Universidad Zaragoza
1542

Trabajo Fin de Grado

El análisis de datos en departamentos de marketing de entidades financieras

Autor

Daniel Aldea González

Directora

Pilar Olave Rubio

Facultad de Economía y Empresa

2017

Autor: Daniel Aldea González

Directora: Pilar Olave Rubio

Título del trabajo: El análisis de datos en departamentos de marketing de entidades financieras

Titulación: Grado en Marketing e Investigación de Mercados

Modalidad: Análisis de datos

RESUMEN

Este proyecto pretende poner en valor la importancia que supone para las empresas analizar toda la información que obtienen a través de diversas fuentes de datos, especialmente en los departamentos de marketing. Se toman decisiones continuamente, y cuanta más información tengamos para ello, tendremos más probabilidades de que nuestra decisión se aproxime más a la óptima. Las empresas, en la toma de sus decisiones, pueden añadir valor a sus productos y servicios mediante el uso de los datos que disponen convirtiéndolos en información útil gracias a los métodos estadísticos. El objetivo último de este trabajo es analizar una base de datos de una entidad financiera, concretamente estimar un modelo de regresión logística binaria que nos ayude a desarrollar una campaña de comunicación comercial mediante la determinación del público objetivo al que dirigirla, en concreto para rentabilizar la inversión de la campaña de marketing en la contratación de un depósito a plazo, detallándose tanto la metodología empleada como las conclusiones obtenidas. Previamente al análisis, se introduce cómo el Big Data se encuentra cada día más presente en la sociedad de la información y comunicación en la que vivimos, y cómo cada día son más aquellos agentes económicos que emplean el análisis masivo de datos para obtener información valiosa en la toma de sus decisiones.

ABSTRACT

The purpose of this project is highlighting the companies' importance analyzing the available information coming from variated sources, putting special relevance in the marketing department. Decisions are taking all time long, the more information we have the higher probability of taking decisions closer to the optimum. The companies along their decision making process can add value to the products and services they offer by the use of available data once it had been treated to become useful information helped by statistical methods. The last aim of the report is analyzing a financial institution data base; more concretely, estimating a binary logistic regression model which may help us identifying the worthiest target market as previous step to develop a publicity campaign that will maximize the profit of the marketing campaign of a time deposit digging into the methodology and procedures required, as well as the conclusions found. Before the analysis, big data concepts will be introduced and how it had become more and more essential in the information and communication society we are currently living, and how day after day is greater the number of economic agents that use massive data analysis to get important information to make their decisions.

ÍNDICE

INDICE DE GRÁFICOS

ÍNDICE DE TABLAS

1. Sociedad de la información

"*Hemos entrado en lo que algunos expertos han calificado como la revolución de la Información y la comprensión de la estadística, ya que uno de los principales obstáculos del progreso ha sido nuestra incapacidad para prever el futuro y tomar decisiones políticas sabias, basadas en una buena información*". [Olave P. (1995)]. Aunque desde aquella publicación en los Cuadernos Aragoneses de Economía han pasado ya más de 20 años, perfectamente podrían formar parte de un artículo de mera actualidad, y es que la imparable revolución tecnológica mundial que estamos viviendo desde la aparición de internet en 1969, y especialmente a partir de mediados de la década de los 90 hasta nuestros días, ha cambiado el modelo de sociedad en la que vivimos, hasta ser denominada "sociedad de la información", que a pesar de ser un término que lleva empleándose varias décadas, no puede decirse que se haya quedado anticuado. Así pues, en este contexto, el análisis de datos toma una relevancia especial.

Podemos entender como tal, una sociedad en la que todas aquellas actividades que realizamos los seres humanos tienen como eje principal la información: ésta es generada por las personas para posteriormente ser difundida y procesada por otros usuarios a los que genera utilidad en diversos ámbitos como el trabajo, el ocio, o simplemente el conocimiento de lo desconocido como vía de superación y evolución personal. Dado el carácter generalista y abstracto del concepto de "*Sociedad de la Información*", así como su uso en campos del conocimiento tan dispares, no existe una definición que sea universalmente aceptada, no obstante, puede entenderse como "*Un estadio de desarrollo social caracterizado por la capacidad de sus miembros (ciudadanos, empresas y sector público) para obtener y compartir cualquier información, instantáneamente, desde cualquier lugar, y en la forma que se prefiera*."

Este nuevo modelo social que rige la forma de relacionarse, trabajar, y en definitiva, de vivir, no puede ser explicado sin comprender el principal pilar que lo sustenta: las tecnologías de la información y la comunicación.

1.1. Las tecnologías de la información y la comunicación

"*Son el conjunto de medios de comunicación y las aplicaciones de información que permiten la captura, producción, almacenamiento, tratamiento, y presentación de informaciones en forma de voz, imágenes y datos contenidos en señales de*

naturaleza acústica, óptica o electromagnética.” Entendidas desde una dimensión social, las TIC facilitan la transmisión de información entre los diferentes agentes de una economía o sociedad, proporcionándoles las herramientas necesarias para la creación, difusión y procesamiento de la información con el fin de conseguir una comunicación más completa, rápida y eficiente.

Fuente: https://tecnologasdelainformacinycomunicacin.wordpress.com/

Ya en el año 1994, el profesor Javier Ordoñez Rodríguez, dijo las siguientes palabras en un acto celebrado para la presentación de la revista Fuentes Estadísticas: “*Las informaciones empíricas proporcionadas por la estadística han contribuido a cambiar esencialmente la percepción de la realidad económica, social, física, y por supuesto, de otro orden material*”. A lo que Fernando de Esteban, por aquel entonces director de difusión de EUROSTAT, apuntó la frase “*cuanta más democracia haya, más información ha de darse*”.

Por lo tanto, la estadística debe ser empleada como herramienta para ordenar y procesar dicha información, otorgándole la forma necesaria para convertirla en útil para la sociedad. Si este procesamiento de los datos se efectúa de la forma adecuada, el resultado será una "fotografía" de la realidad que nos rodea, y que podremos emplear para conocer mejor la situación en la que nos encontramos y tomar mejores decisiones gracias a la disponibilidad de información fiable de gran calidad.

Como consecuencia de ello, la preocupación de los analistas se encuentra en los medios y herramientas que les permitan llevar a cabo un procesamiento de los datos teniendo en

cuenta su carácter masivo y desestructurado, pues desde que el uso de internet se ha convertido en algo generalizado, la información que se genera y almacena está creciendo de forma exponencial, de modo que los métodos estadísticos deben avanzar procesarla.

Según un estudio publicado en *Science* en 2011, la información generada por la humanidad hasta el año 2007 se llegó a duplicar en tan solo cuatro años. Otro dato llamativo es que el total de información que se generó hasta el año 2003 en el mundo, es el equivalente a lo que generamos en dos días actualmente. Tal cantidad de datos serían imposibles de convertir en información útil con el software empleado tradicionalmente, consecuencia de ello entra en juego un nuevo concepto: el Big Data.

El Big Data es un término introducido por el informático teórico estadounidense John Mashey en 1998 en su artículo "*Big Data and the Next Wave of Infrastress*" en el cual alertaba sobre la cantidad de información que se generaría en los próximos años, así como la necesidad de nuevas herramientas que faciliten su procesado.

Fuente: http://computerhoy.com/noticias/internet/que-es-big-data-37627

Hoy día, existen multitud de definiciones para este concepto. Algunos analistas, para explicar este fenómeno hacen referencia a tres uves (3V): Volumen, Velocidad y Variedad. Es decir una gran cantidad de datos de muy diversa naturaleza generados a tal velocidad que supera la capacidad del software tradicional para ser procesados. No obstante, conforme ha ido creciendo la cantidad de personas e instituciones que lo emplean y que se preocupan por otorgarle la importancia que merece, se han ido introduciendo más Vs, como Veracidad, Viabilidad, Visualización y Valor poniendo de

manifiesto la necesidad de garantizar que dicha información sea fiable, la capacidad de los usuarios para generar un uso eficaz a partir de ella, la forma en la que esta es presentada para facilitar su interpretación, así como la capacidad que tiene la información de ser convertida en conocimiento.

Otra forma de entender el Big Data, trata del conjunto de medios y herramientas capaces de convertir volúmenes masivos de datos en información útil para la toma de decisiones, es decir, la recopilación ordenada y procesada de grandes cantidades de datos para hacer posible su interpretación y justificar en ella las decisiones que tomamos. Aquí es donde entraría en juego los métodos estadísticos, herramientas capaces de contrastar hipótesis y de establecer relaciones de causalidad con el fin de dar forma y añadir valor a la información de la que se dispone.

