VUELVE A CASA

PROYECTO DE SEGMENTACIÓN Y CALIFICACIÓN DE CLIENTES PARA MEJORAR FIDELIZACIÓN EN BNP PARIBAS

JOEL GALLEGOS VALDERRABANO
SILVIA MARTIL GONZÁLEZ
JUAN MARCO PÉREZ YNOA
JOHANNY GONZÁLEZ
JAVIER ENCINAR REVILLA

Master en Business Intelligence y Big Data
Grupo 2, 2018/2019
Escuela de Organización Industrial, EOI.

Introducción

La cantidad de información que tenemos hoy en día aumenta de manera exponencial a cada segundo, de igual manera las capacidades de procesamiento y almacenamiento para gestionar datos también se han visto obligadas a aumentar. La combinación de todas estas variables ha llevado a que se desarrollen técnicas, modelos y métodos de aprendizaje automático para la gestión de grandes cantidades de datos, este fenómeno se verá impulsado en mayor medida en los próximos años.

Figura 1. Uso de las nuevas tecnologías

Big data

Cuando hablamos de Big data nos referimos a conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles.

Aunque el término “big data” es relativamente nuevo, la acción de recopilar y almacenar grandes cantidades de información para su posterior análisis se viene realizando desde hace muchos años. El concepto cobró impulso a principios de la década del 2000 cuando el analista de la industria Doug Laney articuló la definición ahora muy popular del big data como las tres Vs:

- **Volumen.** Las organizaciones recopilan datos de diversas fuentes, incluyendo transacciones comerciales, medios sociales e información de sensores o que se transmite de una máquina a otra. En el pasado, almacenarlos habría sido

un problema – pero nuevas tecnologías (como Hadoop) han aligerado la tarea.

- **Velocidad.** Los datos se transmiten a una velocidad sin precedentes y se deben distribuir de manera oportuna. Etiquetas FID, sensores y la medición inteligente crean la necesidad de distribuir torrentes de datos casi en tiempo real
- **Variedad.** Los datos vienen en toda clase de formatos – desde datos numéricos estructurados en bases de datos tradicionales hasta documentos de texto no estructurados, correo electrónico, video, audio, datos de teletipo bursátil y transacciones financieras.

En SAS, consideramos otras dos dimensiones cuando se trata del big data:

- **Variedad.** Además de las velocidades y variedades de datos cada vez mayores, los flujos de datos pueden ser muy inconsistentes con picos periódicos. Las cargas de datos máximas diarias, de temporada y desencadenadas por eventos pueden ser difíciles de controlar. Y más aún con datos no estructurados.
- **Complejidad.** Los datos de la actualidad provienen de múltiples fuentes, lo que hace difícil vincular, empatar, depurar y transformar datos entre diferentes sistemas. Sin embargo, es necesario conectar y correlacionar relaciones, jerarquías y múltiples vínculos de dato.

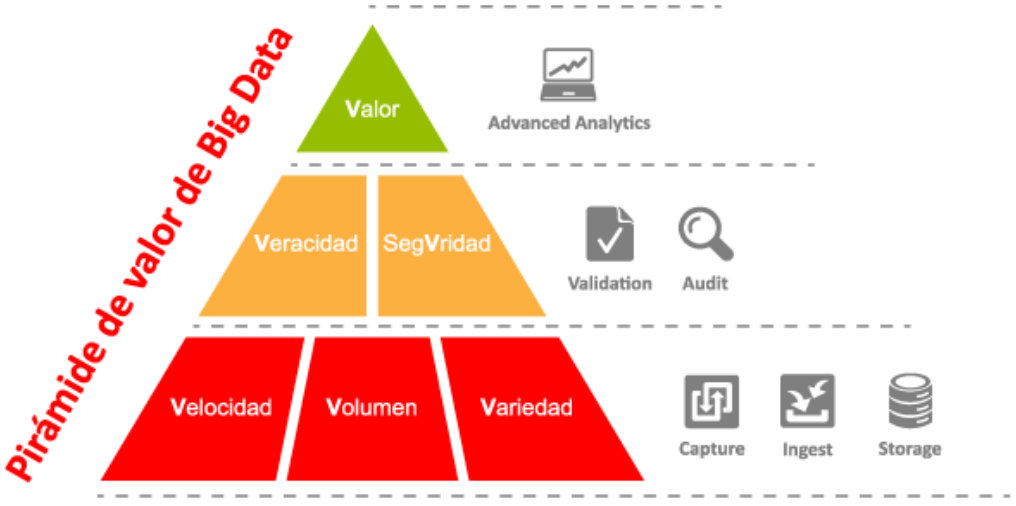


Figura 2. Pirámide de valor de Big Data

Por qué Big data es tan importante

Lo que hace que Big Data sea tan útil para muchas empresas es el hecho de que proporciona respuestas a muchas preguntas que las empresas ni siquiera sabían que tenían. En otras palabras, proporciona un punto de referencia. Con una cantidad tan grande de información, los datos pueden ser moldeados o probados de cualquier manera que la empresa considere adecuada. Al hacerlo, las organizaciones son capaces de identificar los problemas de una forma más comprensible.

La recopilación de grandes cantidades de datos y la búsqueda de tendencias dentro de los datos permiten que las empresas puedan cambiar rápidamente, adaptándose mejor y de manera más eficiente. También les permite eliminar las áreas problemáticas antes de que los problemas acaben con sus beneficios o su reputación.

El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades. Eso, a su vez, conduce a movimientos de negocios más inteligentes, operaciones más eficientes, mayores ganancias y clientes más felices. Las empresas con más éxito con Big Data consiguen valor de las siguientes formas:

- **Reducción de coste**. Las grandes tecnologías de datos, y el análisis basado en la nube, aportan importantes ventajas en términos de costes cuando se trata de almacenar grandes cantidades de datos, además de identificar maneras más eficientes de hacer negocios.
- **Más rápido, mejor toma de decisiones**. Con la velocidad y la analítica en memoria, combinada con la capacidad de analizar nuevas fuentes de datos, las empresas pueden analizar la información inmediatamente y tomar decisiones basadas en lo que han aprendido.
- **Nuevos productos y servicios**. Con la capacidad de medir las necesidades de los clientes y la satisfacción a través de análisis viene el poder de dar a los clientes lo que quieren. Con la analítica de Big Data, más empresas están creando nuevos productos para satisfacer las necesidades de los clientes.

1. Contexto del proyecto

Historia de BNP

BNP Paribas tiene su origen en los bancos de descuento creados en 1848 para afrontar la grave crisis económica y bancaria que estaba sufriendo Francia en ese momento. Concretamente en dos bancos:

- El Comptoir National d'Escompte de Paris (CNEP), que se especializó en la financiación del comercio internacional y en 1860 empezó a crear su pionera red bancaria internacional.
- El Banque Nationale pour le Commerce et l'Industrie (BNCI), que destacó por su dinámico enfoque comercial y por su carácter innovador.

En 1966 ambos bancos se fusionaron para constituir el primer banco estatal francés, el Banque Nationale de Paris (BNP), propiciando así el acceso masivo de la población francesa al sistema bancario. En 1999, BNP se hace con el control de Paribas, que tenía participaciones en un gran número de empresas y se especializó en los mercados financieros y en la financiación de infraestructuras. Actualmente, el Grupo BNP Paribas fue consolidado en el año 2000, donde ocupa una posición de liderazgo en el mercado europeo, siendo el mayor banco de la eurozona en total de activos y el segundo en la capitalización bursátil.

El Grupo tiene posiciones clave en sus tres actividades principales:

- Mercados domésticos y servicios financieros internacionales (cuyas redes de banca minorista y servicios financieros están cubiertos por Banca y servicios minoristas)
- Banca corporativa e institucional, que atiende a dos franquicias de clientes: clientes corporativos e inversores institucionales.
- Soluciones de inversión, que incluye la gestión de activos, seguros, comercio electrónico y real estate.

BNP Paribas en México

El banco tiene una larga trayectoria como multinacional, teniendo presencia en 77 países, y como fruto de esta estrategia de internacionalización, consiguió una alianza con KIA una de las marcas más importantes del país.

- KIA Motors es un fabricante surcoreano de automóviles. Su sede central está ubicada en Seúl, Corea del Sur. La compañía, perteneciente ahora al conglomerado de Hyundai Motor Group.
- En julio de 2015, KIA arranca operaciones en México, con 21 distribuidores en las 10 principales ciudades del país. Rápidamente KIA se ganó la confianza del consumidor mexicano, ya que es la única marca que ofrece garantía de 7 años o 150,000 kilómetro, así como una llamativa financiación con BNP Paribas.

