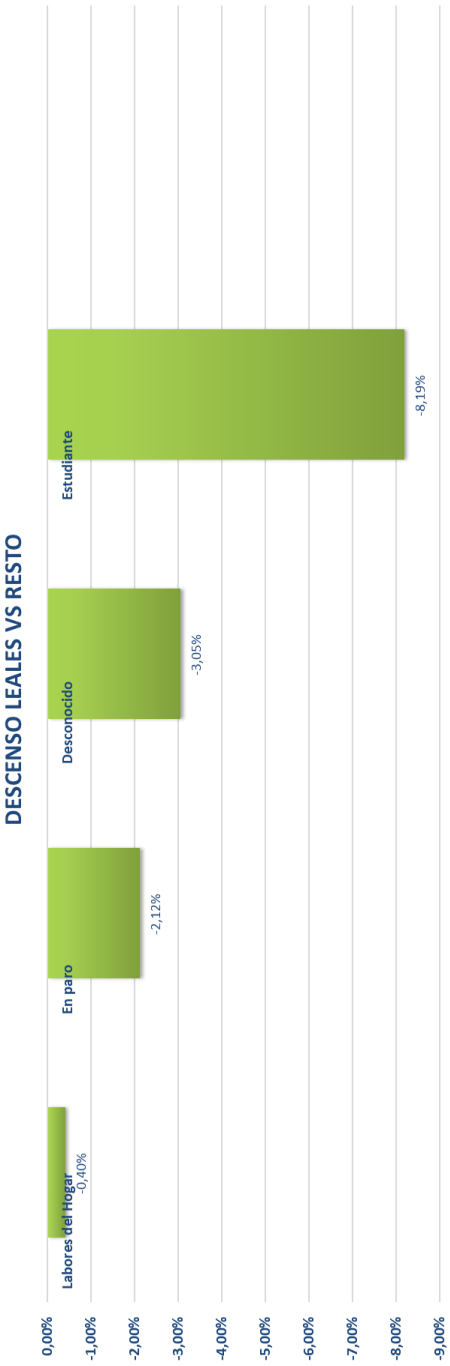


Figura 26: Descenso leales por profesiones

Las figuras 18 y 19 indican los principales incrementos y descensos de clientes leales en lo que a profesiones se refiere. Los datos dicen que el mayor incremento de clientes leales se produce en los pensionistas (5,84 %), seguido por los trabajadores cualificados (4,86 %), autónomos de servicios (2,43 %), funcionarios (2,36 %) y con un incremento más leve profesores y médicos, 1,53 % y 0,55 % respectivamente.

Respecto a las profesiones en las que se produce una mayor caída de clientes leales son principalmente estudiantes, clientes en paro y clientes que se dedican a labores del hogar, lo que parece bastante coherente.

- USO OFICINA

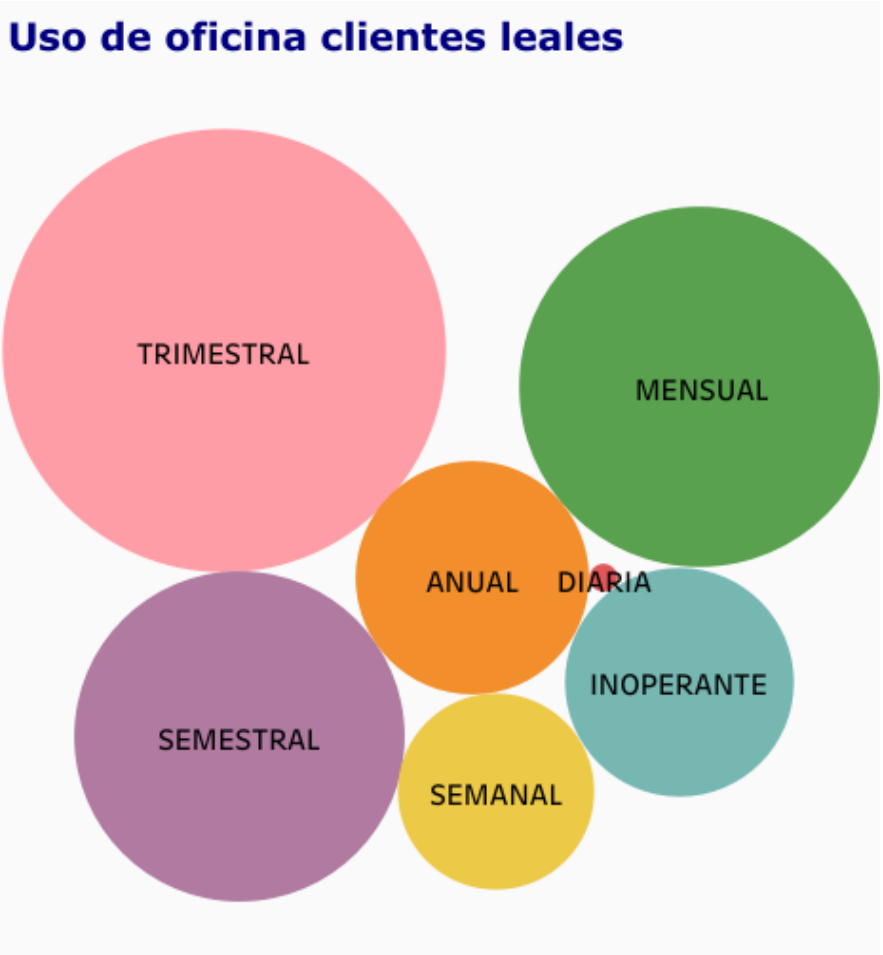


Figura 27: Uso de oficina clientes leales

Respecto a la variable uso de oficina, el principal uso que hacen de ella los clientes leales es trimestral, en segundo lugar mensual y por último semestral. Se observa que las frecuencias más cortas (semanal y diaria) son las mas pequeñas, lo que indica que el cliente leal no visita con mucha frecuencia la oficina.

- TERRITORIAL

Figura 28: Territorio clientes leales

Se observa a través del gráfico que el cliente leal procede principalmente de Aragón y de la zona Madrid y norte, seguidos por la zona de La Rioja, Burgos y Guadalajara y por el Arco Mediterráneo.

- PROVINCIA

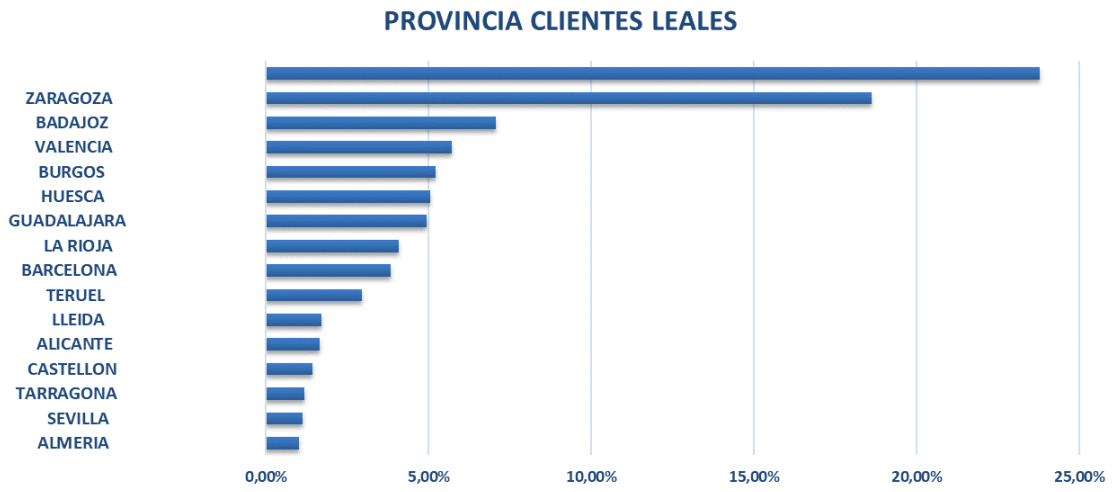


Figura 29: Provincia clientes leales

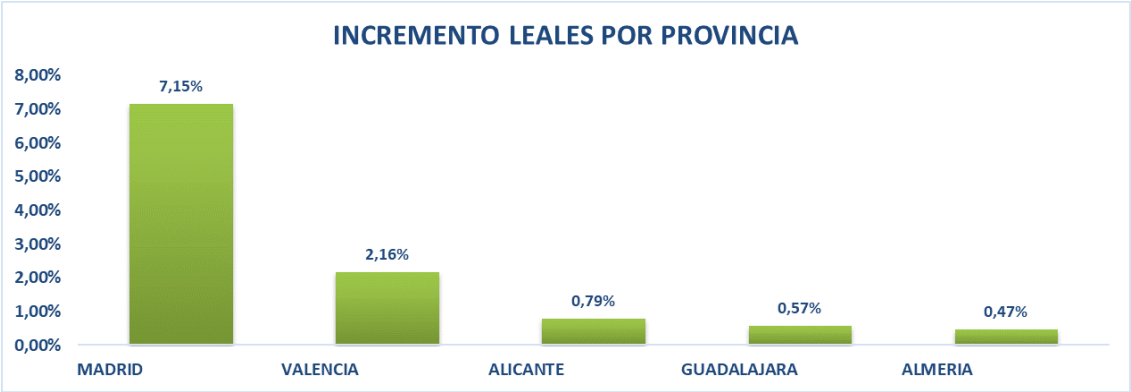


Figura 30: Principales incrementos de clientes leales por provincia

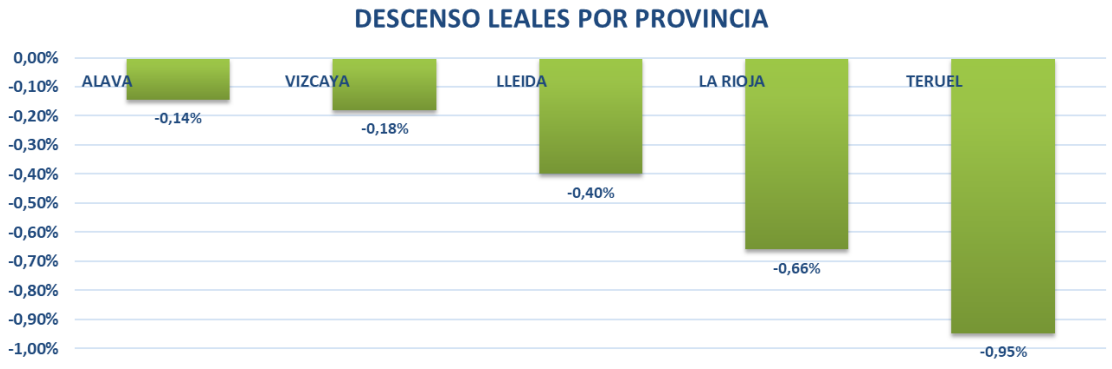


Figura 31: Principales descensos de clientes leales por provincia.