No es necesario buscar una definición universal para definir el Big Data dado que es algo que se encuentra en constante evolución y puede ser interpretado de diferentes formas dependiendo del uso que se va a hacer de él. Lo que si debemos tener claro es que es un concepto estrechamente relacionado con las tecnologías de la información, pudiéndose entender como una consecuencia de estas, y que a su vez necesita de más tecnología para su procesamiento. En otras palabras, el Big Data nace del desarrollo tecnológico, pero ambos necesitan del otro para seguir creciendo: el aumento de la tecnología en la sociedad es el desencadenante de grandes volúmenes de información, que son recopilados y analizados gracias a los métodos estadísticos para implantar y optimizar mejoras tecnológicas.

Ahora toca preguntarse lo siguiente: ¿Quién genera información? ¿Quién la emplea? ¿Con qué finalidad?

1.2. ¿Quién genera información?

Todas las personas y entidades, mediante las diferentes acciones que desempeñamos, emitimos información continuamente, dejando entrever nuestras necesidades, preferencias, gustos, comportamientos y actitudes ante distintas situaciones. La tecnología con la que actualmente compartimos nuestras vidas facilita que esta información pueda ser recogida por nosotros mismos o por otras personas o empresas mediante diferentes softwares informáticos. Un ejemplo sería todo aquello que realizamos con nuestro teléfono móvil, desde enviar un *sms*, hasta subir una foto a las redes sociales, incluso simplemente tener la localización activada. Otro ejemplo puede

ser la navegación web: estamos emitiendo información con cada clic que hacemos en la red, cuando pinchamos en un anuncio, cuando introducimos una búsqueda en google o cuando enviamos un correo electrónico.

Fuente: https://www.emaze.com/@ACCTZIQF/TICs

1.3. ¿Quién la emplea? ¿Con qué finalidad?

Estamos emitiendo información sobre qué nos gusta, con qué personas mantenemos relación, dónde nos encontramos en cada momento, qué cosas tienen interés para nosotros, etc. Los proveedores de dichos servicios (telefonía móvil, página web, redes sociales) pueden emplear estos datos para la toma de sus decisiones, incluyendo mejoras en sus aplicaciones y ofreciendo una experiencia más personalizada para el usuario. Pero, ¿solo en beneficio del cliente? No, ni mucho menos. Las empresas también usan dicha información en beneficio propio ya que conocer mejor a sus clientes les permite **establecer perfiles de usuarios que emplean para dirigir sus acciones de marketing y captación de clientela hacia un público similar**.

Un ejemplo de ello sería cuando una web para realizar compras online nos pide registrarnos con nuestro correo electrónico, nombre, edad… Con ello se consigue asociar un perfil sociodemográfico a las diferentes acciones que emprendemos dentro de la página de forma que se obtengan conclusiones del tipo de "las mujeres son más indecisas que los hombres comprando aparatos informáticos" debido a la medición de la cantidad de clics en productos relacionados previamente a la acción de compra. Otra conclusión puede ser "existe una alta correlación entre la compra de ropa deportiva y de material escolar" porque se ha observado que existe un perfil definido de personas que suelen

comprar ambos tipos de productos. En base a esta información, la empresa recomendará muchos ordenadores a los perfiles de mujeres que hayan mostrado interés en este producto, o anunciará material escolar al lado de su ropa deportiva y viceversa. Estos son solo algunos ejemplos de cómo una empresa de venta online utiliza los datos que les ofrecemos en su beneficio, pero existen utilidades diferentes según la naturaleza de la empresa. Estos ejemplos pueden verse como una demostración adelantada de la idea principal que persigue este proyecto: **la importancia del estudio estadístico de correlaciones marginales entre variables**.

No son sólo las empresas los únicos agentes interesados en la información, también lo son los poderes públicos e instituciones de diversa naturaleza, por motivos tanto económicos como no económicos, pero indistintamente, siempre se pretende realizar diagnósticos fiables que permitan anticiparse al futuro, es decir, tomar decisiones precisas en base a una estimación de lo que ocurrirá.

De este modo se puede realizar una predicción sobre la evolución de la tasa de desempleo de una localidad, o estudiar cuales son las calles de Los Ángeles más propensas a ser escenarios de delitos y crímenes, o determinar qué personas son más propensas a sufrir un determinado efecto secundario de cierto medicamento, con sus correspondientes utilidades.

Este uso de la información que generamos las personas por los diferentes agentes se enfrenta a un dilema ético, el cual, Joan Cwaik, coordinador del Centro de Divulgación Tecnológica de la Facultad de Ingeniería y Ciencias Exactas de la Universidad de Buenos Aires, sintetiza de la siguiente forma: "*Resolver estos dilemas no parece una tarea fácil, especialmente por la frontera difusa entre privacidad y bien público que dificulta su reglamentación legal*." Además, hace referencia a la facilidad de copia y difusión de los datos, así como al aumento de valor que les otorga a los mismos ser conectados con otros como argumentos para poner de manifiesto el crecimiento y evolución de estas prácticas en un futuro inminente y por consiguiente la aparición de los correspondientes dilemas sobre la ética del Big data. ¿Estamos ante una oportunidad de optimizar el modo en el que llevamos a cabo las diferentes tareas de nuestra vida cotidiana? ¿O nos encontramos ante una verdadera amenaza a nuestra privacidad alimentada por el instinto de negocio de las grandes empresas?

2. Importancia del Big Data en el Marketing.

El marketing es una de las áreas de la empresa donde el Big Data juega un papel fundamental. Marketing puede definirse según la RAE como "*el conjunto de principios y prácticas que buscan un aumento del comercio,* ***especialmente de la demanda***", como sabemos, estos principios y prácticas pueden agruparse según su relación con cada uno de los 4 elementos que componen el marketing mix: producto, precio, comunicación y distribución.

En un contexto socioeconómico globalizado donde los consumidores cada vez disponen de más información sobre el mercado y los productos y servicios que se ofrecen, éstos se vuelven más exigentes, desarrollan gustos y preferencias muy dispares, lo que conlleva que las empresas busquen constantemente la diferenciación e innovación para destacar sobre la competencia y ampliar así su cuota de mercado.

Esta necesidad, relativamente reciente, de diferenciar los productos y servicios respecto de los competidores es lo que hace que las empresas se preocupen cada día más por sus labores de marketing. Muchas han visto en el Big Data la oportunidad perfecta para explotar la información de la que disponen para realizar exhaustivos estudios de mercado y **determinar con precisión qué producto quiere el consumidor**, a qué precio lo quiere, dónde lo quiere, y cómo y **a quién hay que comunicárselo para que lo quiera**. Nos encontramos en un momento en que ninguna empresa importante funcionaría sin llevar a cabo estas prácticas.

Entre las infinitas aplicaciones que puede tener el Big Data en el campo del marketing, encontramos aquellas que pretenden determinar qué atributos son valorados por el consumidor y qué necesidades puede tener este para aplicarlas al desarrollo y mejora de los productos, otras aplicaciones estarían enfocadas a medir la sensibilidad de la demanda y otras magnitudes ante variaciones del precio, otras van dirigidas hacia la optimización de la función de distribución.

Pero entre las aplicaciones más interesante y útiles que tiene el análisis de datos en el marketing encontraríamos aquellas estrechamente relacionadas con la labor de comunicación comercial, especialmente las dedicadas a la **determinación del público objetivo** o "***target***". La utilidad que aporta el Big Data en ello es la posibilidad de ofrecer un determinado producto sólo a quien realmente es un consumidor potencial y no a todo el mercado, ya que de ser así, la saturación de publicidad sería tal que el consumidor la

percibiría como intrusiva y molesta, además de que la inversión por parte de la empresa sería muy costosa al afectar de manera positiva a tan solo un pequeño porcentaje del público recepto

Fuente: http://www.emprender-facil.com/es/publico-objetivo-definelo-para-el-exito-de-tu-negocio/

2.1. Modelos de negocio basados en el Big Data

Cada día son más las empresas que ven en el Big Data una herramienta ideal para sacarle partido a estos grandes flujos de datos al obtener beneficios adicionales derivados de un mayor nivel de información a la hora de tomar decisiones. Así, nos encontramos con casos como los supermercados, que emplean la tarjeta de fidelización como instrumento para recoger información personal, que asocian a cada compra que hace el individuo **estableciendo perfiles de consumidores** a tener en cuenta en sus posteriores decisiones sobre marketing.

Este tipo de prácticas han demostrado resultados tan sorprendentes que cada vez son más aquellas empresas que nacen con la principal intención de generar valor añadido a su producto o servicio basándose en el análisis de grandes volúmenes de datos. Aquí es donde reside el principal objetivo de este proyecto de final de grado, en la demostración de que **un adecuado análisis de la información que dispone una empresa puede ser de gran utilidad para obtener mayor rentabilidad de su actividad comercial.**

Un ejemplo de este tipo de empresas es Spotify. La conocida plataforma de música en *streaming* emplea un algoritmo que determina nuestro perfil para recomendarnos música acorde a nuestros gustos, momentos del día e incluso estados de ánimo. Además, permite la opción de vincular la cuenta

a Facebook, con ello consigue, además de asociar unos gustos musicales a un perfil sociodemográfico, basar las recomendaciones en los gustos de tus amigos y, en definitiva, ofrecer una experiencia más personalizada.