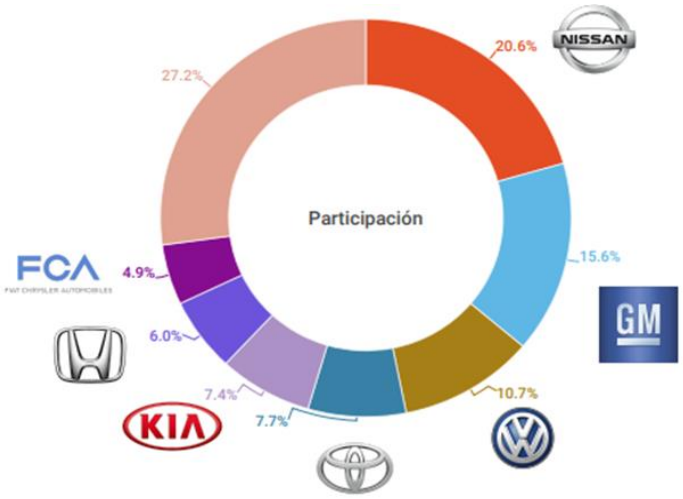


Figura 3. Participación de KIA en el mercado automotriz mexicano.

En 2017, BNP Paribas presentó una solicitud para aumentar su negocio en México y operar como institución de banca múltiple ante los reguladores financieros, por lo que la entidad ha realizado una maniobra para poder crecer en el futuro y abrirse a diferentes productos.[1]

Como toda empresa orientada al cliente, BNP Paribas Personal Finance Mexico, a partir de ahora BNP PPFM, deberá abordar tanto la captación de nuevos clientes como la fidelización de los ya existentes, así como valorar la apertura a nuevos productos.

Actualmente, en México su actividad primordial es la de ofrecer financiación para préstamos al consumo, específicamente dentro del sector automovilístico, que representa su única línea de negocio. Dicha entidad no posee ningún acuerdo de exclusividad con un concesionario específico, no obstante, su principal partner es la marca automovilística KIA, de la cual obtiene el 80% de sus clientes, que recurren a BNP para obtener su financiación.

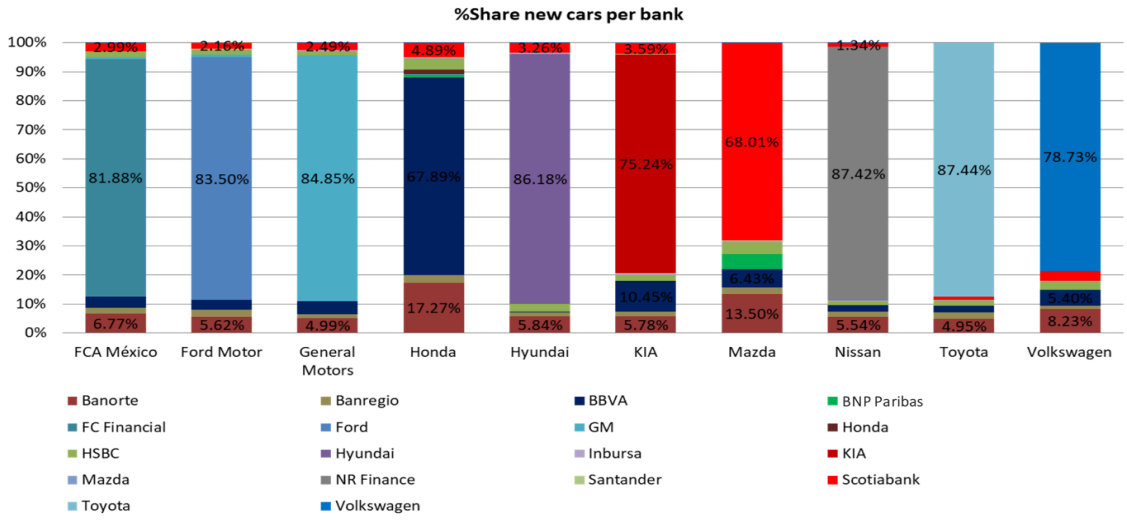


Figura 4. Participación en el mercado de préstamos para vehículos nuevos.

[1] https://expansion.mx/empresas/2018/03/13/bnp-paribas-lanzara-su-banco-en-mexico

2. Definición del problema y oportunidad de negocio

Definición del problema

En la línea de negocio automovilística los diferentes concesionarios son los que generan la oportunidad de negocio, se encargan de recoger la documentación al cliente para el estudio del préstamo y su posterior sanción favorable o desfavorable, según cada tipo de cliente. En caso de que se llegase a formalizar el préstamo, en el propio concesionario se rellena el contrato del préstamo con BNP, de manera que no hay ningún tipo de interacción del cliente con la compañía.

Añadida a esta escasa interacción en la parte comercial, hay que sumarle el hecho de que BNP, a pesar de poseer total acceso a los datos del cliente, no está explotando dicha información que conforma su principal activo.

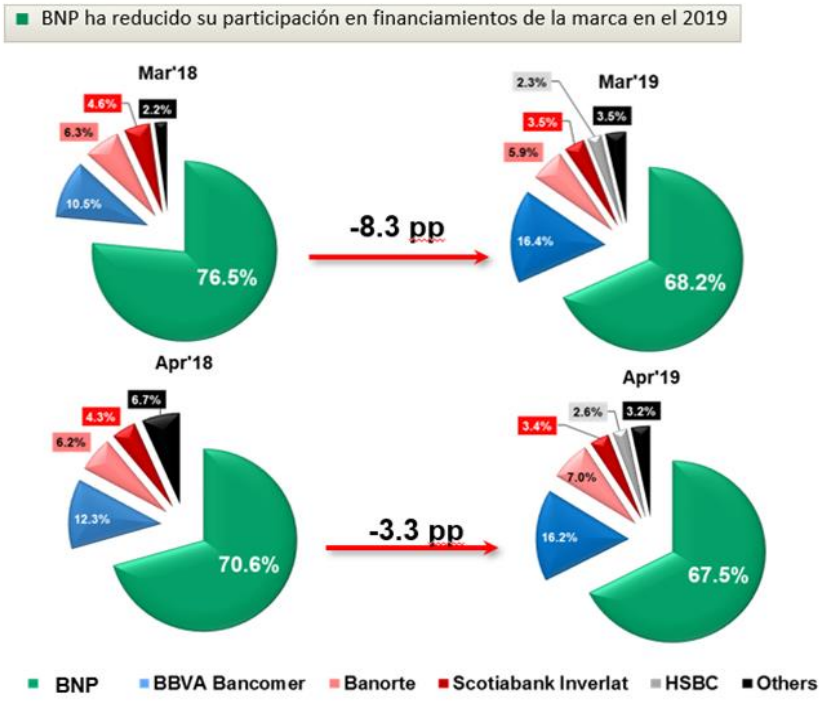


Figura 5. Participación de BNP en financiamientos de Kia, abril 2018 - abril 2019.

Actualmente BNP PPFM, es consciente de su falta de fidelización de clientes, lo que se traduce en un aumento de su tasa de abandono de clientes (tasa Churn). Los clientes recurren a la entidad para obtener financiación para la adquisición de vehículos. Desde BNP PPFM, una vez se le otorga el préstamo al cliente no se le contacta durante la vida útil del préstamo, por consiguiente, no se conoce al cliente, ni sus inquietudes o necesidades. Por tanto, no generamos permanencia en la mente del cliente, lo que conlleva a que una vez finalizada su financiación también finalice su relación con la entidad.

En un entorno tan competitivo como el de las entidades financieras, con una amplia oferta para un mismo tipo de producto, es necesario que se realice una correcta segmentación de los clientes, y así poder recompensarles mediante promociones personalizadas, descuentos o trato preferencial. La compañía no está realizando una ejecución eficiente de este recurso, quedando en posición de desventaja frente a la competencia.

Otro problema a afrontar es la dependencia de las empresas automovilísticas para la obtención de nuevos clientes, por tanto, se debería valorar la apertura a nuevos productos desvinculados del sector automovilístico, tales como préstamos al consumo para otros fines (reformas, estudios...). El préstamo al consumo es un producto que conoce bien, ya que cuenta con una política de riesgos, por lo que no conllevaría una apuesta muy arriesgada al respecto, y le permitiría ampliar su presencia en el mercado.

En relación a esta dependencia con el sector automovilístico, y tras conocer los datos de ventas de automóviles registradas en México durante este año, únicamente nos hacen reafirmarnos en la idea de BNP PPFM debe abrirse a nuevos productos. A pesar de, que están siendo meses difíciles para la industria automovilística, con un decrecimiento del 10,4 % en comparación con el año pasado, y de ser los peores datos de ventas de los últimos 4 años, la marca Kia ha sido la menos afectada, obteniendo números de ventas similares a los del año pasado, lo que demuestra la inclinación y preferencia del consumidor mexicano por la marca.

Figura 6. Diagrama Ishikawa, causas de la alta tasa de abandono de clientes.

Oportunidad de negocio

Como consecuencia de los continuos y progresivos avances tecnológicos en el almacenamiento y en el procesamiento de datos, surge la oportunidad de realizar un proyecto de análisis del cliente que nos permitirá segmentar y aislar del conjunto de clientes únicamente a aquellos con mejores datos o que posean el perfil que BNP PPFM está buscando para su negocio.

A través de la segmentación pueden surgir nuevas oportunidades de negocio. En el caso, que nos ocupa, existe una tipología de productos derivados del sector automovilístico, tales como, el renting o leasing de vehículos que se podrían ofertar a clientes autónomos o empresas. Y que actualmente, no se contempla por desconocimiento del cliente.

Esto permitiría al banco, alineados con la estrategia de expansión en México, disponer de una base de datos que permita ofrecer nuevos servicios y fidelizar de esta forma al cliente.