Entre las profesiones de procedencia de los clientes leales destacan: Zaragoza, Badajoz,Valencia... Donde se producen un mayor incremento de clientes leales es en Madrid, Valencia, Alicante, Guadalajara. El mayor descenso por el contrario se da en Teruel y la zona norte en su mayoría (La Rioja, Lleida, Vizcaya y Álava).

- CARTERIZACIÓN

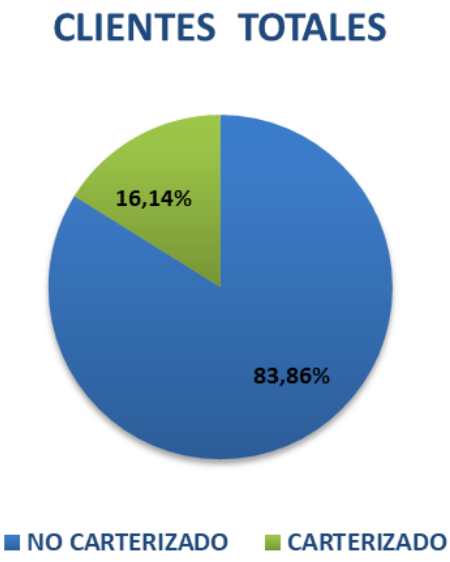


Figura 32: Carterización clientes entidad

Figura 33: Carterización clientes leales en negocio

Figura 34: Carterización clientes leales en comunicaciones

A través de la variable carterización se observa que existe un mayor porcentaje de clientes carterizados dentro de los clientes leales que en la entidad en general, el 55 % y el 47 % de los clientes leales en negocio y leales en comunicaciones son carterizados, mientras que dentro de los clientes totales solo el 16 % de ellos son carterizados.

- RENTA

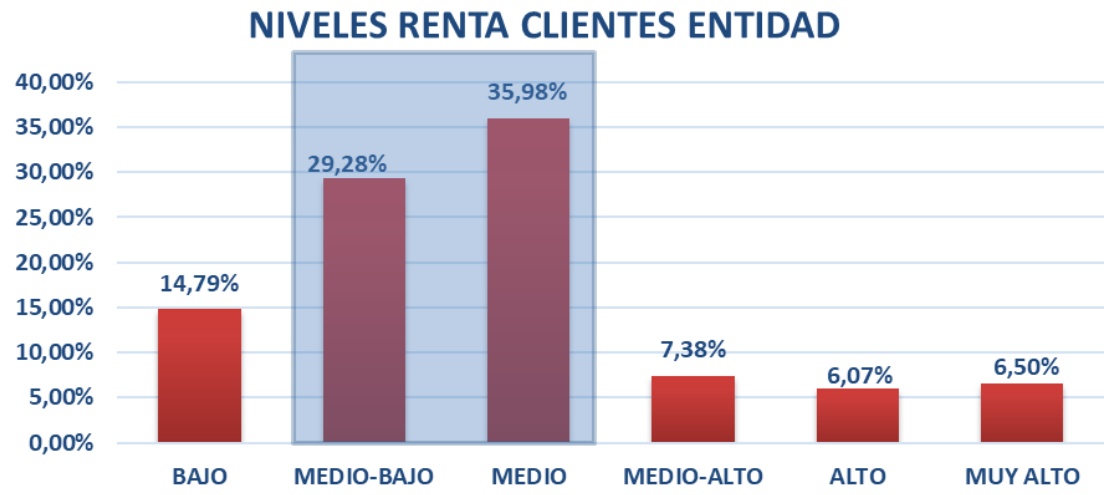


Figura 35: Niveles de renta clientes entidad

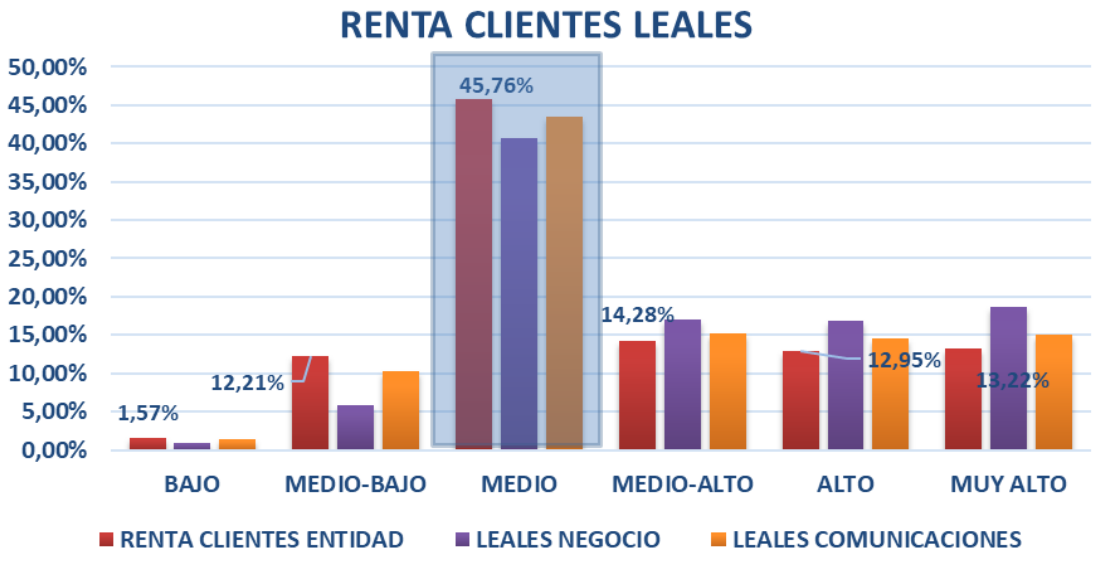


Figura 36: Niveles de renta clientes leales entidad

Se comprueba a través de los gráficos que los clientes leales tienen unos niveles de renta más altos que el resto de clientes, ya que mientras los clientes en su mayoría se sitúan en los tramos de renta medio-bajo y medio; los clientes leales se sitúan en tramos medios principalmente seguidos por los tramos altos.