Otro ejemplo de empresa que dedica al Big Data una parte importante de sus recursos es Air BnB: conocen características tanto del huésped como del propietario que emplean para recomendar encuentros exitosos, priorizan resultados con mayor índice de éxito en sus búsquedas para asegurar que el cliente realice la reserva antes de abandonar la página, así como para asegurar su satisfacción etc. Bernard Marr en su interesante libro "*Big Data in Practise, (Wiley, 2016*)" afirma que la compañía desde sus inicios ha acumulado hasta 1,5 petabytes sobre hábitos vacacionales y preferencias de alojamiento. Además, desde 2015, Air BnB se ha decidido a dar un paso más: aplicar el Big Data a las imágenes. La empresa pretende analizar la distribución de los muebles, la luz y otros elementos decorativos con altas probabilidades de éxito según sus registros, de forma que pueda recomendar, a los usuarios que anuncien sus apartamentos, una distribución alternativa de dichos elementos a fin de que la fotografía sea más atractiva aumentando así las probabilidades de reserva.

Facebook, la red social más usada en el mundo, es el caso perfecto de empleo del Big Data para otorgar un valor añadido al servicio que ofrece, y es que son innumerables las funciones de la red social que serían imposibles de incorporar de no ser por estas técnicas. Desde las más sencillas, como recomendar personas con las cuales tienes amigos en común, hasta las más complejas como recomendar páginas y grupos acorde a tus intereses. También hay que destacar las posibilidades que ofrece Facebook de contratar anuncios, y es que, **para rentabilizar una inversión publicitaria, la primera condición necesaria es hacer llegar el mensaje al público objetivo deseado**. Facebook lo consigue a la perfección. Cuando queremos anunciar nuestro producto o empresa en Facebook, podemos escoger un perfil de destino muy concreto, segmentando por sexo, edad, localización geográfica, intereses etc.

Por último, no podemos hablar de empresas que emplean el Big Data sin mencionar el ejemplo de Google. Google tiene como objetivo organizar la información mundial para hacerla universalmente accesible y útil. Ello conlleva ordenar y segmentar en base a múltiples criterios una gran

cantidad de datos de manera constante, para ello emplea un complejo algoritmo encargado de posicionar en primeros lugares aquellos sitios web de calidad que mejor información estiman que va a ofrecer al usuario para unas determinadas "*keywords*" introducidas en su buscador. No obstante, Google es mucho más que un buscador, y entre sus productos se encuentra la red social Google+, YouTube, Google Drive, Google Maps y un largo etcétera. Todas estas aplicaciones requieren de un permanente análisis de la información que envían sus usuarios para su funcionamiento y mantenimiento, y el hecho de haberlo realizado con éxito hasta ahora es la razón principal por la que Google es una de las empresas más importantes del mundo.

Estos son solo algunos casos concretos de empresas cuyo éxito radica en el análisis de la información que les proporcionan sus clientes, compartiendo el objetivo de ofrecer una experiencia más personalizada y añadir así un valor extra al servicio que ofrecen. Esta tendencia lleva observándose de manera creciente en los últimos años incluso en empresas que no son de naturaleza tecnológica, pero que encuentran interesante el estudio de su mercado mediante estas técnicas de análisis. Por tanto, queda más que justificada la confianza de las empresas en el Big Data como herramienta para apoyar y basar la toma de decisiones de diversa índole, con especial importancia hacia aquellas relacionadas con el marketing.

2.2. Importancia del Big data en el Marketing de banca

El sector bancario es uno de los que más tiempo lleva empleando el análisis de datos para la toma de sus decisiones, y es que, a diferencia de otros sectores, los empleados de la banca "deciden" acerca de prestar dinero o no a sus clientes, con su correspondiente riesgo de crédito. Sin embargo, desde hace ya varias décadas no es la persona que encontramos al otro lado de la mesa la encargada de decidir si pone en riesgo el capital de la entidad. **Ahora es un ordenador, que mediante un complejo algoritmo realiza una estimación de la probabilidad que un determinado cliente tiene de incumplir un préstamo, en base a diferentes características como su sexo, edad, salario, anteriores préstamos, etc.**

Como hemos comentado anteriormente, cada vez se encuentra a disposición de los agentes económicos una mayor cantidad de información de distinta naturaleza, así como avanzadas herramientas tecnológicas que garantizan cierta precisión en su análisis. Ello ha supuesto para las entidades financieras un aumento de las razones por las que resulta

de gran utilidad este empleo del análisis de datos que, además de agilizar la toma de decisiones, es capaz de **personalizar la oferta comercial de manera sencilla con el fin de dirigirla al cliente potencial con mayores probabilidades de compra del producto o servicio.**

En este marco de actuación nacen consultorías estratégicas como *AIS Group* que emplean modelos estadísticos y matemáticos para crear sistemas informáticos que sirvan como soporte a la toma de decisiones gracias a una mezcolanza de tecnología, finanzas y visión estratégica. Estos servicios son demandados cada vez más por las empresas, entre ellas las entidades bancarias, que ven la oportunidad de optimizar su actividad y obtener mayores beneficios.

Entre los servicios que ofrecen estas consultorías encontraríamos:

- ➔ Gestión integral del riesgo de crédito: Permite labores como la determinación de un "scoring" para cada préstamo que se solicita de forma que las decisiones crediticias sean tomadas a partir de un sistema automático, siendo estas más acertadas debido a su capacidad predictiva y a la eliminación de cualquier sesgo humano.
- ➔ Investigación de mercados: se basa en un estudio detallado del cliente y su entorno para saber qué necesita y quién lo necesita. Con ello se consigue ofrecer un producto o servicio adecuado solo al cliente potencial adecuado, aumentando así el éxito del impacto comercial.
- ➔ Valoración automática de inmuebles: permite calcular el valor de los bienes inmuebles en base a distintos parámetros como su localización geográfica, datos demográficos de la zona donde se ubica, indicadores económicos tanto globales como locales etc. La finalidad es cuantificar el valor presente y futuro de las propiedades tanto de la entidad financiera, como de sus clientes, además de cualquiera que tenga interés para ella por ser un activo interesante para invertir.
- ➔ Valoración de los seguros que ofrecen: muchas entidades de crédito actúan a la vez como aseguradoras de hogar, coche, vida… es necesario realizar una estimación fiable de la probabilidad de que el cliente tenga un problema y que por ello la aseguradora deba pagar. La cuantía de la contratación del seguro es calculada en función de dicha probabilidad así como del valor de los bienes asegurados y del mayor o menor margen que desee tener la compañía.

Estas son sólo algunas de las muchas aplicaciones para las que el sector bancario emplea el Big Data. Para la realización de este trabajo perteneciente al Grado de Marketing e investigación de mercados resulta de especial relevancia la segunda. Por este motivo, la parte analítica se fundamente una base de datos de la cartera de clientes de una institución bancaria, que será analizada con objeto de demostrar cómo se puede emplear el análisis de datos en este sector para obtener información capaz de apoyar a la toma de decisiones y obtener mejores resultados.

2.3. Regresión como herramienta predictiva

La gran utilidad de esta herramienta queda justificada por su intensivo uso por parte de individuos, instituciones y finalidades de muy diversa naturaleza, desde su uso con intenciones puramente económicas hasta la investigación en todos sus campos. Y es que como dice el periodista Juan F.Cía en su artículo para la plataforma *BBVA Open 4U*: "*El análisis predictivo es un cambio en el juego de los negocios*."[…] ***"****Ya no hay excusas. Todas las compañías del mundo tienen a su alcance soluciones de análisis de datos tan avanzadas y sencillas de usar que no saber lo que sucederá dentro de seis meses es pecado mortal".*

La regresión es una herramienta analítica de gran utilidad que se emplea en Big Data para realizar predicciones. Se trata de un proceso estadístico para cuantificar relaciones entre variables, el cual puede abordarse desde dos dimensiones complementarias: la primera consiste en hallar un valor estimado para una variable dependiente en base a la información que nos aportan varias variables independientes, la segunda sería el estudio de esa dependencia, es decir, cómo varía el valor de la variable dependiente cuando se modifica el valor de una independiente.

En el campo del marketing resulta de especial interés el estudio de variables de carácter cualitativo, es decir, aquellas que sus valores no representan cantidades, sino cualidades o características, de manera que los casos pueden ser clasificados en distintas categorías a partir de ellas. Ejemplos de variables cualitativas pueden ser: el sexo, la ciudad de residencia de un individuo, o cualquiera obtenida a partir de una encuesta cuya respuesta no sea un dato numérico.