Para conseguir un correcto estudio y gestión de los datos se requerirá de infraestructuras, tecnologías y servicios especializados, que nos permita convertir todos esos datos en información útil para la toma de decisiones, obteniendo un beneficio para la empresa.

3. Planteamiento de la solución, beneficios esperados e hipótesis

Actualmente no existe un modelo o proceso en el negocio que permita mejorar la relación con los clientes existentes, por lo que se realizara un modelo de scoring que nos permita segmentar a aquellos clientes que deseamos conservar y sobre los que se emplearán las estrategias marcadas por la entidad. Se buscará segmentar aquellos clientes que sean los mejores, lo que nos permitirá reducir el esfuerzo del equipo de ventas y a su vez mitigar los posibles impagos originados por los préstamos.
Los beneficios que esperamos obtener son:

- Obtener una Base de Datos de clientes depurada con aquellos clientes que sean de vital importancia para la entidad.
- Identificar principales estrategias de marketing a llevar a cabo para fidelizar a los clientes de la compañía.
- Propuestas de mejora y eficiencia para atraer a nuevos clientes.
- Establecer un modelo de control y seguimiento del cliente, para poder anticiparnos a sus necesidades.

La hipótesis que planteamos es que los clientes mejores son los que:

- Tengan o hayan tenido préstamos con la entidad y siempre hayan estado al corriente de pago durante su financiación.
- Posean un mayor capital inicial en sus préstamos o una aportación inicial elevada, ya que se le otorgará un mayor endeudamiento a aquellos clientes que posean más recursos, y aquellos que pueden realizar un mayor desembolso inicial en la compra del vehículo ya que denotará que el cliente tiene una economía saneada para poder realizar dicho desembolso inicial.
- Posean mayores ingresos derivados de su actividad profesional.
- Tienen interés en un segundo vehículo después de terminar de pagar el primero o cuando están a punto de liquidarlo.

Las hipótesis de problema que se plantean en el estudio:

- Si se logra crear un modelo de negocio dirigido a los clientes existentes las utilidades de la empresa incrementaran, así como las relaciones comerciales con KIA.
- El mercado de automóviles en México es bastante competido por lo que la fidelización correcta de los clientes posiciona a las marcas manufactureras de autos como a las socias financieras.
- Actualmente BNP es la que tiene un mayor financiamiento de autos KIA, pero la competencia ha incrementado su participación derivado de falta de estrategias para la retención de clientes, por lo que un modelo que nos ayuda a segmentar, identificar e incrementar el número de financiamientos por medio de los clientes existentes, asegurara a BNP Paribas el primer lugar y pudiese generar nuevos socios comerciales.

Por último, tenemos las hipótesis de solución las cuales es la manera en la que se resolverá el problema considerando impactos y riesgos adheridos a las mismas

.

- La primera hipótesis es sobre los datos, considerar un proceso ETL exitoso que nos ayude a brindar datos de calidad al modelo.
- El mejor modelo, al ser un proyecto finalmente de negocio se necesita que sea simple y permita ser automatizado para cada una de las campañas que se necesiten sobre los clientes existentes.
- La selección de clientes por medio del modelo y criterios de negocio, donde la estadística y el negocio se fusionan para seleccionar a los mejores clientes y mostrarles la mejor opción cara a sus necesidades.
- Por último, el seguimiento del modelo y los resultados para cada campaña mejorar y lograr el mejor modelo de negocio para clientes existentes.

Cliente – Problema – Solución

Tras realizar diferentes entrevistas con las áreas interesadas del cliente, se han identificado áreas de actividad en las que será necesario actuar para conseguir los objetivos buscados. Gracias a estas entrevistas logramos identificar a los early adopters del proyecto, todas las áreas interesadas en que el proyecto triunfe de la mejor manera.

En estas entrevistas se identificaron varios perfiles, de negocio, riesgo, operativos, etc., los cuales con su experiencia brindan diferentes enfoques de cómo abordar el problema.

Las entrevistas realizadas a los perfiles de negocio identifican a BNP Paribas con un rezago en cuestión de modelo de negocio para clientes existentes, mostrando interés en la anticipación a las necesidades de los mismos, es decir no esperar hasta que el cliente liquide su automóvil para ofrecerle otro, si no brindar opciones anticipadas al cliente para lograr que en la etapa de decisión de su nuevo vehículo nos encontremos presentes, estos serían los principales early adopters ya que son los más interesados en incrementar el número de financiamientos en la empresa.

Sobre las entrevistas realizadas al área de riesgos, identificamos problemas de bases de datos para poder construir el análisis, llaves de identificación de clientes, seguimiento interno de los mismos, y algo muy importante como determinar a un buen cliente, si es posible permitir que haya tenido algún retraso en su tiempo de vida, el monto promedio a financiar, el apetito de riesgo de la institución, así como los diferentes modelos que se pueden utilizar para poder identificar a los mejores clientes (score, clusters, árboles de decisión, etc.).

Las entrevistas realizadas a los equipos operativos, analistas de documentos y proceso de financiación, personas de tecnologías de la información, personal administrativo, etc., muestran el panorama actual del tratamiento a clientes existentes, la falta de ofertas, procesos diferenciados y baja contractibilidad de los clientes, así como los sistemas utilizados para poder afrontar las necesidades y mejorar la contactabilidad de los clientes.

Derivado de las diferentes entrevistas concluimos la necesidad de un modelo de segmentación de clientes, y el interés mostrado por la empresa para poner en práctica el modelo de negocio.

Análisis preliminar de datos

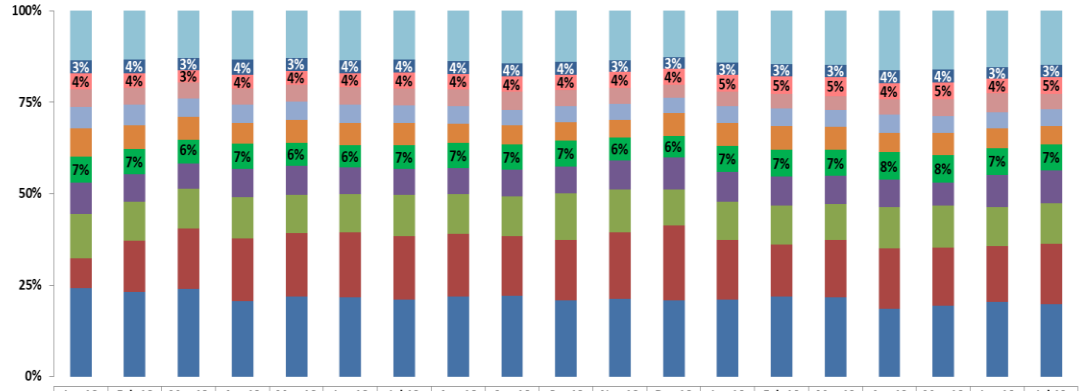
Se cuentan con diferentes bases de datos que permitirían el realizar una correcta segmentación de clientes:

- Base de originación de créditos: en ella se encuentra toda la información al momento del financiamiento del vehículo, datos como pago inicial, endeudamiento, información de buró de crédito, datos demográficos, etc., estas bases de datos no se actualizan conforme la vida del crédito avanza por lo que solo son una fotografía del cliente al momento de la financiación.
- Bases de comportamiento crediticio: en ellas se encuentra el histórico de pagos de cada uno de los clientes, la información se encuentra desde el primer pago del cliente después de la financiación hasta la liquidación del crédito de manera mensual, incluyendo saldos, pagos, días de impago, etc.

4. Análisis de entorno y análisis competitivo

Como ya lo mencionamos BNP Paribas es el principal socio comercial de la marca KIA, por lo que entender cuál es el desempeño de las ventas de manera histórica y su trascendencia en el mercado automotriz mexicano ayudara a entender de mejor manera el problema.

La siguiente grafica muestra el desempeño de la marca KIA sobre las ventas totales de automóviles en México, así como los principales competidores y su desempeño.