- EDAD

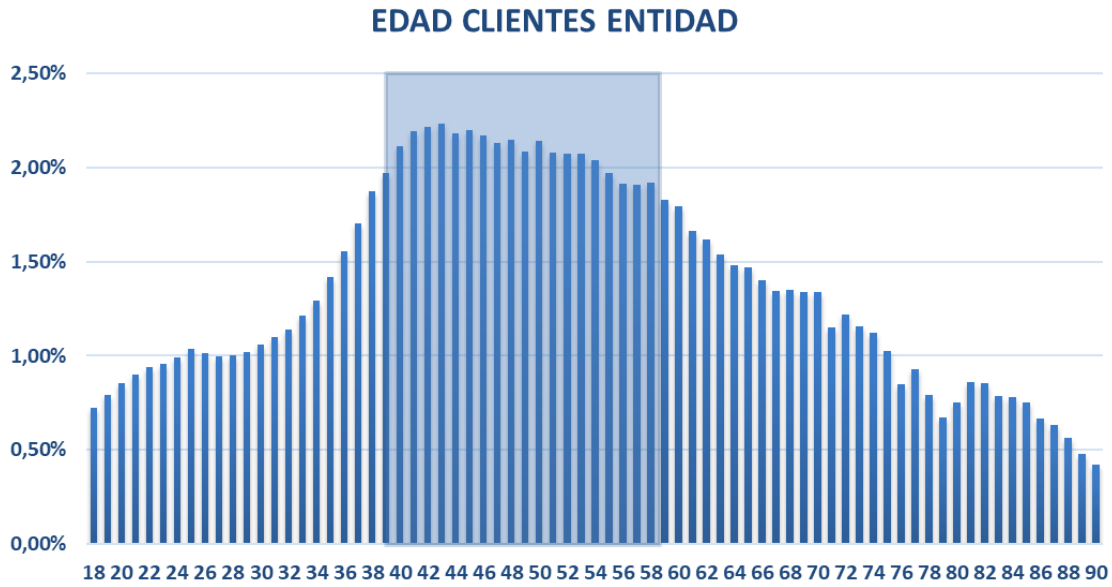


Figura 37: Edad clientes entidad

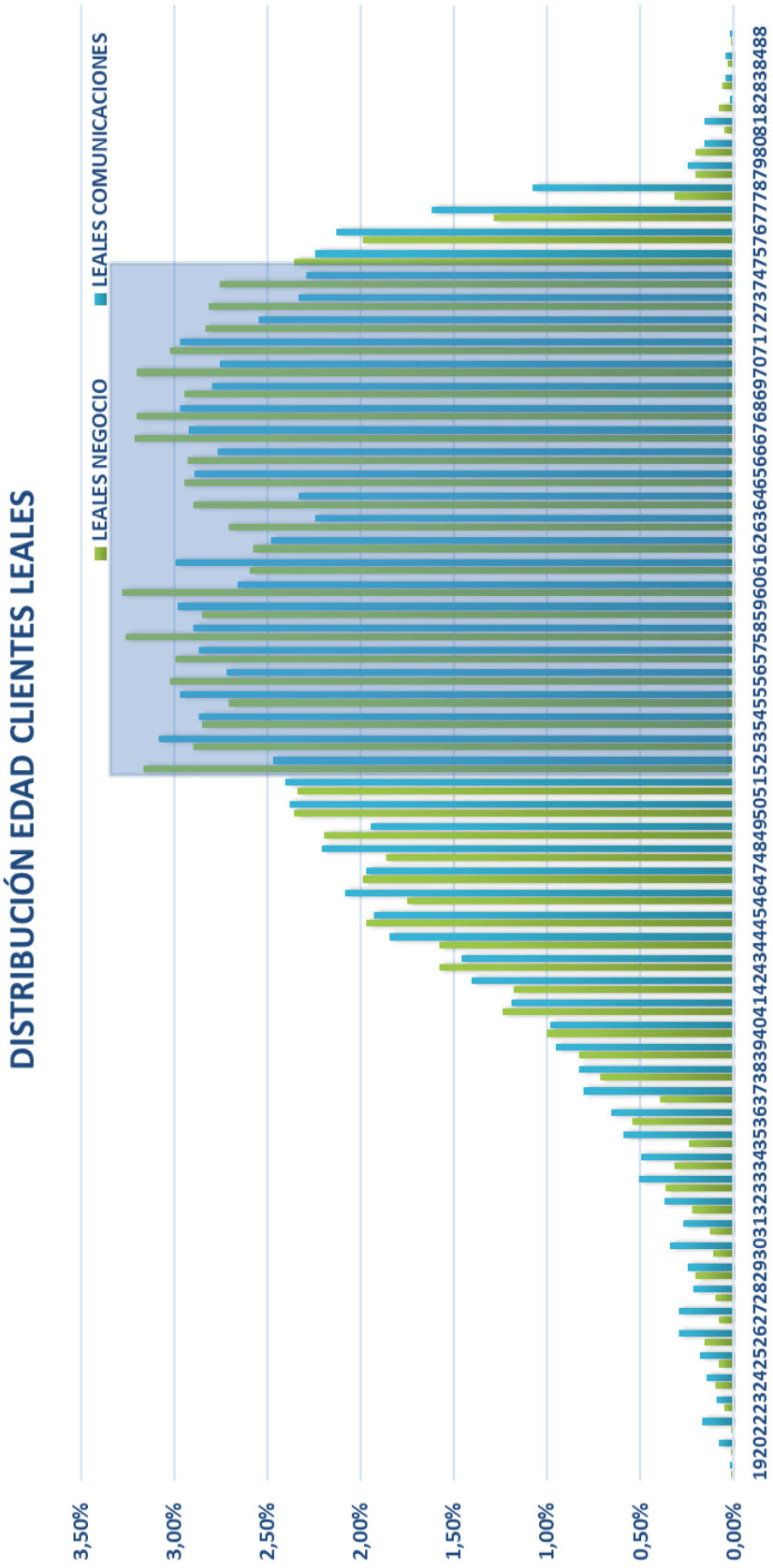


Figura 38: Edad clientes leales

El grueso de los clientes de la entidad se concentra entre los 40 y 60 años, mientras que los clientes leales se concentran entre los 50 y 75 años por lo que se concluye que los clientes leales son más maduros.

- GRUPO ESTRATÉGICO

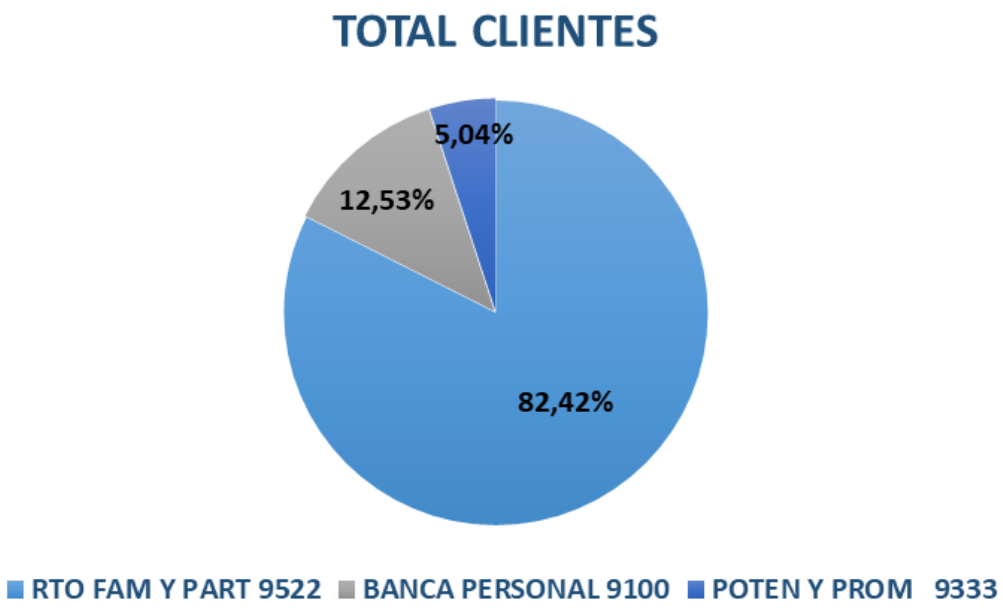


Figura 39: Grupo estratégico clientes entidad

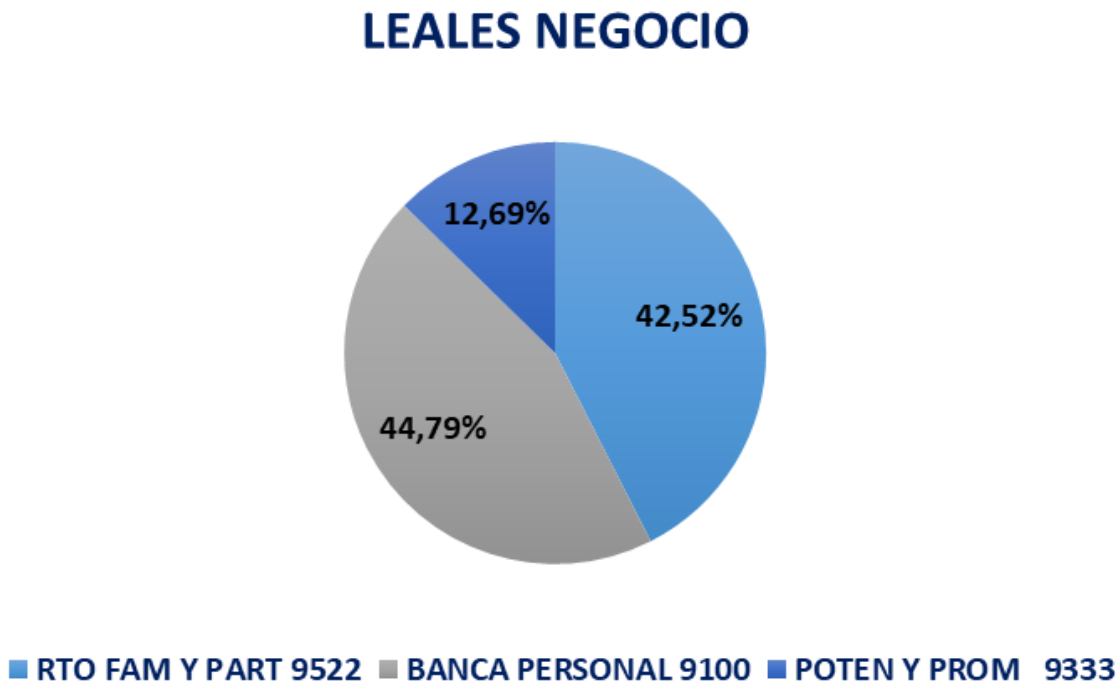


Figura 40: Grupo estratégico clientes leales en negocio

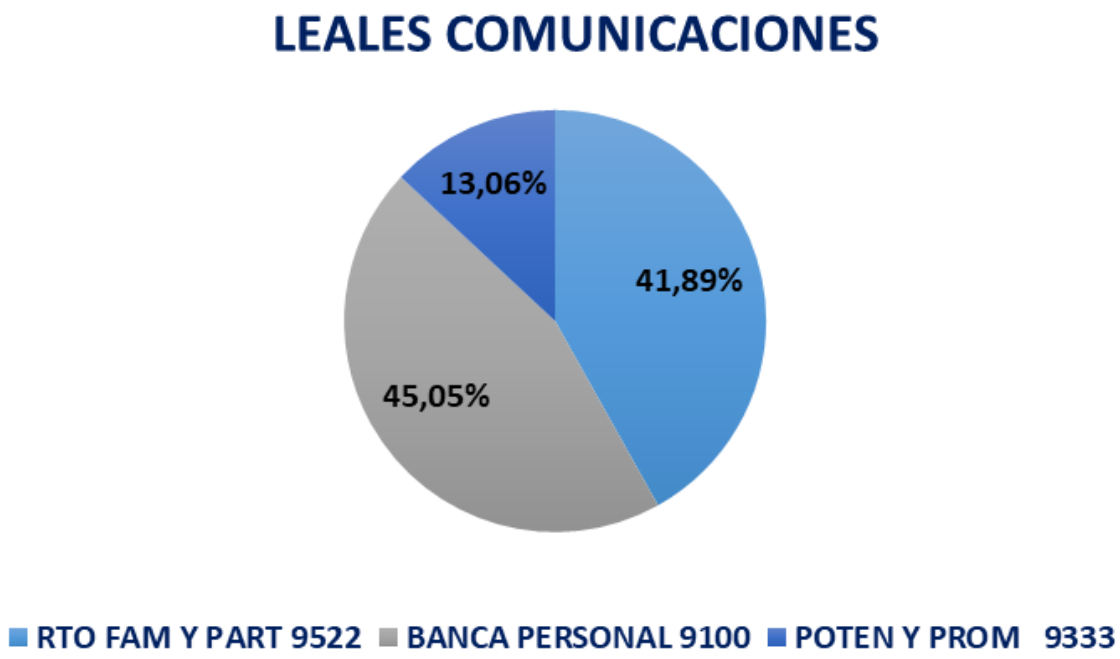


Figura 41: Grupo estratégico clientes leales en comunicaciones

Respecto a la variable grupo estratégico, se observan bastante diferencias entre clientes leales y no leales. El 82 % del total de clientes de la entidad son familias y particulares, mientras que en los clientes leales en el ámbito de negocio, lo que más predomina son los clientes de banca personal (44,7 %) seguido por las familias y particulares. Los clientes leales en comunicaciones, muestran una distribución muy semejante a los de negocio, en primer lugar se sitúan los clientes de banca personal seguido de las familias y particulares.