Cuando se realizan regresiones para explicar este tipo de variables, debido a su carácter cualitativo, lo que se pretende estimar es la probabilidad de cumplimiento de una de las categorías respecto a otra de diferencia. Cuando solo existen dos categorías, es decir, la variable cualitativa es dicotómica, se emplea un modelo de regresión logística binaria para predecir la probabilidad de que se cumpla una de las categorías, siendo el valor complementario la probabilidad de que se cumpla la categoría de referencia.

Su forma funcional es la siguiente:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

Es decir, **se trata de la estimación de la probabilidad de que se cumpla un determinado suceso para un individuo/caso con determinadas características, frente a aquellos que no las cumplen.**

En nuestro caso, para demostrar la utilidad que el análisis masivo de datos tiene en el sector bancario, el suceso cuya probabilidad se intentará predecir es la suscripción a un depósito de plazo fijo.

Los datos a tratar han sido importados de una institución bancaria portuguesa anónima para un proyecto de investigación de la Universidad de Minho (Portugal) en el cual los profesores Sérgio Moro, Raul Laureano y Paulo Cortez demostraban la efectividad del Buisiness Intelligence y Data Mining en las campañas de comunicación directa. El artículo fue publicado en las Actas de la Conferencia Europea de Simulación y Modelización, en Octubre del año 2011.

3. Tratamiento empírico

Como hemos comentado, la base de datos analizada pertenece a una entidad bancaria portuguesa, en la cual se recoge diferente información sobre su cartera de clientes. Parte de estas variables son de carácter sociodemográfico: edad, estado civil y nivel educativo. Otras, son información que posee la entidad debido a la relación que tiene con el cliente: si tiene crédito en incumplimiento, su saldo medio anual, si tiene contratado préstamo personal, y si tiene contratado préstamo hipotecario. El último bloque de variables, están relacionadas con la campaña actual de comunicación comercial que se está llevando a cabo para ofrecer un depósito a plazo, se incluye la fecha del último contacto, su duración en segundos (de la llamada telefónica), el número de contactos realizados para cada cliente en la campaña actual, y el número de contactos realizados para cada cliente en la campaña anterior, además, el resultado de esta campaña anterior. Para finalizar, la variable clave en el análisis: el éxito de la campaña de comunicación, es decir, si el cliente ha suscrito el depósito o no lo ha hecho.

La estrategia actual de la entidad es establecer contacto telefónico con toda su cartera de clientes. Se ha seleccionado una muestra aleatoria de la base de datos principal que cuenta con 4521 clientes, de los cuales 4000 no han suscrito el depósito, y 521 sí. Nuestro objetivo es analizar esta información para determinar la cartera de clientes que más probabilidades tenga de contratar este producto financiero, con objeto de dirigir solo a ellos la campaña y así rentabilizar al máximo el coste que se asume.

Con motivo de poder extender el análisis a nuevos clientes, solo se han escogido aquellas variables que se pueden conocer sin necesidad de haber establecido contacto comercial en una campaña anterior.

En primer lugar se ha procedido a realizar un análisis exploratorio para conocer la muestra de la que se dispone, así como estudiar la dependencia entre la variable a predecir y aquellas que usaremos para explicarla. Posteriormente se realizará un modelo de regresión logística para predecir la probabilidad de éxito de la variable dependiente y determinar las variables más influyentes. Para finalizar, se hará una recomendación de cómo la entidad bancaria debería gestionar su campaña de comunicación comercial directa en base a las conclusiones obtenidas.

3.1. Estudio de las variables y análisis exploratorio

Suscripción: es la variable que representa el suceso que se quiere predecir, es decir, si el cliente ha suscrito un depósito a plazo o no. Representa el resultado de la campaña comercial.

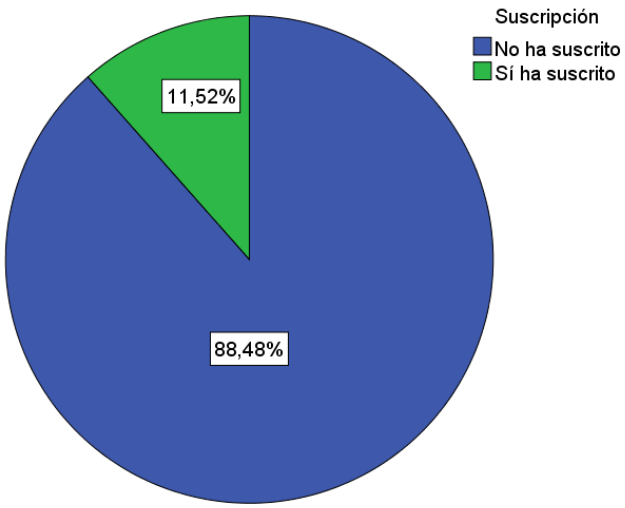

Gráfico A - 3.1 Distribución de suscripción.
***Fuente**: elaboración propia*

Para comenzar el análisis exploratorio se ha procedido a realizar un analisis discriminante de las dos variables cuantitativas que se disponen respecto de la variable a predecir: el saldo medio anual y la edad del individuo. El discriminante clasifica correctamente el 56,8% de los casos, además, mediante el contraste de la Lambda de Wilks, se afirma que las medias y las desviaciones típicas de estas variables son significativamente diferentes según los individuos que contratan el depósito y los que no. Con esta información ya se puede anticipar el hecho de que la muestra tiene características diferentes para los dos subgrupos (suscribir depósito / no suscribir depósito).

Lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,998	10,105	2	,006

Tabla A - 3.1 Lambda de Wilks.
***Fuente**: elaboración propia*

Resultados de clasificación[a]

			Pertenencia a grupos pronosticada		
		Suscripción	No ha suscrito	Sí ha suscrito	Total
Original	Recuento	No ha suscrito	2334	1666	4000
		Sí ha suscrito	285	236	521
	%	No ha suscrito	58,4	41,7	100,0
		Sí ha suscrito	54,7	45,3	100,0

a. 56,8% de casos agrupados originales clasificados correctamente.

Tabla B - 3.2 Clasificación según el discriminante.
***Fuente**: elaboración propia*

Las variables escogidas para explicar la probabilidad del suceso son aquellas que individualmente son significativas mediante un contraste Chi-cuadrado. (Las tablas de correspondencias con sus respectivos contrastes y residuos estandarizados corregidos se encuentran en Anexo 1.2.) A continuación se presentan los gráficos de cada variable condicionados por la "Suscripción".

Estado civil: Se puede observar como hay menos porcentaje de personas casadas entre los que sí han suscrito el depósito respecto a aquellos que no, es decir, mayor proporción de solteros y divorciados o viudos entre los suscriptores del depósito que entre los que lo han rechazado. Por tanto, en ausencia de más variables, se intentaría recomendar el producto a personas solteras y divorciadas o viudas.

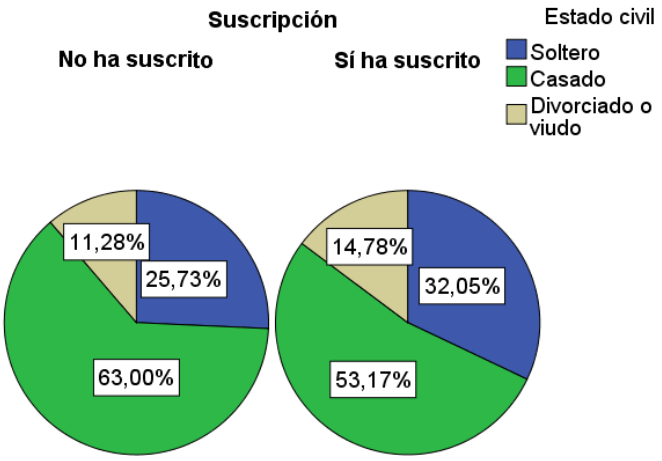

Gráfico B - 3.2. Distribución de Estado civil.
***Fuente**: elaboración propia*

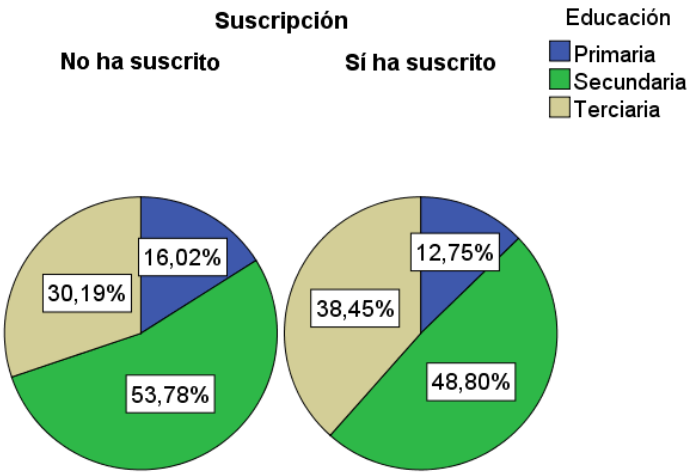

Gráfico C - 3.3. Distribución de Educación.
***Fuente**: elaboración propia*

- **Educación**: esta variable agrupa los tres niveles posibles de educación: primaria, secundaria y terciaria. Como se puede observar en el gráfico, el porcentaje de personas con educación superior es mayor entre aquellos que sí han suscrito el depósito respecto a aquellos que no lo han suscrito, por lo tanto es más fácil encontrar clientes entre la población con nivel terciario (superior) de estudios.