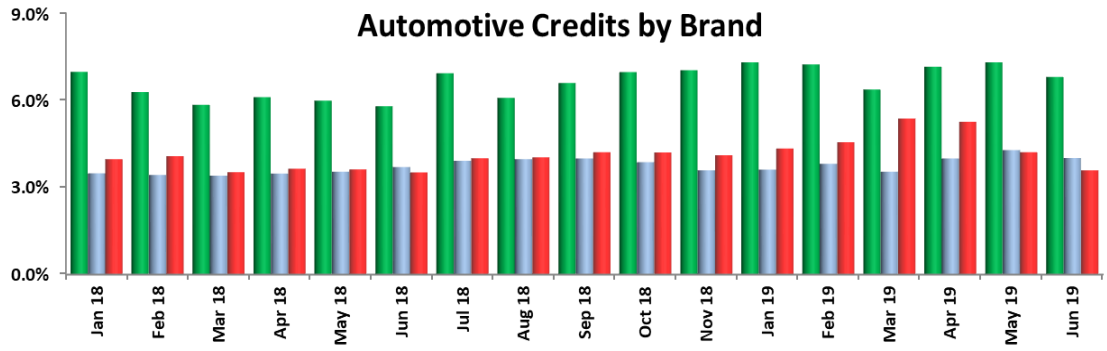
	Jan 18	Feb 18	Mar 18	Apr 18	May 18	Jun 18	Jul 18	Aug 18	Sep 18	Oct 18	Nov 18	Dec 18	Jan 19	Feb 19	Mar 19	Apr 19	May 19	Jun 19	Jul 19
Others	14%	13%	13%	13%	13%	13%	13%	14%	14%	14%	13%	13%	14%	15%	15%	16%	16%	15%	15%
Hyundai	3%	4%	3%	4%	3%	4%	4%	4%	4%	4%	3%	3%	3%	3%	3%	4%	4%	3%	3%
Mazda	4%	4%	3%	4%	4%	4%	4%	4%	4%	4%	4%	4%	5%	5%	5%	4%	5%	4%	5%
Chrysler	5%	5%	4%	4%	5%	5%	5%	5%	5%	4%	4%	4%	4%	4%	4%	4%	5%	5%	4%
Ford	6%	6%	5%	5%	5%	5%	5%	5%	4%	5%	4%	4%	5%	5%	5%	5%	5%	4%	5%
Honda	8%	6%	6%	6%	6%	6%	6%	5%	5%	5%	5%	6%	6%	6%	6%	5%	6%	5%	5%
KIA	7%	7%	6%	7%	6%	6%	7%	7%	7%	7%	6%	6%	7%	7%	7%	8%	8%	7%	7%
Toyota	8%	7%	7%	8%	8%	7%	7%	7%	7%	7%	8%	9%	8%	8%	8%	7%	6%	9%	9%
Volkswagen	12%	11%	11%	11%	10%	10%	11%	11%	11%	13%	12%	10%	10%	11%	10%	11%	11%	11%	11%
General Motors	8%	14%	17%	17%	17%	18%	17%	17%	16%	17%	18%	21%	16%	14%	16%	16%	16%	15%	16%
Nissan	24%	23%	24%	21%	22%	22%	21%	22%	22%	21%	21%	21%	21%	22%	22%	19%	19%	20%	20%

Figura 7. Participación de las marcas en el total de ventas de automóviles, enero 2018 - julio 2019.

KIA participa en promedio con un 7% de todas las ventas de autos en México, situada en el quinto lugar, lo cual es muy importante ya que tiene solo 4 años en México, comparado con marcas como Toyota o Volkswagen que tienen más de 10 años en el país.

Con esto podemos entender la importancia del crecimiento de las ventas de la marca y lo que importa esto para el modelo.

Pero esta grafica solamente nos muestra las ventas totales, incluyendo las de crédito y contado, la siguiente muestra la participación de 3 marcas en México sobre el total de ventas de automóviles a crédito.

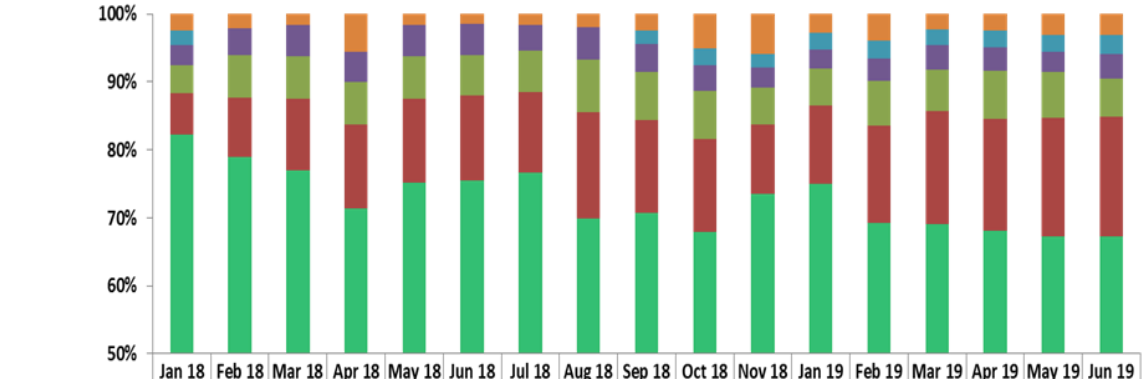


	Jan 18	Feb 18	Mar 18	Apr 18	May 18	Jun 18	Jul 18	Aug 18	Sep 18	Oct 18	Nov 18	Jan 19	Feb 19	Mar 19	Apr 19	May 19	Jun 19
KIA	7.0%	6.3%	5.9%	6.1%	6.0%	5.8%	6.9%	6.1%	6.6%	7.0%	7.0%	7.3%	7.2%	6.4%	7.2%	7.3%	6.8%
Hyundai	3.5%	3.4%	3.4%	3.5%	3.5%	3.7%	3.9%	4.0%	4.0%	3.9%	3.6%	3.6%	3.8%	3.5%	4.0%	4.3%	4.0%
Mazda	4.0%	4.1%	3.5%	3.6%	3.6%	3.5%	4.0%	4.0%	4.2%	4.2%	4.1%	4.3%	4.6%	5.4%	5.3%	4.2%	3.6%

Figura 8. Participación de las marcas en las ventas de automóviles a crédito en México, enero 2018 - julio 2019.

De la misma forma KIA participa con alrededor del 7% sobre las ventas a crédito, colocándola en el número 5, las marcas Hyundai y Mazda son los principales competidores en segmentos de auto y precio de los mismos, con un porcentaje mucho menor.

Estas 2 gráficas nos ayudan a entender cómo funciona el primer socio comercial de BNP Paribas en México, pero nos falta entender a los competidores sobre este socio.



	Jan 18	Feb 18	Mar 18	Apr 18	May 18	Jun 18	Jul 18	Aug 18	Sep 18	Oct 18	Nov 18	Jan 19	Feb 19	Mar 19	Apr 19	May 19	Jun 19
Others	2%	2%	2%	6%	2%	1%	2%	2%	2%	5%	6%	3%	4%	2%	2%	3%	3%
HSBC	2%	0%	0%	0%	0%	0%	0%	0%	2%	2%	2%	3%	3%	2%	3%	2%	3%
Scotiabank	3%	4%	5%	4%	5%	5%	4%	5%	4%	4%	3%	3%	3%	4%	3%	3%	4%
Banorte	4%	6%	6%	6%	6%	6%	6%	8%	7%	7%	6%	5%	7%	6%	7%	7%	6%
BBVA Bancomer	6%	9%	11%	12%	12%	12%	12%	16%	14%	14%	10%	12%	14%	17%	16%	17%	18%
BNP	82%	79%	77%	71%	75%	75%	77%	70%	71%	68%	74%	75%	69%	69%	68%	67%	67%

Figura 9. Participación de BNP Paribas en los créditos de KIA en México, enero 2018 - julio 2019.

Podemos observar que BNP tiene más del 65% de los créditos de KIA, pero con una caída desde el mes de febrero del 2019, principalmente absorbida por BBVA, la segmentación de clientes podrá ayudar a incrementar esta diferencia y regresar a los niveles de 75% que se tenían en los meses de mayo y junio del 2018 o incluso incrementarlos.

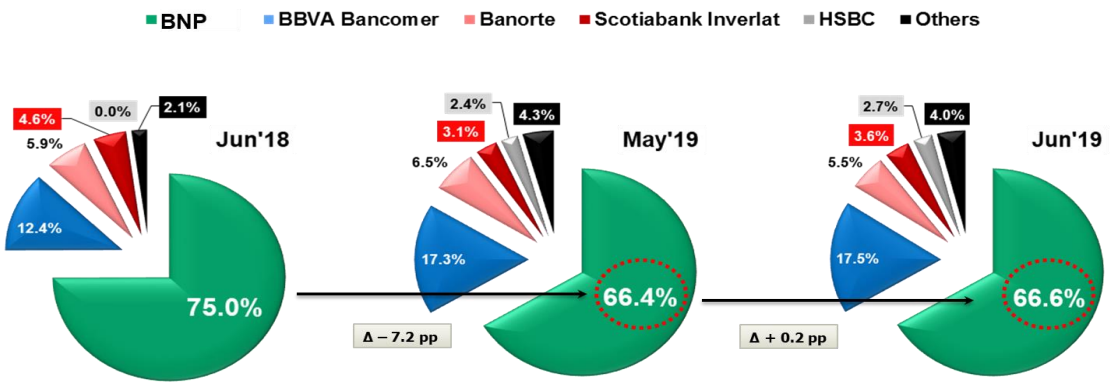


Figura 10. Cuota de participación de BNP en financiamientos de Kia, abril 2018 - abril 2019

Adicional existe una caída en el 2019 comparado con el mismo en el 2018, esta caída es de casi 10 puntos porcentuales lo que hace que BNP necesite focalizar sus esfuerzos hacia los clientes existentes.

Análisis del sector de la consultoría

México tenía uno de los mercados más desarrollados de América Latina, sólo por detrás de Brasil, según el estudio "Perspectivas del mercado geográfico de consultoría 2014: América Latina" realizado por Kennedy Consulting Research & Advisory. En este estudio se marca a Deloitte como el líder del mercado, teniendo una cuota de mercado del 11,1% y una tasa de crecimiento del 19,1%[2].