- SEGMENTO ESTRATÉGICO

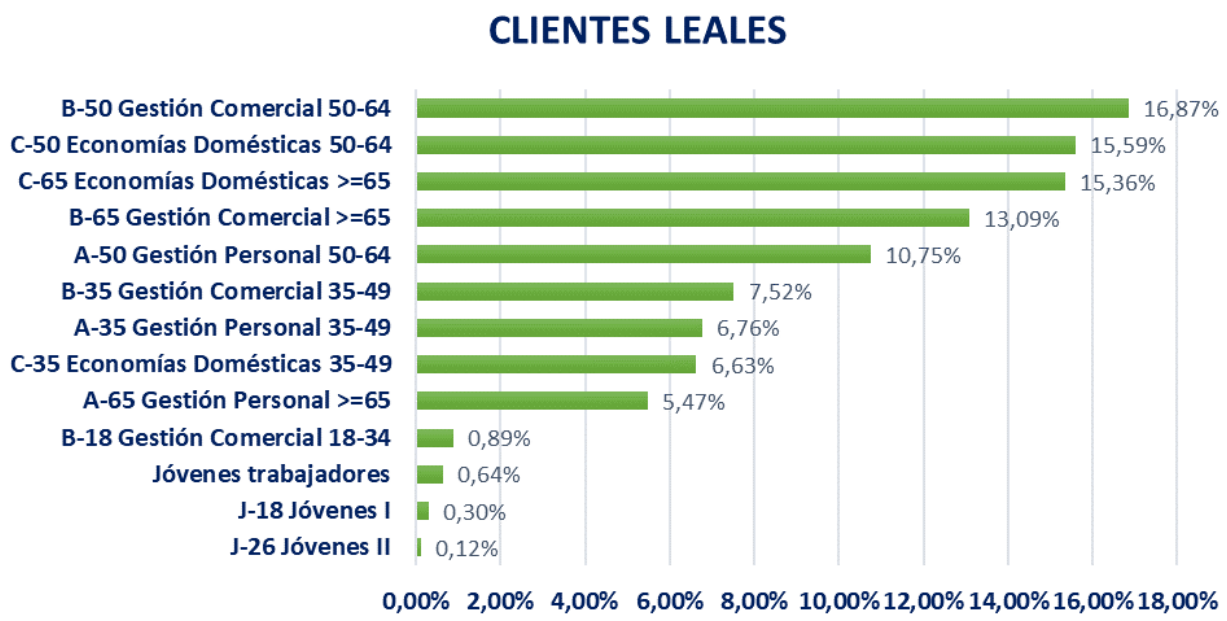


Figura 42: Segmento estratégico clientes leales

La variable segmento estratégico, que muestra los distintos grupos de clientes que establece la entidad, muestra como dentro de los clientes leales, el primer segmento más numeroso es B-Gestión Comercial 50-64 seguido por economías domésticas. En general, los grupos más numerosos son los de clientes mayores de 50 años, mientras que donde se observan menos clientes leales es en los segmentos de clientes jóvenes.

6. Conclusiones

Tras la realización de este Trabajo Fin de Master, se muestran en este apartado las conclusiones que se han obtenido en base al grado de cumplimiento de los objetivos marcados al comienzo del proyecto. Los objetivos que se han conseguido alcanzar con el desarrollo del TFM son:

- El análisis de la estructura de la base de datos general de la entidad financiera así como la construcción de una base de datos propia para la realización del análisis a partir de fuentes de datos heterogéneas.

- La construcción de un modelo de lealtad general que permite clasificar a los clientes de la entidad en base a su nivel de fidelidad con esta.

- El análisis de los clientes leales que tiene la entidad, en lo que a factores demográficos se refiere según los ámbitos que se establecen y en los que se clasifican a estos: ámbito de negocio, ámbito de comunicaciones y ámbito general de lealtad.

- La utilización de distintos programas para la realización del estudio. En primer lugar la herramienta de software SQL server 2016 13.0, como sistema de manejo de bases de datos y en segundo lugar, la herramienta de software libre R-3.4.2., utilizada para el análisis estadístico y la estimación de los modelos. Este último programa se introduce como novedad en la entidad y ofrece muchas más opciones que la herramienta utilizada hasta entonces en el ámbito de modelización, conocida como SPSS.

  El trabajo fin de master realizado, permite la creación de un modelo de lealtad del cliente con alta proyección del futuro. El modelo se plantea desde el enfoque de la lealtad medida desde distintos ámbitos y que por tanto, permite su ampliación a tantos ámbitos como se quieran considerar en el futuro.

  Como reflexión personal, decir que este trabajo ha supuesto para mi un proceso de aprendizaje continuo por varios motivos:

  En primer lugar porque me ha permitido adquirir conocimiento en lo que a competencias informáticas y de bases de datos se refiere, trabajando con uno de los programas de bases de datos más utilizados en las empresas actuales.

  En segundo lugar, la utilización del programa R durante estos meses me ha permitido comprender mejor la estadística y descubrir las múltiples posibilidades que esta herramienta ofrece, ya no solo como herramienta estadística sino como herramienta de programación.

  Además decir que el trabajar con estas herramientas ha despertado en mi gran interés por continuar mi formación en esta línea del análisis de datos unido a la ciencia económica.

  Por último me gustaría agradecer a mis directores del trabajo fin de master su apoyo. Tanto a mi directora por parte de la universidad Carmen Galé, que me ha ofrecido soluciones y no ha dudado en atenderme en tutorías siempre que he solicitado su ayuda como a Eduardo Velasco, director por parte de la empresa que me ha ayudado en todo lo que estaba a su alcance.

7. Bibliografía

Aluja,T. *Curso sobre árboles de decisión*. Universitat Politècnica de Catalunya.

Araújo, P.M., (2015). *Influencia de la experiencia de marca, en el valor de la marca, por via de la satisfacción y lealtad de clientes*

Baptista, M. V., & de Fátima León, M. (2013). *Estrategias de lealtad de clientes en la banca universal*. Estudios Gerenciales, 29(127), 189-203. Science direct.

Buckinx, W., Verstraeten, G., & Van den Poel, D. (2007). *Predicting customer loyalty using the internal transactional database*. Expert systems with applications, 32(1), 125-134.

Chin-Hsue, C., You-Shyang Chen (2009). *Classifying the segmentation of costumer value via RFM model and RS theory*.Expert Systems with Applications, Volume 36, Issue 3, Part 1, April 2009, Pages 4176-4184

Coussement, K., Van den Bossche, F. A., & De Bock, K. W. (2014). *Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees*. Journal of Business Research, 67(1), 2751-2758.

Delgado, M. (2004). *Estado actual de la investigación sobre Lealtad de Marca: Una revisión teórica*. Revista de Dirección, Organización y Organizacion, N0 30. Universidad de Murcia.

Dick, A. S.,& Basu, K. (1994). *Customer loyalty: Toward an integrated conceptual framework.* Journal of the academy of marketing science, 22(2), 99-113.

Documentacion de SQL Server. Microsoft Docs

Dursun, A.,& Caber, M. (2016). *Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis.* Tourism management perspectives, 18, 153-160.

Faraway, Julian J. (2016)*Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models* CRC press.

Goldberg, S.M. (1982). *An Empirical Study of Lifestyle Correlates to Brand Loyal Behavior. Advances in Consumer Research* vol. 9, n.º 1 pp. 456-460

Hastie,T& Tibshirani, R & Friedman, J.(2009) *The Elements of Statistical Learning.* New York, NY, USA; Springer series in statistics.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Jacoby, J., & Chestnut, R. W. (1978). *Brand loyalty: Measurement and management*. John Wiley & Sons Incorporated.

Nelder J. A & Wedderburn R.W. (1972)*General linear models.* Journal of the Royal Statistical Society Series A, 135.

Kristof Coussement, Filip A.M. Van den Bossche, Koen W. De Bock (2014)*Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees.*Journal of Business Research,Volume 67, Issue 1,Pages 2751-2758,ISSN 0148-2963,

Manotas, E. et al. (2014) *Regresión Logística Ordinal Aplicada a la Identificación de Factores de Riesgo para Cáncer de Cuello Uterino* Ingeniare, no 17, p. 87-105.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006).*Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.

Moliner, M.A., Callarisa L.J. *La explicación del comportamiento de lealtad desde la teoría de la actitud: una aplicación a usuarios de hospitales.*Dialnet.

Oliver, R. L. (1999). *Whence consumer loyalty?*. The Journal of Marketing, 33-44.

Park, E., & Bai, B. (2014). *The relationship between brand loyalty and financial performance: an empirical study on the hotel industry.* Journal of Business Research, 8(5), 29-36.

Safary, M. & Safary, Z. (2011). *An Empirical Analysis to design enhanced customer lifetime value based on customer loyalty: evidences from iranian banking sector*. Iranian Journal of Management Studies. Vol 5., No 2

Timofeev, R. (2004). *Classification and Regression Trees (CART).Theory and Applications*, Thesis, Center of Applied Statistics and Economics, Humboldt University, Berlin.

Wei, Jo-Ting & Lin, Shih-Yen & Wu, Hsin-Hung. (2010). *A review of the application of RFM model.* African Journal of Business Management December Special Review. 4. 4199-4206.