Préstamo vivienda: Se trata de una variable dicotómica que diferencia a los clientes según tienen contratado o no un préstamo para su vivienda. Se aprecia claramente que hay más proporción de público con hipoteca entre los que no han contratado el depósito y menos entre los que sí lo han hecho. Por tanto, es más probable encontrar un suscriptor para el depósito entre aquellos que no tienen préstamo hipotecario.

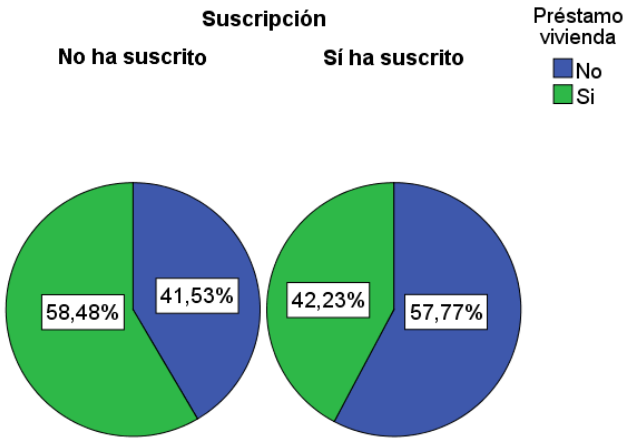

Gráfico D - 3.4. Distribución de Préstamo vivienda. ***Fuente****: elaboración propia*

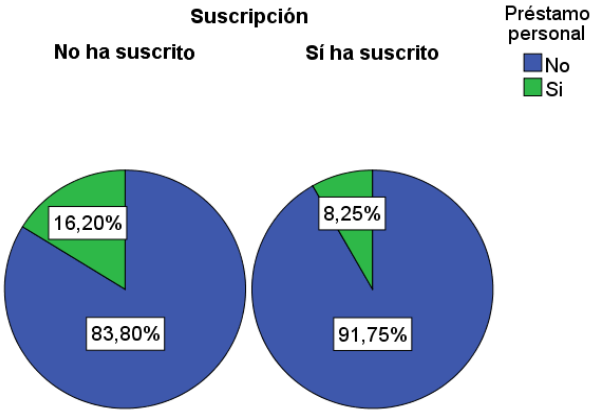

Gráfico E - 3.5. Distribución de Préstamo personal. ***Fuente****: elaboración propia*

Préstamo personal: Es una variable dicotómica que segmenta aquellos clientes que tienen contratado un préstamo personal, de aquellos que no lo tienen. Con préstamo personal se refiere a que no va dirigido al pago de la vivienda. Como puede observarse, es más probable tener un préstamo si el cliente no está suscrito al depósito a plazo que en caso de sí estarlo, por tanto conviene dirigirse a aquellos que no tienen préstamo personal para que se suscriban al depósito a plazo, ya que existen más probabilidades de éxito.

Resultado de campaña anterior: Esta variable dicotómica diferencia aquellos clientes que en la campaña anterior contrataron en depósito de aquellos que no. La muestra empleada para esté análisis de correspondencias es menor dado que solo incluye los clientes a los que ya se les ofreció el producto en la campaña anterior. Es la variable más significativa, ya que es aquella cuya distribución de frecuencias para cada subgrupo es más dispar. De aquellos clientes que han suscrito el depósito, el 56,85% lo hicieron ya el año pasado, por tanto es una variable muy a tener en cuenta a la hora de realizar la comunicación comercial.

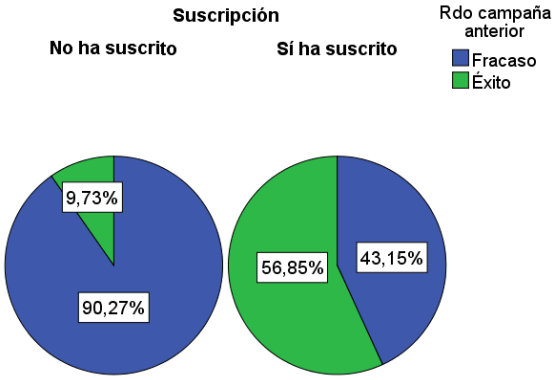

Gráfico F - 3.6. Distribución de Resultado campaña anterior
***Fuente**: elaboración propia*

Edad: Por último, la edad del individuo es una variable que aparece como cuantitativa en la base de datos. Sin embargo, si suponemos que las personas mayores son las que suelen contratar depósitos a plazo, dado que suelen tener más dinero y probablemente piensen en ahorrarlo para su jubilación, recodificaremos esta variable en cuatro intervalos: hasta 40 años de edad, de 41 a 50, de 51 a 60, y más de 60 años. De este modo comprobaremos si realmente aumentan las probabilidades de compra del producto con la edad.

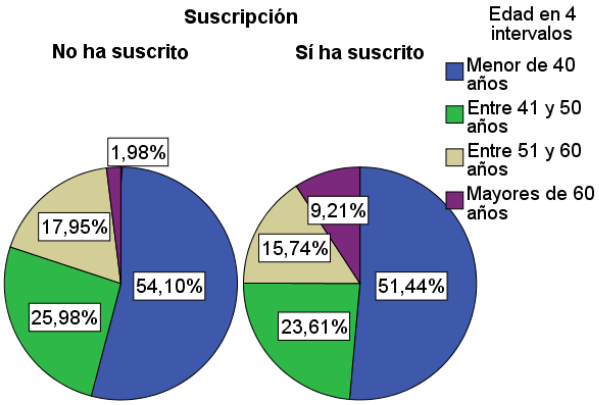

Gráfico G - 3.7. Distribución de Edad.
***Fuente**: elaboración propia*

Y efectivamente, como se muestra en el gráfico condicionado, la proporción de gente mayor de 60 años es notablemente superior entre los que sí suscriben el depósito que entre aquellos que no, por tanto, son un público objetivo muy a tener en cuenta a la hora de realizar las campañas comerciales.

A priori de la realización del modelo, tan solo mediante el análisis exploratorio de los datos ya se puede saber que las características que cumplen los individuos con más probabilidades de éxito son: solteros o divorciados y viudos, con un nivel de educación superior, sin préstamo personal ni préstamo hipotecario, mayores de 60 años, y que el año pasado hayan suscrito el depósito. Estas características, comparadas cada una con las de su respectiva variable, son indicadores generales de tener más dinero disponible para ahorrar, por tanto podemos afirmar que el análisis exploratorio nos ofrece resultados lógicos. La realización del modelo de regresión logística nos ayudará a determinar cuáles de estas son más importantes, así como la estimación de probabilidades de éxito para individuos concretos pertenecientes a una o varias de estas categorías.

3.2. Modelo de regresión logística binaria

Como hemos comentado anteriormente, cuando la variable que se pretende predecir es dicotómica, se emplea una regresión logística binaria capaz de estimar una probabilidad de ocurrencia del suceso a explicar. Se busca un modelo que pueda explicar la probabilidad del éxito de un contacto, es decir, si el cliente se suscribe al depósito. Con ello se pretende conocer qué variables son más influyentes en la suscripción de un depósito plazo, así como la magnitud y sentido de esta dependencia.

La variable dependiente es la **suscripción al depósito**, que toma valor "1" cuando éste es contratado por el cliente, y valor "0" cuando el cliente rechaza el producto. Las variables explicativas son:

Edad: variable compuesta por cuatro categorías que agrupa individuos según intervalos de edad (Toma valor "1" hasta 40 años, "2" entre 41 y 50, "3" entre 51 y 60, y "4" para más de 60).

Estado civil: compuesto por tres categorías ("1" soltero, "2" Casado y "3" Divorciado o viudo)

Educación: codificada con valor "1" para aquellos sin educación o con educación primaria, "2" para quienes tienen hasta educación secundaria y "3" para los que tienen educación superior.

Préstamo personal: variable dicotómica cuyo valor "1" representa a quien tiene contratado préstamo personal, y su valor "0" a quien no.