Según el análisis de mercado del ICEX[3], el mercado de consultoría principalmente está ocupado por empresas internacionales, que se caracterizan por tener equipos multidisciplinares. Las empresas que destacan en el sector mexicano son las siguientes:

Principales empresas de consultoría internacionales con presencia en México
ATKearney
Bain & Company
Bank of America - Merrill Lynch
Barlovento Consultores
Boston Consulting Group
Deloitte
EY
KPMG
ManagementSolutions
Morgan Stanley
Oliver Wyman
PwC
Sintec

Figura 11. El mercado de la consultoría e ingeniería en México. ICEX 2019

[2] https://www.eleconomista.com.mx/empresas/Mexico-mercado-de-consultoria-20150408-0068.html

[3] El mercado de la consultoría e ingeniería en México. ICEX 2019

En el sector financiero, las empresas que destacan son las Big Four y Management Solutions, que son las que identificamos como posibles competidoras. Por lo comentado en las entrevistas con la empresa, nos identifican a Deloitte y KPMG como los líderes del mercado.

En cuanto a la demanda, el ICEX identifica que se está generando una cultura de contratación de consultoría. Esto sumado al cambio tecnológico, nos genera una oportunidad de entrada en el mercado.

Por qué nosotros

Las razones de que BNP debería hacer el proyecto con nosotros son las siguientes:

- Somos un equipo multidisciplinar, que engloba experiencia acreditada de varios años tanto en conocimientos técnicos como del sector financiero. No sería un equipo de personas sin experiencia.
- Tenemos conocimiento de la empresa y de sus necesidades, ya que un integrante del equipo trabaja allí.
- Somos de diferentes países, por lo que podemos aportar conocimiento sobre las “best practices” en proyectos tecnológicos y puntos de vista diferentes.
- Tenemos experiencia en trabajar con una orientación a resultados en un determinado plazo.
- Al tener una estructura más pequeña y menos piramidal, nos permite dar un servicio a mejor precio que una Big Four.

5. Análisis y Diagnóstico / Plan estratégico-acción

La situación actual de la compañía, nos ha llevado a desarrollar una solución que les ayude a retener la cartera de clientes existentes, al mismo tiempo que les permita impulsar la captación de nuevos clientes u oportunidades de negocio. Esta situación se debe atajar, tanto desde el punto de vista comercial (incremento de ventas), operativo (rapidez y disminución de tiempos de respuesta), y de riesgos (incentivar la mejora en la calidad del portafolio potenciando la llegada de buenos clientes). Estos 3 ejes fundamentales en los que debe apoyarse la organización, le permitirán mejorar la experiencia del cliente, fidelizándolos y consiguiendo de este modo que acudan a la compañía de manera recurrente (cuando vuelvan a necesitar financiación), y atrayendo nuevos clientes (por referencias de los usuarios).

Para profundizar y analizar aquellos puntos a tratar en el proyecto, procederemos a realizar un análisis DAFO.

Análisis DAFO

Para poder realizar un diagnóstico completo de la situación de la empresa, recurrimos al denominado Análisis DAFO, que es una herramienta de estudio fundamental para identificar aquellos factores críticos que nos ayudarán a conocer la situación actual en la que se encuentra la compañía y, a partir de la que se trazará la futura estrategia empresarial. Este estudio analiza aquellos elementos de la organización, Oportunidades y Amenazas, que afectan por igual a todas las empresas del sector y cuya identificación proviene del análisis externo realizado. Mientras que las Fortalezas y Debilidades, son aquellos elementos específicos de la empresa cuya identificación se extrae del análisis interno llevado a cabo.

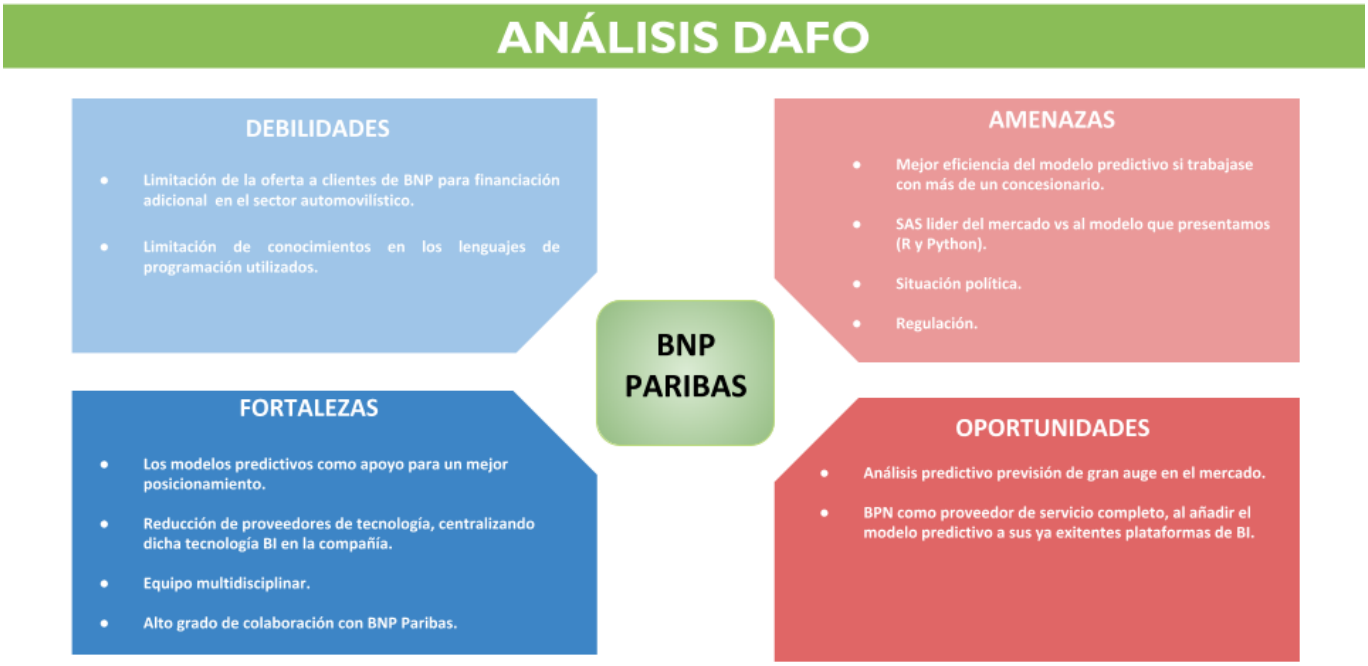


Figura 12. Análisis DAFO. Elaboración propia.

Fortalezas:

Atributos y características propias de la empresa que la posicionan en una situación de ventaja en relación a sus competidores.

- Los modelos predictivos llenan un vacío importante en las capacidades generales de inteligencia de negocios de la compañía y le permite posicionarse como un proveedor de un espectro completo de tecnología de BI. Estos modelos se basan en tecnología probada y comercialmente exitosa.
- Los bancos aumentarán el uso de análisis predictivos en el futuro, en su esfuerzo por identificar comportamientos sospechosos o de prevención del fraude, así como la mejora de la experiencia del cliente, entre otros.
- Muchas organizaciones están tratando de reducir la cantidad de proveedores con los que tratan y esto permite a la empresa una fortaleza única, ya que son ellos mismos los que poseen y desarrollan de manera centralizada la tecnología avanzada de BI.
- Equipo multidisciplinar, los miembros del equipo cuentan con diversos perfiles, aportando experiencia en áreas de IT y negocio lo que enriquece el desarrollo del proyecto.
- Alto grado de implicación por parte de BNP Paribas en el desarrollo del proyecto, para facilitar el uso de los datos como para mantener una comunicación directa, flexible y activa.

Debilidades:

Elementos internos que constituyen barreras u obstáculos para la consecución de los objetivos marcados.

- El modelo propuesto busca clientes de BNP que quieran obtener financiación adicional, pero solo en el sector de la automoción, lo que limita la oferta de otro tipo de productos.
- Como equipo implantador de la solución, con experiencia limitada en los lenguajes de programación utilizados para el desarrollo del proyecto, Python o R. Este punto puede afectar al resultado final del modelo.

Oportunidades:

Factores positivos que hay en el entorno, y que una vez identificados pueden ser utilizados a nuestro favor.

- Aunque muchas organizaciones consideran el análisis predictivo como una potente arma competitiva continúan siendo reacios a discutir su uso, y sin embargo éste puede ser uno de los segmentos que obtenga un mayor crecimiento en el mercado de BI. Solo estamos viendo la punta del iceberg en relación con las oportunidades generales del mercado.
- Al ofrecer un modelo predictivo a BNP como complemento de su plataforma de BI, puede posicionarse como un proveedor de servicio completo que puede proporcionar a los clientes potenciales tecnología de inteligencia empresarial central y avanzada.

Amenazas:

Situación externa y negativa, que puede perjudicar o atentar contra los intereses de la compañía.

- Como únicamente se trabaja con Kia para la obtención de financiación en la venta de sus vehículos, el modelo predictivo será eficiente, aunque podría serlo aún más si se trabajase con otros concesionarios.
- SAS, es el líder en el mercado de análisis avanzado, actualmente BNP lo utiliza, aunque esperamos ofrecer un modelo eficiente y hecho a medida que supere los modelos de SAS.
- La situación política puede añadir incertidumbre constituyendo importantes obstáculos de carácter ajeno al desarrollo del proyecto.
- Un cambio en la regulación de normativas y reglas dirigidas a las entidades financieras puede afectar negativamente a la entidad.