8. ANEXOS

SCRIPT DE SQL SERVER
Modelo negocio script

```
SCRIPT DE MODELO DE NEGOCIO
USE DW001

-- CREACIÓN DE LA TABLA DE CLIENTES
CREATE TABLE [RE001].[dbo].[TMP_LEALTAD_NEGOCIO](
[NICI] [int] NULL,
[ANTIG_COMER] [int] NULL,
[TRASPASOS] [int] NULL,
[RECIBOS_MAS_18] [int] NULL,
[VALORACION_PUNT] [decimal](10, 2) NULL,
[SATIS_DESLEALES] [int] NULL,
[SATIS_LEALES] [int] NULL,
[LEALES_INFERIDOS_MAS_18] [int] NULL,
[DESLEALES_INFERIDOS_MAS_18] [int] NULL,
[LEALTAD_MAS_18] [nvarchar](100) NULL,
[RECEN_ULTIMA_OPER] [datetime] NULL
[FREC_OPERACIONES_3_MESES] [int] NULL,
[FREC_OPERACIONES_12_MESES] [int] NULL,
[PERM_MESES_ALTA] [int] NULL,
[PERM_ANYOS_ALTA] [int] NULL,
[INVOL_PRODUCTOS_DIFERENTES] [int] NULL,
) ON [PRIMARY]


-- BASE DE CLIENTES (TOTAL BDD)
/*SELECCIÓN DE CLIENTES:
1° SEA PRIMER TITULAR --> B.TIREL='T' AND B.NUREL IN (1)
2° ALGÚN CONTRATO VIGENTE --> c.SITUA =''
3° PERSONAS FÍSICAS: NATUR = 'H','M'
4° EDAD COHERENTE: EDAD ENTRE 18 y 90 años
5° NO FALLECIDOS, NI INCAPACITADOS:
EXCLUSELEC NOT IN ('F','O') F = FALLECIDO, O=INCAPACITADO
*/
INSERT INTO  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO] (NICI)
SELECT DISTINCT A.NICI
FROM  [dbo].[T1_AGR_CLIENTE_SELEC_ACT]   T1
INNER JOIN  DW001.dbo.CLIENTES_v_act A  WITH(NOLOCK) ON
(T1.NICI = A.NICI)
INNER JOIN RELACION B WITH(NOLOCK)  ON (A.NICI = B.NICI)
INNER JOIN CONTRATOS_ACT C  WITH(NOLOCK)
ON B.CLACA=C.CLACA AND B.CUENTA=C.CUENTA
WHERE B.TIREL='T' AND B.NUREL IN (1) -- ACOTAMOS A TITULAR1
AND C.SITUA=''
AND A.NATUR in ('H','M')
```

```
AND A.EDAD BETWEEN 18 and 90
AND T1.EXCLUSELEC NOT IN ('F','O')

-- ANTIGUEDAD COMERCIAL:
UPDATE [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET ANTIG_COMER = 0

UPDATE T1
SET ANTIG_COMER = 1
FROM [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]  T1
INNER JOIN DW001..CLIENTES_ANTIG_ACT T2 ON (T1.NICI = T2.NICI)
WHERE T2.ORDEN<>1

-- TRASPASOS
UPDATE [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET TRASPASOS =0

UPDATE T1
SET T1.TRASPASOS=1
FROM [RE001].[dbo].[TMP_LEALTAD_NEGOCIO] T1
INNER JOIN DW001..CLIENTE_FLUJOS_ENTID T2 ON (T1.NICI = T2.NICI)
WHERE T2.ENTORIGE='2085' AND T2.ENTDESTI<>'2085'  and
      T2.traspaso = 'S'

--3° DESLEALES -
--SELECT TOP 10  T1.NICI,T1.EDAD,T3.DESCCPTOAH,*
--FROM CLIENTES_V_ACT T1
--INNER JOIN RELACION T4 ON (T1.NICI = T4.NICI)
--INNER JOIN MOV_AHORRO T2 ON (T2.CUENTA = T4.CUENTA)
--INNER JOIN CAT_CONCEP_AH T3 ON (T2.CONCEPAH = T3.CONCEPAH)
--WHERE T1.EDAD>=30
--  AND T3.DESCCPTOAH IN ('AGUA','LUZ','TELEFONO','GAS','ENERGIA')


UPDATE  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET RECIBOS_MAS_18 = 1

UPDATE T5
SET RECIBOS_MAS_18 = 0
FROM  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]  T5
INNER JOIN CLIENTES_V_ACT T1 ON (T5.NICI = T1.NICI)
INNER JOIN RELACION T2 ON (T1.NICI = T2.NICI)
INNER JOIN MOV_AHORRO T3 ON (T3.CUENTA = T2.CUENTA)
INNER JOIN CAT_CONCEP_AH T4 ON (T4.CONCEPAH = T3.CONCEPAH)
WHERE T1.EDAD>=18
  AND T4.DESCCPTOAH IN ('AGUA','LUZ','TELEFONO','GAS','ENERGIA')
```

```
-- 4° SATISFACCIÓN DEL CLIENTE (CREACÓN DE LAS VARIABLES)
--SELECT COUNT (*), SATISFACCION_CONCEPTO
--FROM DW001..SATISFACCION_CLIENTE_FCLIENTE
--WHERE NOMBRE_CONCEPTO = 'RECOMENDACION' AND
COD_CAMPA IN ('ECAL2017','ECAL2016','ECALBP17','ECALBP16')
--GROUP BY SATISFACCION_CONCEPTO
--ORDER BY SATISFACCION_CONCEPTO

-- -- SATISFACCIÓN VALORACIÓN (CÁLCULO COMO
LA MEDIA DE VALORACIONES DEL ÚLTIMO AÑO POR CLIENTE)
UPDATE T1
SET T1.VALORACION_PUNT = T2.PUNTUACION_MEDIA
FROM [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]  T1
INNER JOIN (
SELECT T2.NICI,AVG(CONVERT(DECIMAL(10,2),
T2.SATISFACCION_CONCEPTO)) PUNTUACION_MEDIA
FROM [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]  T1
INNER JOIN  DW001..SATISFACCION_CLIENTE_FCLIENTE T2 ON
(T1.NICI = T2.NICI)
WHERE NOMBRE_CONCEPTO = 'RECOMENDACION' AND
COD_CAMPA IN ('ECAL2017','ECAL2016','ECALBP17','ECALBP16')
 AND SATISFACCION_CONCEPTO
IN ('0','1','2','3','4','5','6','7','8','9','10')
GROUP BY T2.NICI
) T2 ON (T1.NICI = T2.NICI)


-- SATISFACCION_DESLEALES
UPDATE  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET SATIS_DESLEALES = 0

UPDATE T1
SET SATIS_DESLEALES = 1
FROM  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]  T1
WHERE VALORACION_PUNT  IN ('0','1','2','3','4','5','6','7','8')

-- SATISFACCION LEALES
UPDATE [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET SATIS_LEALES = 0

UPDATE T1
SET SATIS_LEALES = 1
FROM [RE001].[dbo].[TMP_LEALTAD_NEGOCIO] T1
WHERE VALORACION_PUNT  IN ('9','10')

-- LEALES Inferidos

UPDATE  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
```

```
SET LEALES_INFERIDOS_MAS_18 = 0

UPDATE  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET LEALES_INFERIDOS_MAS_18 = 1
WHERE SATIS_LEALES =1 AND ANTIG_COMER = 0
AND TRASPASOS = 0 AND RECIBOS_MAS_18 = 0


-- DESLEALES Inferidos

UPDATE [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET DESLEALES_INFERIDOS_MAS_18 = 0

UPDATE [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET DESLEALES_INFERIDOS_MAS_18 = 1
WHERE ANTIG_COMER = 1 OR TRASPASOS = 1
OR RECIBOS_MAS_18 = 1 OR SATIS_DESLEALES = 1



-- LEALTAD (ANALIZO LEALES Vs. DESLEALES)


update [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
set LEALTAD_MAS_18 = NULL

update [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
set LEALTAD_MAS_18 = 'LEAL'
where   ANTIG_COMER = 0 AND TRASPASOS = 0
AND RECIBOS_MAS_18 = 0 and valoracion_punt>=9

update [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
set LEALTAD_MAS_18 = 'NO_LEAL'
where  ANTIG_COMER = 1 OR TRASPASOS = 1
OR RECIBOS_MAS_18 = 1  or valoracion_punt<9


-- VARIABLES DE CLASIFICACIÓN
UPDATE T1
SET T1.RECEN_ULTIMA_OPER = T2.FECHA
FROM  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO] T1
INNER JOIN (
           SELECT T3.NICI,MAX(T3.FECHAINF) +'01' AS FECHA
FROM OPERACIONES T3
INNER JOIN  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO] T4 ON (T3.NICI = T4.NICI)
WHERE T3.CANAL NOT IN ('02','04') AND
T3.CODOPE NOT IN  ('ACTITARJ','ALTACTAR','CONEXION','DESCONEX','SOLICITS')
GROUP BY T3.NICI
```

```
            ) T2 ON (T1.NICI = T2.NICI)


UPDATE  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET RECEN_MESES_ULTIMA_OPER=
datediff(mm,RECEN_ULTIMA_OPER,'20180501')

-- RECENCIA (PORCENTAJE REPRESENTA ÚLTIMA
 VEZ QUE OPERÓ EL CLIENTE):RECEN_PORC_ULTIMA_OPER
UPDATE  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET RECEN_PORC_ULTIMA_OPER = 0

UPDATE  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET RECEN_PORC_ULTIMA_OPER = (1-(CONVERT
(FLOAT,RECEN_MESES_ULTIMA_OPER)-1)/25)*100



--  FRECUENCIA


-- FRECUENCIA: TOTAL DE OPERACIONES en los últimos 3 meses
UPDATE T1
SET T1.FREC_OPERACIONES_3_MESES = 0
from [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]  t1

UPDATE T1
SET T1.FREC_OPERACIONES_3_MESES = T2.OPERACIONES
FROM [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]  T1
INNER JOIN  (
           SELECT T3.NICI,COUNT(*) OPERACIONES
FROM OPERACIONES T3
INNER JOIN [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
 T4 ON (T3.NICI = T4.NICI)
WHERE T3.CANAL NOT IN ('02','04') AND
           T3.CODOPE NOT IN
('ACTITARJ','ALTACTAR','CONEXION','DESCONEX','SOLICITS')
 AND T3.FECHAINF IN ('201802','201803','201804')
      GROUP BY T3.NICI
           ) T2 ON (T1.NICI = T2.NICI)

-- FRECUENCIA DE OPERACIONES ÚLTIMOS 12 MESES
UPDATE T1
SET T1.FREC_OPERACIONES_12_MESES = 0
from [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]  t1
```

```
UPDATE T1
SET T1.FREC_OPERACIONES_12_MESES = T2.OPERACIONES
FROM [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]  T1
INNER JOIN  (
           SELECT T3.NICI,COUNT(*) OPERACIONES
           FROM OPERACIONES T3
INNER JOIN [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
 T4 ON (T3.NICI = T4.NICI)
WHERE T3.CANAL NOT IN ('02','04') AND
 T3.CODOPE NOT IN
('ACTITARJ','ALTACTAR','CONEXION','DESCONEX','SOLICITS')
  AND T3.FECHAINF IN ('201705','201706','201707','201708','201709'
,'201710','201711','201712','201801','201802','201803','201804')
           GROUP BY T3.NICI
           ) T2 ON (T1.NICI = T2.NICI)


--  PERMANENCIA

UPDATE T1
SET T1.PERM_FECHA_ALTA  = T2.FELICA
FROM  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO] T1
INNER JOIN  DW001..CLIENTES_V_ACT  T2 ON (T1.NICI = T2.NICI)

-- CÁLCULO MESES DESDE LA FECHA DE ALTA HASTA AHORA
UPDATE T1
SET T1.PERM_MESES_ALTA = datediff(MM,T1.PERM_FECHA_ALTA,'20180501')
FROM  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO] T1

-- CÁLCULO DE AÑOS DESDE LA FECHA DE ALTA HASTA AHORA
UPDATE  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET PERM_ANYOS_ALTA = PERM_MESES_ALTA/12


-- INVOLUCRACIÓN:
-- TIREL:
--T TITULAR
--PRESTATARIO
--BENEFICIARIO DE CONFIRMING
--CEDENTE DE CARTERA
--ORDENANTE DE CONFIRMING
--A AVALADO (TITULAR DE LOS AVALES)
--APORTANTE EN PLANES DE PENSIONES.
--B BENEFICIARIO DE SEGUROS
--EL DE CONFIRMING ES TITULAR DE SU CTA.
--BENEFICIARIO DE UN AVAL
--D DISPONENTE
--F FIADOR
```

```
--R REPRESENTANTE
--U USUFRUCTUARIO
--G GARANTIZADO (GARANTE)
--M TOMADOR DE SEGUROS



-- INVOLUCRACIÓN: CÁLCULO DISTINTOS PRODUCTOS DEL CLIENTE

UPDATE  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
SET INVOL_PRODUCTOS_DIFERENTES =0

UPDATE T1
SET T1.INVOL_PRODUCTOS_DIFERENTES = T2.PRODUCTOS_DIFERENTES
FROM  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]  T1
INNER JOIN (
select T1.NICI,COUNT(distinct T3.FAMPRO) AS PRODUCTOS_DIFERENTES
from  [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]  T1
INNER JOIN relacion t2 ON(T1.NICI = T2.NICI)
INNER join contratos t3 on (t2.cuenta = t3.cuenta)
INNER join cat_familias t4 on (t3.fampro = t4.FAMPRO)
group by T1.NICI
           ) AS T2 ON T1.NICI = T2.NICI


--------------------------------------------------------------------------------
```