Préstamo vivienda: variable dicotómica cuyo valor "1" representa a quien tiene contratado préstamos para su vivienda, y "0" a quien no.

Estas han sido las variables escogidas para realizar la regresión por que cumplen tres requisitos básicos:

- Son datos de los que se puede disponer antes de iniciar la campaña comercial. Dado que el banco va a emplear esta información para seleccionar su público objetivo, las decisiones que se tomen gracias a este modelo tendrán lugar antes del inicio de la campaña de comunicación, por ello se excluyen variables recogidas durante la misma, como son el número de contactos realizados, su duración, la fecha etc.
- Todas las variables son significativas por separado en el hecho de contratar o no el depósito, por lo tanto es más fácil que lo sean conjuntamente.
- Ninguna de ellas acapara toda la significatividad del modelo, motivo por el que se ha excluido la variable "Resultado de la campaña anterior". Ésta tiene tanta significatividad que prácticamente explica por si sola la contratación del depósito, por lo que la dejamos fuera para que no reste importancia al resto de variables. Además, ello se debe a que su inclusión obligaría a reducir la muestra sólo a los clientes a los que ya se les ofreció el depósito anteriormente, dejando un modelo inservible para la captación de nuevos clientes. A pesar de que no pueda entrar en el modelo, en ningún caso significa que la variable no deba tenerse en cuenta, sino todo lo contrario, es un indicador de que se debe ofrecer el depósito a todo aquel que ya lo contrató alguna vez.

El método empleado para introducir las variables es "introducción por pasos hacia delante", el cual filtra la entrada/salida según el estadístico de Wald. El funcionamiento consiste en introducir en el modelo la variable más significativa en cada paso.

En primer lugar es necesario establecer una categoría de referencia de aquellas variables categóricas para que las demás puedan ser comparadas con esta. En este caso hemos escogido la de nivel inferior, de modo que, en el caso de la educación por ejemplo, se estudiará cómo influye en la probabilidad de suscripción teniendo estudios secundarios o terciarios respecto de cómo influye tener primarios. En otras palabras, no se compara el hecho de tener estudios secundarios con tener terciarios.

			Codificación de parámetro		
		Frecuencia	(1)	(2)	(3)
Edad en 4 intervalos	Menor de 40 años	2366	,000	,000	,000
	Entre 41 y 50 años	1108	1,000	,000	,000
	Entre 51 y 60 años	743	,000	1,000	,000
	Mayores de 60 años	117	,000	,000	1,000
Educación	Primaria	678	,000	,000	
	Secundaria	2306	1,000	,000	
	Terciaria	1350	,000	1,000	
Estado civil	Soltero	1150	,000	,000	
	Casado	2680	1,000	,000	
	Divorciado o viudo	504	,000	1,000	
Préstamo personal	No	3650	,000		
	Si	684	1,000		
Préstamo vivienda	No	1858	,000		
	Si	2476	1,000		

Tabla C - 3.3 Codificación de variables categóricas
***Fuente**: elaboración propia*

En la siguiente tabla encontramos las pruebas ómnibus sobre los coeficientes del modelo indican la significatividad del modelo en cada paso según se van incluyendo variables. Podemos observar como en todos los pasos el p-valor es menor que 0,05, y como el estadístico Chi-cuadrado del modelo va aumentando en cada paso. Por lo tanto, todas las variables que se incluyen son significativas, y ayudan a explicar el suceso.

		Chi-cuadrado	gl	Sig.
Paso 1	Escalón	55,573	3	,000
	Bloque	55,573	3	,000
	Modelo	55,573	3	,000
Paso 2	Escalón	33,287	1	,000
	Bloque	88,860	4	,000
	Modelo	88,860	4	,000
Paso 3	Escalón	22,288	2	,000
	Bloque	111,148	6	,000
	Modelo	111,148	6	,000
Paso 4	Escalón	18,586	1	,000
	Bloque	129,734	7	,000
	Modelo	129,734	7	,000
Paso 5	Escalón	9,148	2	,010
	Bloque	138,883	9	,000
	Modelo	138,883	9	,000

Tabla D - 3.4. Pruebas ómnibus de coeficientes del modelo
***Fuente**: elaboración propia*

Como es conocido, la prueba de Hosmer y Lemeshow es una medida de bondad del ajuste del modelo que consiste en dividir la muestra en varios intervalos para tratar de contrastar la distribución de los valores observados con la distribución de los esperados mediante una Chi-cuadrado. En otras palabras comprobar si los percentiles de la distribución empírica coinciden con los de la distribución empírica. Su hipótesis nula es que NO haya diferencias, por tanto se busca un nivel de significación alto para no rechazarla. En nuestro caso para el paso 5 contamos con un nivel muy alto de significación. Este es el test de bondad de ajuste en el que nos basamos para afirmar que nuestro modelo es válido.

Escalón	Chi-cuadrado	gl	Sig.
1	,000	1	1,000
2	2,886	4	,577
3	5,565	7	,591
4	8,788	8	,361
5	5,088	8	**,748**

Tabla E - 3.5 Prueba de Hosmer y Lemeshow
***Fuente**: elaboración propia*

Con la tabla de contingencia para la prueba podemos comparar los percentiles de la distribución observaba con la esperada, dado que se obtienen valores muy similares, podemos afirmar que contamos con un buen modelo.

		Suscripción = No ha suscrito		Suscripción = Sí ha suscrito		
		Observado	Esperado	Observado	Esperado	Total
Paso 5	1	392	394,492	22	19,508	414
	2	330	328,427	21	22,573	351
	3	404	397,629	26	32,371	430
	4	382	382,910	34	33,090	416
	5	366	375,042	48	38,958	414
	6	367	364,304	41	43,696	408
	7	402	403,951	56	54,049	458
	8	351	349,571	52	53,429	403
	9	303	298,025	46	50,975	349
	10	535	537,649	156	153,351	691

Tabla F - 3.6 Contingencia para Hosmery Lemeshow.
***Fuente**: elaboración propia*

SPSS nos da el resumen del modelo que indica el valor del estadístico -2 log de la similitud, útil para realizar comparaciones entre los pasos, y dos valores del R^2. Interpretamos el R^2 de Nagelkerke ya que esté varía entre 0 y 1. Este valor refleja el porcentaje de variabilidad de los datos que explica nuestro modelo, en este caso muy pequeño, (del 0,062 en el paso 5) que no siendo óptimo, aceptaremos el modelo dado que el objetivo final no es utilizarlo de forma predictiva, sino de forma exploratoria para lanzar la campaña de marketing con ciertas garantías de éxito.

Resumen del modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	3052,165[a]	,013	,025
2	3018,878[a]	,020	,040
3	2996,590[a]	,025	,049
4	2978,004[a]	,029	,058
5	2968,855[a]	,032	**,062**

a. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de ,001.

Tabla G - 3.7 Resumen del modelo
***Fuente**: elaboración propia*

Para finalizar, la tabla de las variables en la ecuación es la más importante dado que nos muestra qué variables se han introducido en cada paso, la significación de cada una en el modelo y cómo afecta cada una en el resultado estimado.

Las variables categóricas de más de dos categorías se introducen mediante una recodificación en “ceros” y “unos” según las observaciones pertenezcan o no al grupo. Así pues, el estado civil tiene tres categorías, hemos establecido “soltero” como referencia, por tanto se introducen dos variables al modelo, una que toma valor “1” cuando el individuo está casado (y el resto ceros) y otra que toma valor “1” cuando el individuo está divorciado o viudo” (y el resto ceros). Ocurre lo mismo para el resto de variables.

Como podemos observar, en el último paso de nuestro modelo todas las variables son significativas excepto la variable correspondiente a “Estadocivil(2)” que corresponde con la categoría “Divorciado o viudo”, “Educación(1)” que corresponde con la categoría “Estudios secundarios” y los dos intervalos de edad centrales correspondientes a los individuos entre 41 y 50 años, y entre 51 y 60 años, por tanto, procederemos a cuantificar

la importancia de las demás variables en el proceso de predicción. (Ver tabla completa en Anexo II)

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 5[e]	Estadocivil			18,732	2	,000	
	Estadocivil(1)	-,437	,118	13,644	1	,000	,646
	Estadocivil(2)	,008	,165	,002	1	,962	1,008
	Educación			9,040	2	,011	
	Educación(1)	,237	,156	2,297	1	,130	1,267
	Educación(2)	,457	,163	7,875	1	,005	1,580
	Préstamo_vivienda(1)	-,502	,101	24,713	1	,000	,605
	Préstamo_personal(1)	-,659	,167	15,485	1	,000	,518
	EdadCuali4			49,106	3	,000	
	EdadCuali4(1)	,042	,126	,109	1	,742	1,042
	EdadCuali4(2)	,032	,151	,044	1	,834	1,032
	EdadCuali4(3)	1,483	,221	45,242	1	,000	4,408
	Constante	-1,806	,178	103,397	1	,000	,164

Tabla H - 3.8 Variables en la ecuación
***Fuente**: elaboración propia*

Para determinar la incidencia de las distintas variables en el modelo nos fijaremos en la columna "Exp(B)", este valor también es conocido como ODD ratio, se trata de la ventaja relativa que tiene el suceso a explicar para cada variable (entendiendo ahora categoría como una variable), es decir, muestra un cociente entre la probabilidad de que ocurra el suceso (en nuestro caso suscripción al depósito) y la probabilidad de que no ocurra. Por tanto, las variables que más información aportan son aquellas cuyo ODD ratio más se aleja del valor 1, si este es superior, la probabilidad de que ocurra el suceso es mayor del 50%, y si es menor de uno, la probabilidad de que ocurra el suceso es menor del 50%.