Modelo de Negocio

El modelo de negocio que hemos desarrollado para BNP Paribas se centra en la realización de un modelo de scoring que les permita segmentar adecuadamente a sus clientes. Con el firme propósito de ayudarles a conseguir este objetivo ha surgido nuestro proyecto.

Para ello, definiremos el modelo de negocio de la empresa utilizando el Modelo CANVAS que nos permitirá definir las estrategias claves a seguir. El modelo se divide en 9 secciones básicas que reflejan la lógica que sigue la empresa para generar los ingresos que cubrirán las principales áreas del negocio: clientes, oferta, infraestructura y viabilidad económica del modelo.

En primer lugar, se deberán conocer y analizar los aspectos externos de la empresa, es decir, su entorno. Que hace referencia a los módulos que aparecen a la derecha, tales como: segmento de clientes, propuesta de valor, canales, relación con clientes y fuentes de ingresos.

Una vez conozcamos el entorno de la empresa, deberemos adaptar los aspectos internos, localizados en la parte izquierda del modelo, para que la propuesta de valor se realice de la mejor manera posible, mediante la creación de asociaciones con los agentes necesarios, centrándonos en las actividades clave de la empresa e identificando los recursos que serán necesarios para ello, bajo una adecuada estructura de costes.

MODELO DE NEGOCIO

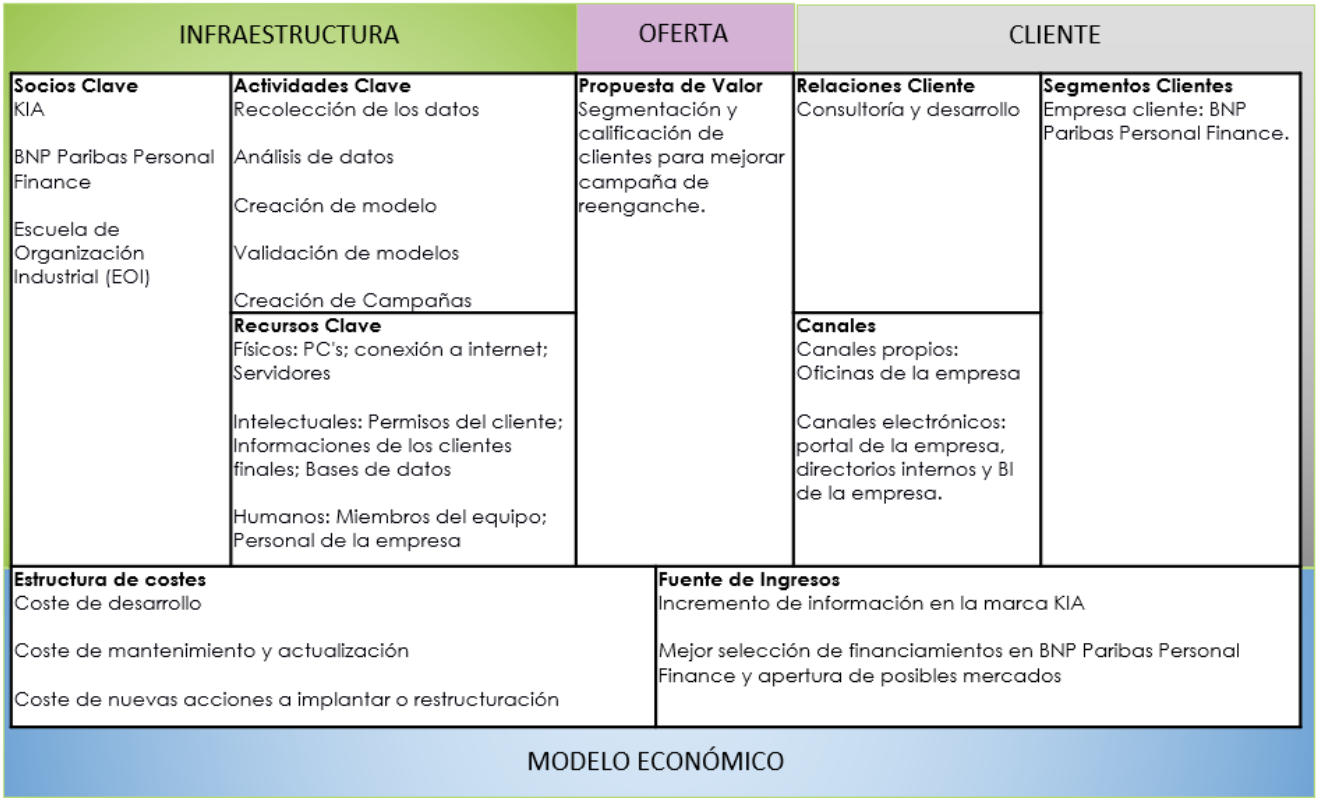


Figura 13. Business Model Canvas. Elaboración propia.

Propuesta de valor

La propuesta de valor que ofrecemos para BNP Paribas nace de la necesidad por parte del grupo de tratar de disminuir su tasa Churn. Actualmente, no se realiza ninguna acción de fidelización de clientes, por lo que una vez finalizada su financiación, el cliente también da por finalizada la relación que tiene con la entidad. Que el cliente vuelva o no será un hecho meramente aleatorio, puesto que al no realizarse acciones de fidelización, el cliente puede percibir que se le trata de manera indiferente y no tendrá motivos para repetir su experiencia con la entidad.

Por ese motivo, la propuesta de valor que realizamos se basa en la elaboración de una base de datos de clientes depurada, que junto con la realización de un modelo de scoring nos permita extraer de entre todos los clientes únicamente a aquellos que BNP Paribas desea mantener, y sobre los que realizará acciones comerciales y de marketing para cada segmento específico. Adicionalmente, de esta segmentación se pueden generar nuevas oportunidades de negocio, tales como préstamos personales para otras finalidades distintas a la adquisición de un vehículo, renting, leasing, etc. Lo que se traducirá en una menor dependencia del sector automovilístico, pudiendo incrementarse su volumen de préstamos y/o clientes.

Segmento de clientes

El proyecto va dirigido exclusivamente a nuestro cliente BNP Paribas con los que hemos trabajado en la extracción y limpieza de datos para buscar una solución focalizada en satisfacer sus necesidades. A través del procesamiento y análisis de sus datos serán capaces de obtener información relevante que les permita reorientar su estrategia empresarial, segmentando a sus clientes adecuadamente mediante un modelo de scoring que les ayude a tomar mejores de decisiones en el futuro y puedan reducir así, su tasa de abandono de clientes.

Canales de distribución

Los canales de distribución empleados con el grupo BNP Paribas, debido a que uno de los componentes del grupo del proyecto trabaja en la entidad serán aquellos que permitan una comunicación lo más fluida posible mediante la utilización de los siguientes canales de distribución: conference call, vía e-mail y reuniones presenciales con las áreas implicadas e interesadas en la realización y el éxito del proyecto.
También se planteará tener un directorio compartido dentro de la empresa, así como una parte en las aplicaciones de BI de la empresa.

Relaciones con clientes

La entidad BNP Paribas ha colaborado activamente con el proyecto, por lo que la comunicación ha sido regular y muy fluida. Esto nos ha permitido estar en consonancia y realizar una solución adaptada a las necesidades del cliente.

Adicionalmente, como uno de los integrantes del proyecto trabaja en la entidad, éste ofrece a BNP Paribas una relación de asistencia personal dedicada (KAM). Para poder entender correctamente este término, procederemos a realizar una breve definición de dicha figura, la cual, asigna a un responsable a la atención específica de un cliente (BNP Paribas). Se trata de una relación más íntima y profunda con el cliente y suele prolongarse durante un largo período de tiempo.

Socios Clave

En cuanto a los socios clave del proyecto podemos destacar a 3 grupos de socios:

- BNP Paribas: Entidad financiera en México que otorga préstamos a clientes físicos o empresas para la financiación de los vehículos de la marca comercial KIA, y del cual obtenemos la fuente de datos con los que poder trabajar.

- KIA: marca dedicada a la venta de vehículos con auge en México y proveedor de clientes a BNP, puede beneficiarse del tratamiento en los datos de los clientes que han financiados sus vehículos con BNP Paribas, ya que podrán hacerse con una base de datos con datos depurados y útiles para la realización de sus campañas comerciales.

- EOI: fundamental en el apoyo tanto técnico como de negocio brindado durante la vida del proyecto, para ampliarnos conocimientos en el uso de las herramientas de BI y ayudarnos a tener una visión más amplia del proyecto.

Actividades Clave

Nuestra propuesta de actividades clave es el tratamiento y análisis de datos que, en combinación con el uso de la tecnología y la analítica avanzada, nos permite obtener información relevante para la toma de decisiones, la segmentación de clientes y su posterior fidelización.

Recursos clave

Los recursos clave que identificamos en el proyecto son:

- La información sin tratar del cliente, de cuya calidad en la información depende el servicio que se va a prestar.