Modelo comunicaciones script

```
use DW001
-- COMUNICACIONES
 drop table #aux
-- SE CREA LA TABLA
CREATE TABLE #AUX (
[NICI] [int] NULL,
[ANTIG_COMER] [int] NULL,
[TRASPASOS] [int] NULL,
[RECIBOS_MAS_18] [int] NULL,
[VALORACION_PUNT] [decimal](10, 2) NULL,
[SATIS_DESLEALES] [int] NULL,
[SATIS_LEALES] [int] NULL,
[LEALES_INFERIDOS_MAS_30] [int] NULL,
[LEALES_INFERIDOS_MAS_18] [int] NULL,
[DESLEALES_INFERIDOS_MAS_30] [int] NULL,
[DESLEALES_INFERIDOS_MAS_18] [int] NULL,
[RECEN_ULTIMA_COMUNICACION] [INT] NULL,
[RECENCIA] [int] NULL, -- MESES ULTIMA COMUNICACIÓN
```

```
[FRECUENCIA][FLOAT] NULL, -- COMUNICACIONES QUE RECEPCION
[PERMANENCIA][INT] NULL, -- TIEMPO QUE LLEVA RECIBIENDO COMUNICACIONES Y RE
[LORTAD][INT]NULL,
[INVOLUCRACION][FLOAT] NULL, -- COMUNICACIONES EN LAS QUE HA EXISTIDO RESPU
)


-- NOS BASAMOS EN LA TABLA CONSTRUIDA PARA NEGOCIO
INSERT INTO #AUX (
NICI ,
ANTIG_COMER ,
TRASPASOS ,
RECIBOS_MAS_18 ,
VALORACION_PUNT ,
SATIS_DESLEALES ,
SATIS_LEALES,
LEALES_INFERIDOS_MAS_18,
DESLEALES_INFERIDOS_MAS_18,
)
SELECT
[NICI],
[ANTIG_COMER],
[TRASPASOS] ,
[RECIBOS_MAS_18] ,
[VALORACION_PUNT],
[SATIS_DESLEALES],
[SATIS_LEALES],
[LEALES_INFERIDOS_MAS_18],
[DESLEALES_INFERIDOS_MAS_18],
[LEALTAD_MAS_18]
FROM [RE001].[dbo].[TMP_LEALTAD_NEGOCIO]
--WHERE VALORACION_PUNT >0

-- CONSULTAS BASE:
-------------------------------------------------------------------------
---- ENVIADO
--select * from  [RE001].[DBO].[TMP_LEALTAD_TABLA_COMUNICACIONES_TOTAL]
--where DISTRIBUCION = 1 AND DESC_CANAL <> 'AVISOS'


---- RECEPCIONADO
--select * from [RE001].[DBO].[TMP_LEALTAD_TABLA_COMUNICACIONES_total]
--where DISTRIBUCION = 1 AND IMPACTO = 1 AND DESC_CANAL <> 'AVISOS'


---- ACCESOS O COMUNICACIONES CON ÉXITO
--select DESC_CANAL,DESCRESUL,ALFABLOQUE,COUNT(*)
--from [RE001].[DBO].[TMP_LEALTAD_TABLA_COMUNICACIONES_total]
```

```
--where DISTRIBUCION = 1 AND IMPACTO = 1
--  AND DESC_CANAL <> 'AVISOS'
--  and ALFABLOQUE LIKE 'éXITO%' OR DESCRESUL in ('dudoso','interesado')
--GROUP BY DESC_CANAL,DESCRESUL,ALFABLOQUE
--ORDER BY ALFABLOQUE

-------------------------------------------------------------------------
-- RECEN_ULTIMA_COMUNICACION (MES ÚLTIMA COMUNICACIÓN)
UPDATE T1
SET RECEN_ULTIMA_COMUNICACION = NULL
FROM #AUX T1

UPDATE T1
SET T1.RECEN_ULTIMA_COMUNICACION = T2.MESES
FROM #AUX T1
INNER JOIN (
           select T1.NICI,MAX(T1.ANIO_MES)AS MESES
FROM  [RE001].[DBO].[TMP_LEALTAD_TABLA_COMUNICACIONES_total] T1
WHERE DISTRIBUCION = 1 AND IMPACTO = 1 AND DESC_CANAL <> 'AVISOS'
GROUP BY T1.NICI
           )  T2
ON (T1.NICI = T2.NICI)

-- RECENCIA: (MESES QUE HAN PASADO DESDE LA ÚLTIMA COMUNICACIÓN) (SE TOMA C
REFERENCIA EL 201802 - últimos datos registrados)
UPDATE T1
SET RECENCIA = NULL
FROM #AUX T1

UPDATE T1
SET RECENCIA = DATEDIFF(mm,CONVERT(NVARCHAR(6),RECEN_ULTIMA_COMUNICACION)+'
FROM #AUX T1



-- FRECUENCIA: (% Comunicaciones
que se recepcionan sobre el total de comunicaciones realizadas)

drop table #ENVIADOS


UPDATE #AUX
SET FRECUENCIA = NULL

             -- ENVIOS
SELECT T1.NICI,COUNT(*) AS COMUNICACIONES_ENVIADAS
INTO #ENVIADOS
FROM #AUX T1
```

```
INNER JOIN  [RE001].[DBO].[TMP_LEALTAD_TABLA_COMUNICACIONES_TOTAL]
T2 ON (T1.NICI = T2.NICI)
WHERE DISTRIBUCION = 1 AND DESC_CANAL <> 'AVISOS'
GROUP BY T1.NICI

DROP TABLE #RECEPCIONES
             -- RECEPCIONES
SELECT T1.NICI,COUNT(*) AS COMUNICACIONES_RECEPCIONADAS
INTO #RECEPCIONES
FROM #AUX T1
INNER JOIN  [RE001].[DBO].[TMP_LEALTAD_TABLA_COMUNICACIONES_TOTAL]
 T2 ON (T1.NICI = T2.NICI)
WHERE DISTRIBUCION = 1 AND IMPACTO = 1 AND DESC_CANAL <> 'AVISOS'
GROUP BY T1.NICI

-- RATIO DE FRECUENCIA (RECEPCIONADAS/ ENVIADAS)
UPDATE T1
SET FRECUENCIA = (CONVERT(FLOAT,COMUNICACIONES_RECEPCIONADAS)/CONVERT
(FLOAT,COMUNICACIONES_ENVIADAS))*100
FROM #AUX T1
INNER JOIN #ENVIADOS T2 ON (T1.NICI = T2.NICI)
INNER JOIN #RECEPCIONES T3 ON (T2.NICI = T3.NICI)

-------------------------------------------------------------------------
-- PERMANENCIA: A MODO ESTIMACIÓN DIFERENCIA EN DÍAS ENTRE
 LA FECHA MINIMA DEL PRIMER ENVIO QUE SE HA REALIZADO AL CLIENTE
 Y LA FECHA MÁXIMA DE RECEPCIÓN
-------------------------------------------------------------------------
DROP TABLE #FECHA_MINIMA_RECEPCION
DROP TABLE #FECHA_MAXIMA_RECEPCION

-- FECHA MINIMA ENVIO
select T1.NICI,MIN(T2.ANIO_MES)AS MESES_MINIMO
INTO #FECHA_MINIMA_RECEPCION
FROM  #AUX T1
INNER JOIN [RE001].[DBO].[TMP_LEALTAD_TABLA_COMUNICACIONES_total]
 T2 ON (T1.NICI = T2.NICI)
WHERE T2.DISTRIBUCION = 1 AND T2.IMPACTO = 1
AND T2.DESC_CANAL <> 'AVISOS' AND DESC_CANAL in ('Correo electrónico','Opo

GROUP BY T1.NICI



-- FECHA MINIMA ENVIO
select T1.NICI,MAX(T2.ANIO_MES)AS MESES_MAXIMO
INTO #FECHA_MAXIMA_RECEPCION
FROM  #AUX T1
```

```
INNER JOIN [RE001].[DBO].[TMP_LEALTAD_TABLA_COMUNICACIONES_total
] T2 ON (T1.NICI = T2.NICI)
WHERE T2.DISTRIBUCION = 1 AND T2.IMPACTO = 1
AND T2.DESC_CANAL <> 'AVISOS' AND DESC_CANAL in
 ('Correo electrónico','Oportunidades de oficina')
GROUP BY T1.NICI

-- CÁLCULO DEL INDICADOR
UPDATE T1
SET T1.PERMANENCIA = DATEDIFF(MM,CONVERT(NVARCHAR(6),MESES_MINIMO)+
'01',CONVERT(NVARCHAR(6),MESES_MAXIMO)+'01')
FROM #AUX T1
INNER JOIN #FECHA_MINIMA_RECEPCION  AS T2 ON (T1.NICI = T2.NICI )
INNER JOIN #FECHA_MAXIMA_RECEPCION AS T3 ON (T2.NICI = T3.NICI)


-- LOPD --> LORTAD3
-------------------------------------------------------------------------
--NULL si Tácito
--0 Si Tácito
--1 Sí Expreso ***
--2 Solo de la Caja
--3 NO
--4 Solo correo: cualquier publicidad
--5 Solo correo: publicidad de la caja

UPDATE #AUX
SET LORTAD = NULL

UPDATE t1
SET T1.LORTAD = T2.LORTAD3
from #AUX t1
inner join CLIENTES T2 ON (t1.NICI = T2.NICI)


-- INVOLUCRACIÓN: (ÉXITO DE LAS COMUNICACIONES DE DISTINTOS CANALES)
-----------------------------------------------------------------------
UPDATE T1
SET INVOLUCRACION = NULL
FROM #AUX T1

-- SE CALCULAN LAS OPERACIONES CON ÉXITO, PARA LAS
RECEPCIONADAS UTILIZAMOS LA TABLA ANTERIOR #RECEPCIONES

DROP TABLE #RECEPCIONES_INVOLUCRACION

            -- RECEPCIONES
SELECT T1.NICI,COUNT(*) AS COMUNICACIONES_RECEPCIONADAS
```

```
INTO #RECEPCIONES_INVOLUCRACION
FROM #AUX T1
INNER JOIN  [RE001].[DBO].[TMP_LEALTAD_TABLA_COMUNICACIONES_TOTAL]
T2 ON (T1.NICI = T2.NICI)
WHERE DISTRIBUCION = 1 AND IMPACTO = 1 AND
DESC_CANAL <> 'AVISOS' AND DESC_CANAL <> '' AND DESC_CANAL in
('Correo electrónico','Oportunidades de oficina')
and ANIO_MES >=201703
GROUP BY T1.NICI



DROP TABLE #COMUNICACIONES_EXITO_INVOLUCRACION

SELECT T1.NICI,COUNT(*) AS COMUNICACIONES_EXITO
INTO #COMUNICACIONES_EXITO_INVOLUCRACION
FROM #AUX T1
INNER JOIN [RE001].[DBO].[TMP_LEALTAD_TABLA_COMUNICACIONES_total]
T2 ON (T1.NICI = T2.NICI)
where T2.DISTRIBUCION = 1
  AND T2.IMPACTO = 1
  AND T2.DESC_CANAL <> 'AVISOS' AND DESC_CANAL <>
 '' AND DESC_CANAL in('Correo electrónico','Oportunidades de oficina')
  AND (T2.ALFABLOQUE LIKE 'éXITO%' OR T2.DESCRESUL in ('dudoso','interesado
  and ANIO_MES >=201703
GROUP BY T1.NICI

-- ACTUALIZANDO CAMPO
UPDATE T1
SET INVOLUCRACION = (CONVERT(FLOAT,COMUNICACIONES_EXITO)
/CONVERT(FLOAT,  COMUNICACIONES_RECEPCIONADAS))*100
FROM #AUX T1
inner JOIN  #RECEPCIONES_INVOLUCRACION AS T2 ON (T1.NICI = T2.NICI)
LEFT JOIN  #COMUNICACIONES_EXITO_INVOLUCRACION  AS T3 ON (T2.NICI = T3.NICI
```