- **Estadocivil(1)**: Su ODD ratio es de 0,646, lo cual nos indica que los casados tienen una probabilidad de no suscribir el depósito del 61% y una probabilidad de suscribirlo del 31% , frente a la categoría de referencia que es soltero (véase el 3.2 en el cual el 63 % de los no suscritos son casados mientras sólo un 53% de los que sí suscriben los son). En definitiva, hay más probabilidades de éxito entre los solteros que entre los casados.
- **Educación(2):** Su ODD ratio es de 1,58, lo que indica que la probabilidad de éxito para esta categoría es mayor que la de fracaso, en concreto, la probabilidad de éxito

es del 61% y la de fracaso es el 39% respecto a tener estudios primarios (véase el gráfico 3.3 en el que 30% de aquellos que no suscribieron tienen estudios superiores, frente al 38% de aquellos que si suscribieron). Hay más probabilidades de éxito entre aquellos que tienen estudios superiores.

- **Préstamo_vivienda_(1):** su ODD ratio es de 0,605, lo que nos indica que entre aquellas personas que tienen contratado un préstamo para pagar su vivienda hay probabilidad de éxito del 37% frente a una probabilidad de fracaso del 63%, en comparación con aquellas que no lo tienen (véase el gráfico 3.4. donde la proporción de aquellos que tienen préstamo es notablemente menor entre los que sí han suscrito que entre los que no). Por tanto debe ofrecerse el depósito a plazo a aquellos que no tengan contratado préstamo hipotecario.
- **Préstamo_personal(1):** ocurre algo parecido con las personas que tienen un préstamo personal, su ODD ratio es de 0,518, es decir, la probabilidad de éxito es un 34% frente a un 66% de fracaso en comparación con aquellas personas que no lo tienen (véase el gráfico 3.5 donde la proporción de aquellos que tienen préstamo contratado es mayor entre los que no han suscrito que entre aquellos que sí). Por tanto debe ofrecerse el depósito a plazo a aquellos que no tengan contratado préstamo personal.
- **EdadCuali4(3)**: Su ODD ratio tiene un valor de 4,408, el valor más alto del modelo. Esto quiere decir que la probabilidad de éxito es 4,408 veces mayor que la de fracaso, 81% frente a 19% frente a la categoría de referencia (menor de 40 años). Por tanto las personas mayores de 60 años tienen una probabilidad muy superior de contratar el depósito que aquellos menores de 40 (véase gráfico 3.7, en el que la proporción de casados entre los que sí han suscrito es del 9%, mientras que entre aquellos que no lo han hecho es del 2%. Debe ofrecerse el depósito a todas las personas mayores de 60 años debido a las altas probabilidades de éxito que se tienen.

La interpretación global del modelo consiste en afirmar que los individuos que más probabilidades tienen de suscripción de un depósito a plazo tienen las siguientes características: son personas solteras (o divorciadas o viudas, ya que vemos que no existe diferencia con esta categoría, pero en ningún caso casadas), su nivel de estudios es superior, no tienen contratado un préstamo ni personal ni de vivienda, y además son mayores de 60 años. Los coeficientes para estas variables son: 0 por estar soltero, 0,457 por tener estudios superiores, 0 por no tener préstamo personal, 0 por no tener préstamo para vivienda y 1,483 por tener más de 60 años, además el coeficiente de la constante es -1,806.

La cuantía de esa probabilidad máxima se obtiene a partir de la siguiente expresión:

$$Pmax(Suscripción\ de\ depósito) = \frac{1}{1+e^{-(-1{,}806+0+0{,}457+0+0+1{,}483)}} = 0{,}535$$

Es decir, la probabilidad máxima es de un 53,5%.

Si se quisiera calcular la probabilidad de suscripción al depósito de cualquier individuo, solo haría falta conocer a qué categoría pertenece de cada una de las variables incluidas en el modelo y reproducir la ecuación anterior con los coeficientes el nuevo individuo. Por ejemplo, suponemos que el banco cuenta con dos nuevos clientes y quiere conocer que probabilidad tienen de aceptar el depósito a plazo para tenerlo en cuenta en su campaña de marketing:

	Estado civil	**Nivel de estudios**	**¿Tiene préstamo personal?**	**¿Tiene préstamo para vivienda?**	**Edad**
Individuo 1	Casado	Secundarios	No	Sí	Entre 51 y 60
Coeficientes individuo 1	-0,437	0,237	0	-0,502	0,032
Individuo 2	Divorciado	Primarios	Si	Sí	Entre 41 y 50
Coeficientes individuo 2	0,008	0	-0,659	0,605	0,042
Coeficiente de la constante: -1,806					

Tabla I - 3.9 Ejemplo de características de clientes.
***Fuente**: elaboración propia*

$$P1(Suscripción\ de\ depósito) = \frac{1}{1+e^{-(-1{,}806-0{,}437+0{,}237+0-0{,}502+0{,}032)}} = 0{,}077$$

Es decir, el individuo 1 tiene una probabilidad de suscripción del 7,7%.

$$P2(Suscripción\ de\ depósito) = \frac{1}{1+e^{-(-1{,}806+0{,}008+0-0{,}659+0{,}605+0{,}042)}} = 0{,}14$$

Es decir, el individuo 2 tiene una probabilidad de suscripción del 14%.

Para finalizar el análisis, estudiaremos cómo nuestro modelo clasifica los individuos según la variable que se quiere pronosticar, de modo que se realiza una comparación entre lo observado y lo esperado para determinar el porcentaje de aciertos y de errores que comete el modelo. Un aspecto muy importante a tener en cuenta es que las probabilidades de éxito (suscribir un depósito a plazo) para cada individuo son muy bajas, por tanto, no podemos considerar como pronóstico acertado aquellos que tienen probabilidad mayor del 50% porque apenas los hay. A pesar de ello, es interesante saber cuáles son los casos con más probabilidades de éxito (aunque ésta sea baja), para poder dirigir hacia ese público las acciones de comunicación comercial.

Suponiendo que nuestra cartera de clientes es de 4521, si deseamos dirigir la campaña de comunicación a aquellos con más del 10% de probabilidades de éxito, tendríamos un público objetivo de prácticamente la mitad, de los que nos comprarían el producto 338 personas (el 15% del público objetivo) entre los cuales se incluirían el 67% de aquellos que aceptarían (en caso de que nos dirigiésemos a toda la cartera), perdiendo el 37% restante.

Tabla de clasificación[a]

			Pronosticado		
			Suscripción		Corrección de
	Observado		No ha suscrito	Sí ha suscrito	porcentaje
Paso 5	Suscripción	No ha suscrito	1994	**1838**	52,0
		Sí ha suscrito	164	**338**	**67,3**
	Porcentaje global				53,8

a. El valor de corte es ,100

Tabla J - 3.10 Tabla de clasificación con corte en 0,10
***Fuente**: elaboración propia*

Si deseamos dirigir la campaña comercial a aquellos con más del 20% de probabilidades de éxito, obtendríamos un público objetivo de tan solo 396 personas, de las cuales contratarían el depósito 102 (el 25% del público objetivo) que representarían el 20,3% de aquellos que lo contratarían (de la cartera total), perdiendo el 79,7% restante.

Tabla de clasificación[a]

			Pronosticado		
			Suscripción		Corrección de porcentaje
	Observado		No ha suscrito	Sí ha suscrito	
Paso 5	Suscripción	No ha suscrito	3538	**294**	92,3
		Sí ha suscrito	400	**102**	**20,3**
	Porcentaje global				84,0

a. El valor de corte es ,200

Tabla K - 3.11 Tabla de clasificación con corte en 0,20
***Fuente**: elaboración propia*

3.3. Toma de decisiones

Ahora la entidad bancaria debe escoger cual es el público objetivo al que dirigir la campaña, para ejemplificar la situación vamos a suponer que se enfrenta a un proceso de decisión en el cual cuenta con tres escenarios: dirigir la campaña comercial a toda la cartera de clientes, hacerlo solo a quien tiene más de un 10% de probabilidades de contratar el depósito, o hacerlo solo a quien tiene más de un 20% de probabilidades de contratarlo.