- Es el capital humano formado por los miembros del equipo del proyecto, que deberán poseer un alto conocimiento de programación y analítica avanzada junto con el uso de herramientas de BI y Big Data, para que el proyecto pueda albergar los mejores resultados.

- Recursos tecnológicos serán necesarios ordenadores con conexión a internet, servidores para almacenar los datos facilitados por la empresa.

Fuentes de ingresos

Más que una fuente de ingresos, va a ser un enriquecimiento de la información actual que se tiene del cliente. Esto va a generar los siguientes beneficios:

- Posibilidad de abrirse a nuevos productos, ya que se tendrá una base de potenciales clientes con los que trabajar.
- Análisis y seguimiento del tipo de clientes.
- Focalización del esfuerzo de fidelización en clientes, ahorrando tiempo en fidelizar a quien no interesa.
- Aportación de información valiosa a KIA.

Estructura de costes

La estructura de costes necesaria para el desarrollo del proyecto estará compuesta por los siguientes costes:

- Infraestructura: equipos informáticos con conexión a internet, servidores para almacenar los datos facilitados por la empresa, etc.

- Actualización y almacenamientos de los datos.

Reestructuración del modelo y de las estrategias llevadas a cabo.

6. Plan de Acción

Primero revisaremos como se encuentra en la compañía el proceso para los clientes existentes, así como la participación dentro del total de financiados.

La metodología actual reconoce a un cliente existente por su RFC (registro federal de contribuyentes), clave única por cliente.

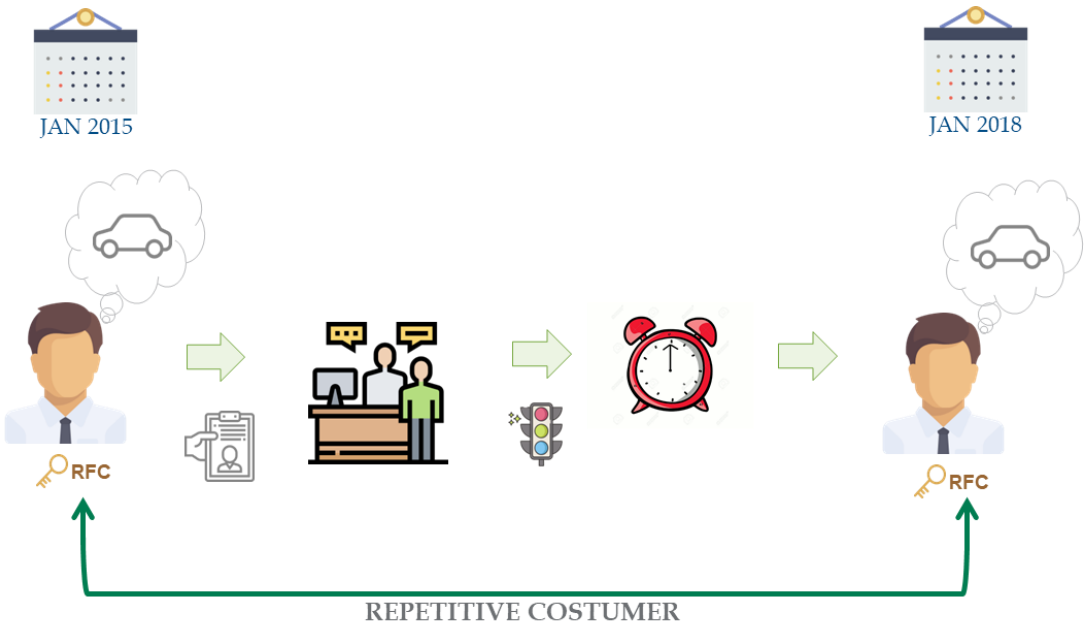


Figura 14. Ciclo de clientes recurrentes. Elaboración propia.

Como el diagrama muestra un cliente viene por primera vez, realiza una solicitud se aprueba y en meses posteriores a su regreso vuelve a pedir un segundo crédito.

El proceso se divide en tres, las solicitudes de estos clientes, la decisión de aprobar o rechazar a estos clientes y los créditos financiados:

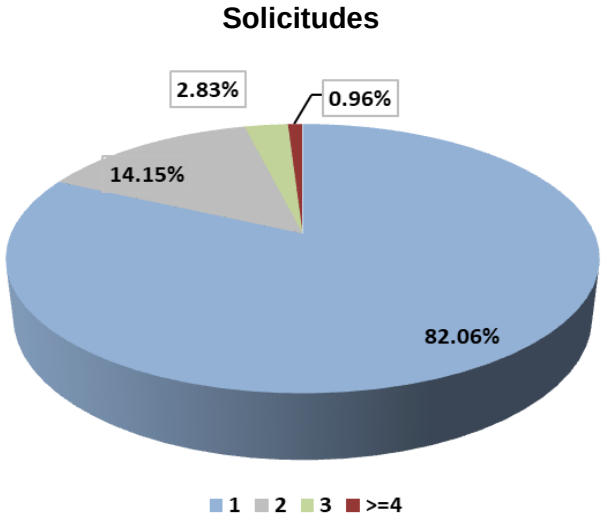


Figura 15. Proceso de tramitación de financiación. Elaboración propia.

El 18% de todos los clientes que solicitan un crédito ya lo había solicitado previamente, esto es un porcentaje importante de la demanda total de clientes, pero también es importante conocer el tiempo en el que estos clientes regresan a solicitar un nuevo crédito.

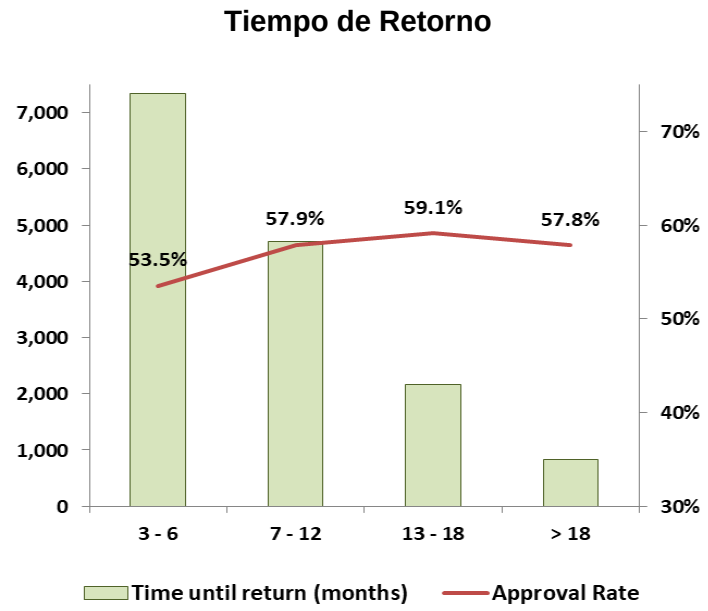


Figura 16. Tiempo de retorno de inversión. Elaboración propia.

La siguiente parte corresponde a la decisión que arroja el sistema de aprobar o rechazar a estos clientes:

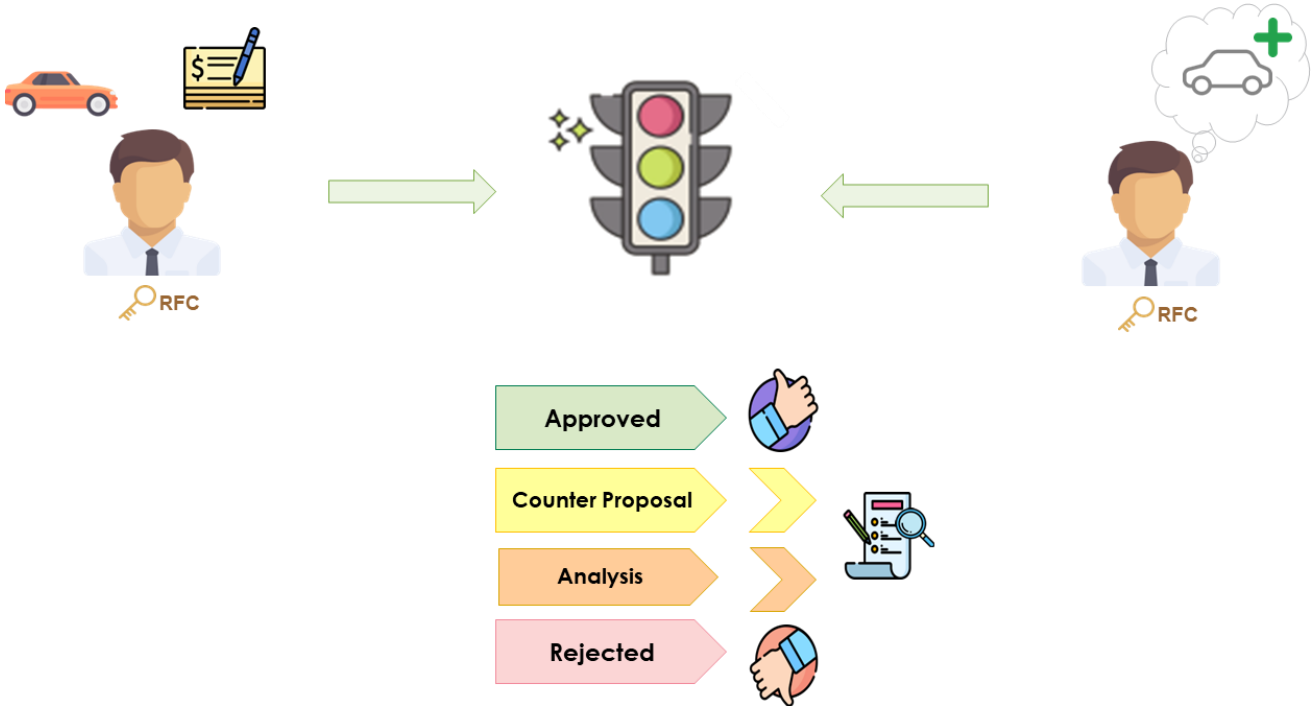


Figura 17. Fases de la sanción de financiación. Elaboración propia.