SCRIPT DE R

```
library(readxl)


MUESTRA1 <- read_excel("MUESTRA1.xlsx",
   sheet = "MUESTRA", na = "NULL")


##MODELO LEALTAD ÁMBITO NEGOCIO

MUESTRA1$LEALTAD<-as.factor(MUESTRA1$LEALTAD)
```

```
MUESTRA1$RECENCIA_N<-as.numeric(MUESTRA1$RECENCIA_N)

MUESTRA1$FRECUENCIA_N<-as.numeric(MUESTRA1$FRECUENCIA_N)

MUESTRA1$PERMANENCIA_N<-as.numeric(MUESTRA1$PERMANENCIA_N)

MUESTRA1$INVOLUCRACION_N<-as.numeric(MUESTRA1$INVOLUCRACION_N)

library(ggplot2)
library(lattice)
library(caret)
library(AppliedPredictiveModeling)

set.seed(2018)

trainIndex <- createDataPartition(MUESTRA1$LEALTAD, p = .8,
                                  list = FALSE,
                                  times = 1)
datosTrain <- MUESTRA1[ trainIndex,]
datosTest  <- MUESTRA1[-trainIndex,]

modelo1<-glm(LEALTAD~RECENCIA_N+FRECUENCIA_N+PERMANENCIA_N+INVOLUCRACION_N,
summary(modelo1)


library(ResourceSelection)


step(modelo1prueba,scope=list(upper=LEALTAD_TOTAL),direction="forward",data


LEALTAD0NEGOCIO<-MUESTRA1[MUESTRA1$LEALTAD == "0",]

LEALTAD1NEGOCIO<-MUESTRA1[MUESTRA1$LEALTAD == "1",]

summary(LEALTAD1NEGOCIO)

summary(LEALTAD0COM)

sd(MUESTRA1$RECENCIA_N)


#gráficos de caja en el ámbito de negocio

par(mfrow=c(1,2))

boxplot(RECENCIA_N~LEALTAD , data=MUESTRA1, id.method="y", xlab="Lealtad",
```

```
View(MUESTRA1$RECENCIA_N)

hist(MUESTRA1$RECENCIA_N,freq = FALSE, col="lightcyan", main=" ", xlab="",

boxplot(FRECUENCIA_N~LEALTAD, data=MUESTRA1, id.method="y", xlab="Lealtad",

boxplot(PERMANENCIA_N~LEALTAD, data=MUESTRA1, id.method="y", xlab="Lealtad

boxplot(INVOLUCRACION_N~LEALTAD, data=MUESTRA1, id.method="y", xlab="Lealta

#Test de máxima verosimilitud

library(zoo)
library(lmtest)

lrtest(modelo1)

#H0:el modelo es adecuado sólo con la constante
#Ha:el modelo no es adecuado sólo con la constante
#como se obtiene que el pvalor es <0.05, se rechaza la hipótesis nula,
el modelo mejora con las variables introducidas

prob_exito<- predict(modelo1,newdata=datosTest,type="response")

prediccionRL<- ifelse(prob_exito>=0.5, 1, 0)


mc1 <- table(datosTest$LEALTAD,prediccionRL, dnn = c("Real","Predicción"))
mc1
library(lattice)
library(ggplot2)
library(caret)

confusionMatrix(datosTest$LEALTAD,prediccionRL)

aciertos <-sum(diag(mc1)) / sum(mc1) * 100

summary(mc1)


library(ROCR)

pred1 <- prediction(prob_exito,datosTest$LEALTAD)

roc.perf1 = performance(pred1, measure = "tpr", x.measure = "fpr")
#tpr: True positive rate: verdaderos clasificados como verdaderos
#fpr: False positive rate: negativos clasificados como positivos
```

```
str(roc.perf3)
plot(roc.perf3)
abline(a=0, b= 1)
roc.perf3
summary(roc.perf3)


#Árbol de decisión modelo negocio
MUESTRAARBOL<-MUESTRA1[,c(10,13,16,23,24)]

library(MASS)
library(rpart)

set.seed(1684)

trainIndex1 <- createDataPartition(MUESTRAARBOL$LEALTAD, p = .8,
                                   list = FALSE,
                                   times = 1)
datosTrainarbol1 <- MUESTRAARBOL[ trainIndex1,]
datosTestarbol1 <- MUESTRAARBOL[-trainIndex1,]

library(rpart)

modeltree_N<-rpart( LEALTAD~. ,data=datosTrainarbol1, method="class")
modeltree_N
summary(modeltree_N)
plot(modeltree_N)

prob_exito<- predict(modeltree_N,newdata=datosTestarbol1,type="class")


TAB <- table(datosTestarbol1$LEALTAD, prob_exito)

mcrtree <- 1 - sum(diag(TAB))/sum(TAB)
mcrtree
library(rpart)
library(rattle)
library(rpart.plot)
rpart.plot(modeltree_N)
modeltree_N

aciertos <-sum(diag(TAB)) / sum(TAB) * 100

##Modelo comunicaciones##
library(readxl)
MUESTRA2 <- read_excel("F:/MUESTRA2.xlsx",
  sheet = "Hoja1", na = "NULL")
```

```
set.seed(2018)

library(lattice)
library(ggplot2)
library(caret)

trainIndex2 <- createDataPartition(MUESTRA2$LEALTAD_C, p = .8,
                                   list = FALSE,
                                   times = 1)
datosTrain2 <- MUESTRA2[ trainIndex,]
datosTest2  <- MUESTRA2[-trainIndex,]


datosTrain2$LEALTAD_C<-as.factor(datosTrain2$LEALTAD_C)

datosTrain2$RECENCIA_C<-as.numeric(datosTrain2$RECENCIA_C)

datosTrain2$FRECUENCIA_C<-as.numeric(datosTrain2$FRECUENCIA_C)


datosTrain2$PERMANENCIA_C<-as.numeric(datosTrain2$PERMANENCIA_C)

datosTrain2$INVOLUCRACION_C<-as.numeric(datosTrain2$INVOLUCRACION_C)

modelo2<-glm(LEALTAD_C~RECENCIA_C+FRECUENCIA_C+
PERMANENCIA_C+INVOLUCRACION_C, data=datosTrain2, family = "binomial")

MUESTRA2$LEALTAD_C<-as.factor(MUESTRA2$LEALTAD_C)

MUESTRA2$INVOLUCRACION_C<-as.factor(MUESTRA2$INVOLUCRACION_C)
summary(MUESTRA2)

LEALTAD0COM<-MUESTRA2[MUESTRA2$LEALTAD_C == "0",]

LEALTAD1COM<-MUESTRA2[MUESTRA2$LEALTAD_C == "1",]


#COEFICIENTE DE ASIMETRÍA
library(moments)
skewness(MUESTRA2$INVOLUCRACION_C, na.rm = TRUE)
skewness(LEALTAD1COM$PERMANENCIA_C)
skewness(LEALTAD1COM$INVOLUCRACION_C)

#DESVIACIÓN TÍPICA

sd(LEALTAD1COM$LORTAD__1, na.rm = TRUE)
```

```
#gráficos de caja en comunicaciones
par(mfrow=c(1,2))

boxplot(RECENCIA_C~LEALTAD_C, data=MUESTRA2, id.method="y"
, xlab="Lealtad", ylab="Recencia")

boxplot(FRECUENCIA_C~LEALTAD_C, data=MUESTRA2, id.method="y"
, xlab="Lealtad", ylab="Frecuencia")

boxplot(PERMANENCIA_C~LEALTAD_C, data=MUESTRA2, id.method="y"
, xlab="Lealtad", ylab="Permanencia")

barplot(prop.table(table(MUESTRA2$INVOLUCRACION_C)),col =
c("orange","blue"),ylim=c(0,1), xlim =c(0,3.3))

View(MUESTRA2$INVOLUCRACION_C)
lrtest(modelo2)

View(LEALTAD0COM$LORTAD__1)


library(AppliedPredictiveModeling)