	Toda la cartera	**Clientes con probabilidad mayor del 10%**	**Clientes con probabilidad mayor del 20%**
Clientes potenciales	4521	2176	396
Clientes potenciales a los que se renuncia	0	2345	4125
Casos de éxito	521	338	102
Casos de éxito a los que se renuncia	0	183	419

Tabla L - 3.12 Escenarios para la toma de decisión
***Fuente**: elaboración propia*

Para encontrar el escenario que más beneficia a la entidad bancaria habría que estimar un cálculo de los ingresos y gastos en cada uno de ellos y escoger aquel cuya diferencia sea máxima. Sería complicado realizar estos cálculos de manera exacta dada la cantidad de variables que influyen en la realidad, sin embargo, si podemos plantear un modelo teórico del beneficio que se obtendría gracias a la campaña de comunicación de este producto financiero.

Para ello realizaremos una **simplificación de la realidad** mediante los siguientes supuestos:

- El horizonte temporal es un año, y en el mismo momento que el cliente realiza el depósito la entidad bancaria invierte esa cuantía (*d*).
- Los ingresos que recibe el banco por la contratación del depósito son los intereses que obtiene por invertir el capital que el cliente deposita, es decir la TIR que obtiene de la inversión (TIR_B) multiplicado por la cuantía del depósito (*d*).
- Los costes que asume el banco son los siguientes:
    - La cuantía de intereses que paga a la persona que contrata el depósito, es decir, la TIR que obtiene el cliente (*TIRc*) multiplicado por la cuantía del depósito (*d*).
    - La campaña publicitaria, compuesto por una parte fija (*cf*) debido al diseño completo de la campaña, y otra parte variable (*cv*) que corresponde con el coste de establecer un contacto telefónico con un cliente (*i*).
    - Una vez el cliente ha contratado el depósito, el coste de tramitación de documentos, atención al cliente en oficina, entre otros, también variable (*ct*).
- El número de clientes potenciales que son escogidos para establecer un contacto lo denominaremos ***n*** y el número de clientes que finalmente se suscriben al depósito lo denominaremos ***m***.
- El interés es fijo, y se cobra/paga una sola vez.
- Solo es necesario un contacto telefónico para saber si el cliente contrata o no el depósito.

Por lo tanto, el beneficio de la operación se encuentra representado por la siguiente expresión:

$$B = \left(\sum_{i=1}^{m} (TIR_B - TIR_C) * d_i \right) - (cf + cv * n + ct * m)$$

Bastaría con aplicarla a los tres escenarios y quedarnos con aquel cuyo beneficio será superior. Estos escenarios han sido supuestos para ejemplificar cómo deben ser empleados los resultados obtenidos, pero habría que tener en cuenta que, en caso de que esta ecuación que determina del beneficio realmente fuese válida, **habría que encontrar el punto de corte exacto que arroje los valores *m* y *n* capaces de maximizar su valor, que puede ser cualquiera dependiendo de los valores reales que tomasen los demás parámetros.**

No obstante, en la realidad no es tan sencillo, ya que en el beneficio que hemos calculado también estaría influenciado por otras variables como puede ser la inflación, entre otras. Además, mientras determinamos el tamaño del público objetivo al que nos queremos dirigir, estamos suponiendo que conocemos la cuantía de los depósitos que van a firmar, y es algo imposible.

Como conclusión final extraída del análisis, el departamento de marketing debería tener en cuenta la dependencia tanto individual como conjunta de las variables que explican el éxito de la suscripción, dirigiendo el producto a aquellos consumidores que cumplan el mayor número de ellas y especialmente las más influyentes: **ser mayor de 60 años y no tener contratado ningún tipo de préstamo**. A pesar de no haberla incluido en el modelo por las razones anteriormente comentadas, no hay que olvidarse de la variable "**Resultado de la campaña anterior**", incluyendo en nuestro público a todo aquel que lo contratase en la campaña pasada, independientemente de la predicción del modelo. Otra recomendación importante que hacer sería revisar el modelo periódicamente ya que con el mero paso del tiempo se producen variaciones en la coyuntura social y económica que alteran las preferencias y comportamientos de los consumidores. **Solo aquellas compañías capaces de prever ese cambio y anticiparse a las consecuencias tendrán éxito en su futuro.**

4. Conclusiones

Con este proyecto se pretende hacer una demostración de cómo la información puede ser tratada como un activo estratégico por parte de las empresas si se sabe procesar adecuadamente, de forma que aporte valor a su actividad hasta el punto de ser la principal fuente que determina el éxito de un negocio. Esta idea se encuentra situada en un contexto actual globalizado donde las tecnologías de la información y la comunicación rigen la sociedad y la economía.

Concretamente, se ha querido hacer hincapié en la capacidad del análisis de los datos que se disponen para determinar el público objetivo al que dirigir una campaña de comunicación comercial. Para la entidad bancaria en la que se ha basado este proyecto, dicho público objetivo deberá ser aquel que cumpla el mayor número posible de las siguientes características: personas no emparejadas, con estudios superiores, no tiene contratado préstamo personal ni hipotecario y es mayor de 60 años (especialmente esta última). También deberían incluirse todos aquellos que suscribieron el depósito en la campaña de comunicación anterior.

La preocupación por definir un sector de clientes potenciales determinado radica en que la difusión de las campañas de comunicación comercial tienen un coste que solo será rentabilizado si impacta en el lugar adecuado. Esta relación entre coste del impacto y probabilidad de éxito del mismo es el determinante de la eficacia de la campaña.

Mediante el planteamiento de un modelo de minería de datos se ha llegado a la conclusión de que, si éste es adecuado, se consigue una tasa de éxito por número de impactos mayor que si la comunicación comercial directa se dirige a toda una cartera de clientes, reduciendo coste y en consecuencia aumentando beneficio.

En definitiva, la idea general es poner de manifiesto la gran utilidad de los métodos estadísticos a la hora de la toma de decisiones, siendo estas más objetivas, fiables y óptimas, eliminando el sesgo humano y dejando que la realidad hable por si sola. En la sociedad de la información y la comunicación en la que nos encontramos actualmente, los datos están ahí, solo requieren orden, estructura y un pensamiento analítico por parte de las empresas que les permita aprovecharse de ellos para alcanzar el éxito.

Y es que, a pesar de haberla incluido ya en el marco teórico de este informe, merece la pena reincidir en la frase del periodista Juan F.Cía para resumir de forma simple y concisa todo este juego: "*Todas las compañías del mundo tienen a su alcance soluciones de análisis de datos tan avanzadas y sencillas de usar que no saber lo que sucederá dentro de seis meses es pecado mortal*."

5. Bibliografía

- Bernard, M. (2016): Big Data in practise. Wiley, Cornwall.
- Dans, E. Big Data: una pequeña introducción https://www.enriquedans.com/2011/10/big-data-una-pequena-introduccion.html 19/10/2011
- Martin, H. y López, P. The World's Technological Capacity to Store, Communicate, and Compute Information http://science.sciencemag.org/content/332/6025/60.full 01/04/2011
- Instituto de ingeniería del conocimiento, Las 7 V del Big data: Características más importantes http://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/ 28/06/2016
- Santiago, J. M. (2008) Cuadernos de trabajo de la escuela universitaria estadística: factores de protección y riesgo de infidelidad en la banca comercial, Universidad Complutense de Madrid.
- Web de la consultoría Ais-Group: http://www.ais-int.com/marketing-y-ventas/
- El Big Data permitirá a la banca disponer de mejores modelos predictivos. https://www.estrategiasdeinversion.com/analisis/bolsa-y-mercados/informes/el-big-data-permitira-a-la-banca-disponer-de-mejores-n-321589 09/06/2016
- Consultoría "Ey" (2014) Big Data en el sector financiero español: Resultados de la encuesta sectorial sobre Big Data. http://www.ey.com/Publication/vwLUAssets/EY-big-data-en-el-sector-financiero-espanol/$FILE/EY-big-data-en-el-sector-financiero-espanol.pdf
- Centro de ayuda de IBM Knowledge Center: Regresión logística https://www.ibm.com/support/knowledgecenter/es/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/regression/idh_lreg.htm
- Olave, P. (1995) La información estadística: acceso y uso de bases de datos. Cuadernos aragoneses de economía, pp 5-9.
- Mashey. J (1998) *Big Data and the Next Wave of Infrastress* https://www.usenix.org/legacy/publications/library/proceedings/usenix99/invited_talks/mashey.pdf

- Cwaik J. Los dilemas éticos del Big Data http://www.telam.com.ar/notas/201701/177171-los-dilemas-eticos-del-big-data.html 19/01/2017