featurePlot(x = MUESTRA2[, c(16,17,18)],
            y = MUESTRA2$LEALTAD_18,
            plot = "box")

prob_exito2<- predict(modelo2,newdata=datosTest2,type="response")
View(prob_exito2)
View(datosTest2)
summary(datosTest2)
prediccion<- ifelse(prob_exito2>=0.5, 1, 0)

mc2 <- table(datosTest2$LEALTAD_18,prediccion, dnn= c ("observaciones","pre

library(lattice)
library(ggplot2)
library(caret)

confusionMatrix(datosTest2$LEALTAD_C,prediccion)

aciertos <- sum(diag(mc2)) / nrow(datosTest2) * 100
aciertos
sensibilidad<-(mc2[4]/sum(mc2[2],mc2[4]))
espe<-(mc2[1]/ sum(mc2[1],mc2[3]))
sensibilidad
espe
1-espe
pred2 <- prediction(prob_exito2,datosTest2$LEALTAD_18)
```

```
roc.perf2 = performance(pred2, measure = "tpr", x.measure = "fpr")
plot(roc.perf2)
#tpr: True positive rate: verdaderos clasificados como verdaderos
#fpr: False positive rate: negativos clasificados como positivos
plot(roc.perf2)
abline(a=0, b= 1)


#likelihood test ratio
library(zoo)
library(lmtest)
lrtest(modelo2)
#H0:el modelo es adecuado sólo con la constante
#Ha:el modelo no es adecuado sólo con la constante
#como se obtiene que el pvalor es <0.05, se rechaza la hipótesis nula, el m


#Árbol de decisión modelo de comunicaciones

MUESTRAARBOLCOM<-MUESTRA2[,c(16,17,18,20,21)]

View(MUESTRAARBOLCOM)


set.seed(417)

trainIndex <- createDataPartition(MUESTRAARBOLCOM$LEALTAD_C, p = .8,
                                  list = FALSE,
                                  times = 1)
datosTrainarbolcom <- MUESTRAARBOLCOM[ trainIndex,]
datosTestarbolcom <- MUESTRAARBOLCOM[-trainIndex,]

modeltree_C<-rpart( LEALTAD_C~. ,data=datosTrainarbolcom, method="class")
head(summary(modeltree),60)
plot(modeltree)
text(modeltree)

predict.tree <- predict(modeltree_C,newdata=datosTestarbolcom,type="class"
TAB <- table(datosTestarbolcom$LEALTAD_C, predict.tree)
TAB
mcrtree <- 1 - sum(diag(TAB))/sum(TAB)
library(rpart)
library(rattle)
library(rpart.plot)
rpart.plot(modeltree_C)
```

```
MUESTRA3 <- read_excel("H:/MUESTRA3.xlsx", na = "NULL")

library(readxl)



MUESTRA3$RECENCIA_C<-as.numeric(MUESTRA3$RECENCIA_C)
MUESTRA3$RECENCIA_N<-as.numeric(MUESTRA3$RECENCIA_N)
MUESTRA3$FRECUENCIA_C<-as.numeric(MUESTRA3$FRECUENCIA_C)
MUESTRA3$FRECUENCIA_N<-as.numeric(MUESTRA3$FRECUENCIA_N)
MUESTRA3$PERMANENCIA_C<-as.numeric(MUESTRA3$PERMANENCIA_C)
MUESTRA3$PERMANENCIA_N<-as.numeric(MUESTRA3$PERMANENCIA_N)
MUESTRA3$INVOLUCRACION_C<-as.numeric(MUESTRA3$INVOLUCRACION_C)
MUESTRA3$INVOLUCRACION_N<-as.numeric(MUESTRA3$INVOLUCRACION_N)
MUESTRA3$LEALTAD<-as.factor(MUESTRA3$LEALTAD)



set.seed(1587)

trainIndex <- createDataPartition(MUESTRA3$LEALTAD, p = .8,
                                  list = FALSE,
                                  times = 1)
datosTrain3 <- MUESTRA3[ trainIndex,]
datosTest3  <- MUESTRA3[-trainIndex,]


LEALTAD_TOTAL<-glm(LEALTAD~RECEN_MESES_ULTIMA_OPER+FREC_OPERACIONES_12_MESI
+PERMANENCIA_NEG+INVOL_PRODUCTOS_DIFERENTES+RECENCIA_COM
+FRECUENCIA_COM+PERMANENCIA_COM+INVOLUCRACION_COM,data=datosTrain3,family=b
summary(LEALTAD_TOTAL)

prob_exito<- predict(LEALTAD_TOTAL,newdata=datosTest3,type="response")
prob_exito
View(prob_exito)
View(datosTest$LEALTAD)
prediccionRL3<- ifelse(prob_exito>=0.5, 1, 0)

mc3 <- table(datosTest3$LEALTAD,prediccionRL3, dnn = c("Real","Predicción")
mc3
aciertos <-sum(diag(mc3)) / sum(mc3) * 100
aciertos

#likelihood test ratio
lrtest(LEALTAD_TOTAL)
```

```
trainIndex <- createDataPartition(MUESTRA3$LEALTAD, p = .8,
                                  list = FALSE,
                                  times = 1)

datosTrain3 <- MUESTRA3[ trainIndex,]
datosTest3 <- MUESTRA3[-trainIndex,]

LEALTAD_TOTAL<-glm(LEALTAD~RECENCIA_N+FRECUENCIA_N+PERMANENCIA_N
+INVOLUCRACION_N+RECENCIA_C+FRECUENCIA_C+PERMANENCIA_C+INVOLUCRACION_C, dat

LEALTAD_TOTAL<-glm(LEALTAD~FRECUENCIA_C, data=datosTrain3, family = binomia


lrtest(LEALTAD_TOTAL)

prob_exito<- predict(LEALTAD_TOTAL,newdata=datosTest3,type="response")
prob_exito
View(prob_exito)
View(datosTest$LEALTAD)
prediccionRL<- ifelse(prob_exito>=0.5, 1, 0)
summary(LEALTAD_TOTAL)

library(lattice)
library(ggplot2)
library(caret)

confusionMatrix(datosTest3$LEALTAD,prediccionRL)


library(ROCR)
pred3 <- prediction(prob_exito,datosTest3$LEALTAD)
?performance
roc.perf3 = performance(pred3, measure = "tpr", x.measure = "fpr")
#tpr: True positive rate: verdaderos clasificados como verdaderos
#fpr: False positive rate: negativos clasificados como positivos
View(roc.perf3)
str(roc.perf3)
plot(roc.perf3)
abline(a=0, b= 1)
roc.perf3
summary(roc.perf3)



#Árbol de decisión modelo de lealtad total

MUESTRAARBOLTOTAL<-MUESTRA3[,c(12,15,18,23,24,27,28,29,30)]
```

```
set.seed(417)

trainIndex <- createDataPartition(MUESTRAARBOLTOTAL$LEALTAD, p = .8,
                                  list = FALSE,
                                  times = 1)
datosTrainarboltotal<- MUESTRAARBOLTOTAL[ trainIndex,]
datosTestarboltotal <- MUESTRAARBOLTOTAL[-trainIndex,]

modeltreetotal<-rpart( LEALTAD~. ,data=datosTrainarboltotal, method="class"
head(summary(modeltree),60)
plot(modeltree)
text(modeltree)

predict.tree <- predict(modeltreetotal,newdata=datosTestarboltotal, type="c
TAB <- table(datosTestarboltotal$LEALTAD, predict.tree)
mcrtree <- 1 - sum(diag(TAB))/sum(TAB)
confusionMatrix(predict.tree, datosTestarboltotal$LEALTAD)

rary(rattle)
library(rpart.plot)
```