El motor de decisión de la empresa arroja un decreto para estos clientes:

	Approved	Counter Proposal	Orange	Red	Total
Approved	2.0%	0.3%	57.1%	13.8%	73.2%
Counter Proposal	0.1%	0.0%	0.5%	0.3%	0.8%
Orange	0.6%	0.1%	18.2%	5.1%	24.0%
Red	0.2%	0.0%	0.8%	1.0%	2.0%
Total	2.8%	0.4%	76.6%	20.1%	100%

Figura 18. Cuadro ponderado de las fases de proceso de sanción de la financiación. Elaboración propia.

El 57% de los clientes reciben una decisión de análisis (Orange), es decir no se decide de manera automática la aprobación o rechazo de los mismos, haciendo el proceso de financiación más lento.

Por último, la parte de financiación de los créditos:

Financiación

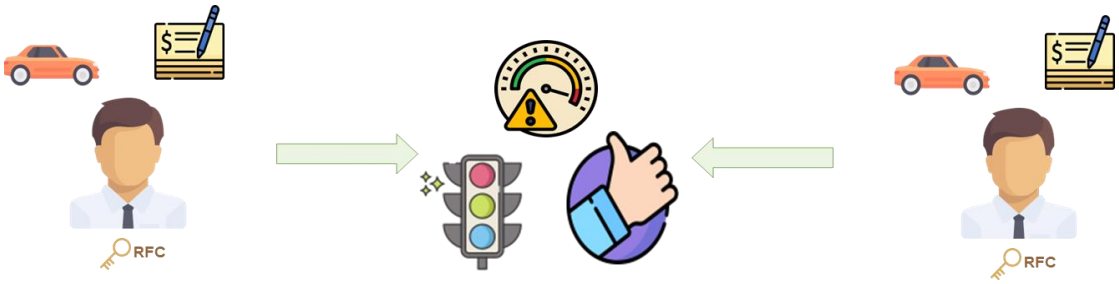


Figura 19. Proceso de financiación. Elaboración propia.

Al ser un proceso más lento y con problemas en el análisis esto aunado a lo competitivo que es el mercado automotriz en México la participación en el portafolio de clientes existentes se reduce considerablemente.

Financiaciones

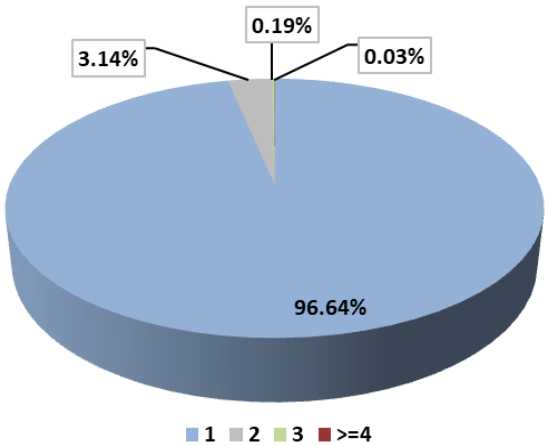


Figura 20. Porcentaje de clientes financiados. Elaboración propia.

El 4% del total del portafolio tiene o ha tenido más de 1 crédito financiado.

Como parte del plan de acción se definen los siguientes objetivos que permitirán centrar el proyecto en las necesidades de la compañía.

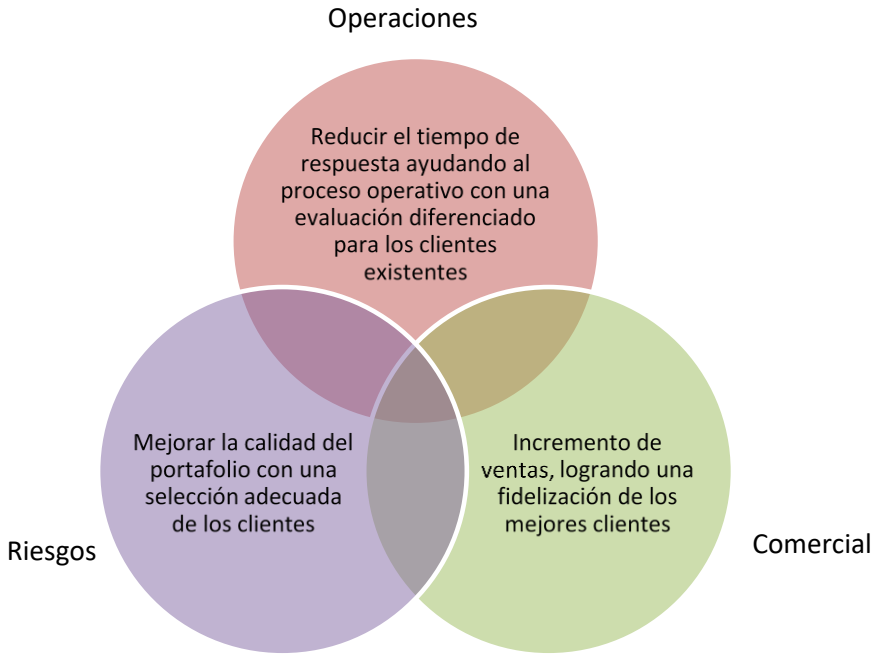
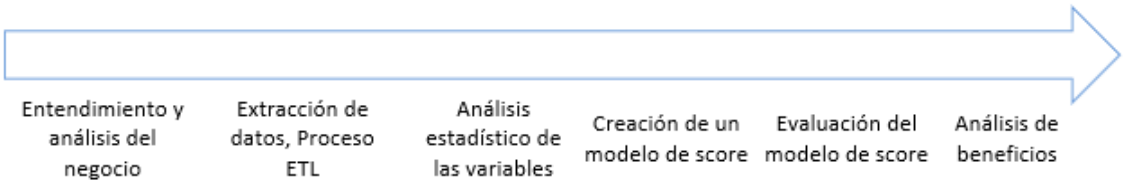


Figura 21. Objetivos de los departamentos implicados en el proyecto. Elaboración propia.

El alcance del proyecto será por medio de machine learning y el análisis de los datos y negocio desarrollar un modelo que permita localizar a los mejores clientes.



Fases del proyecto. Elaboración propia.

Figura 20. Cuadro ponderado de las fases de proceso de sanción de la financiación.Elaboración propia

Para la primera parte del entendimiento y análisis del negocio se entrevistaron las áreas involucradas, se entendieron los procesos de la compañía, la importancia del proyecto y los posibles retos a enfrentar en el desarrollo del mismo.

La segunda parte consta de recopilación de datos sobre los clientes que nos pueden ayudar a mejorar nuestro modelo, sobre estos datos tenemos de 3 tipos.

1. Demográficos del cliente: Información general sobre los clientes, como la edad, sexo, régimen fiscal, lugar de residencia, etc.

2. Buró de crédito: correspondiente al comportamiento externo y reportado en buró de crédito, esta información proporciona líneas de crédito externas

manejadas por el cliente, comportamiento externo, tipos de crédito que el cliente maneja o ha manejado, etc.

3. Datos internos del cliente: correspondientes a información interna del crédito, como enganches, endeudamiento del cliente, plazo del crédito, comportamiento interno, etc.

Los recursos disponibles para alcanzar los objetivos, corresponden principalmente a software utilizados para la creación de la base de datos, así como los recursos intelectuales sobre modelos y conocimiento del negocio.

Las herramientas utilizadas son 2 principalmente: SAS Enterprise Guide para el proceso ETL, creación de la base de datos, enmascarar los datos, creación de variables, dar formato a las mismas, etc. y Python para el análisis estadístico de las variables y el desarrollo del modelo.

Métricas

Como parte del alcance del proyecto, debemos de definir los Indicadores clave del negocio (KPI). Los KPI sirven a las organizaciones para evaluar si están alcanzando sus objetivos.

Las métricas rondan alrededor de la efectividad del modelo, es decir las métricas como KS, ROC y GINI que miden la eficiencia del modelo soportan que el modelo cumpla en gran medida con el objetivo del negocio, adicional a estas se encuentran las de carácter financiero, el incremento en ventas que se espera derivado de una correcta selección de clientes, así como los incrementos en la utilidad de la empresa.

Análisis de actividades: modelo lógico - arquitectura técnica

Como principales actividades y arquitectura técnica del modelo tenemos lo siguiente:

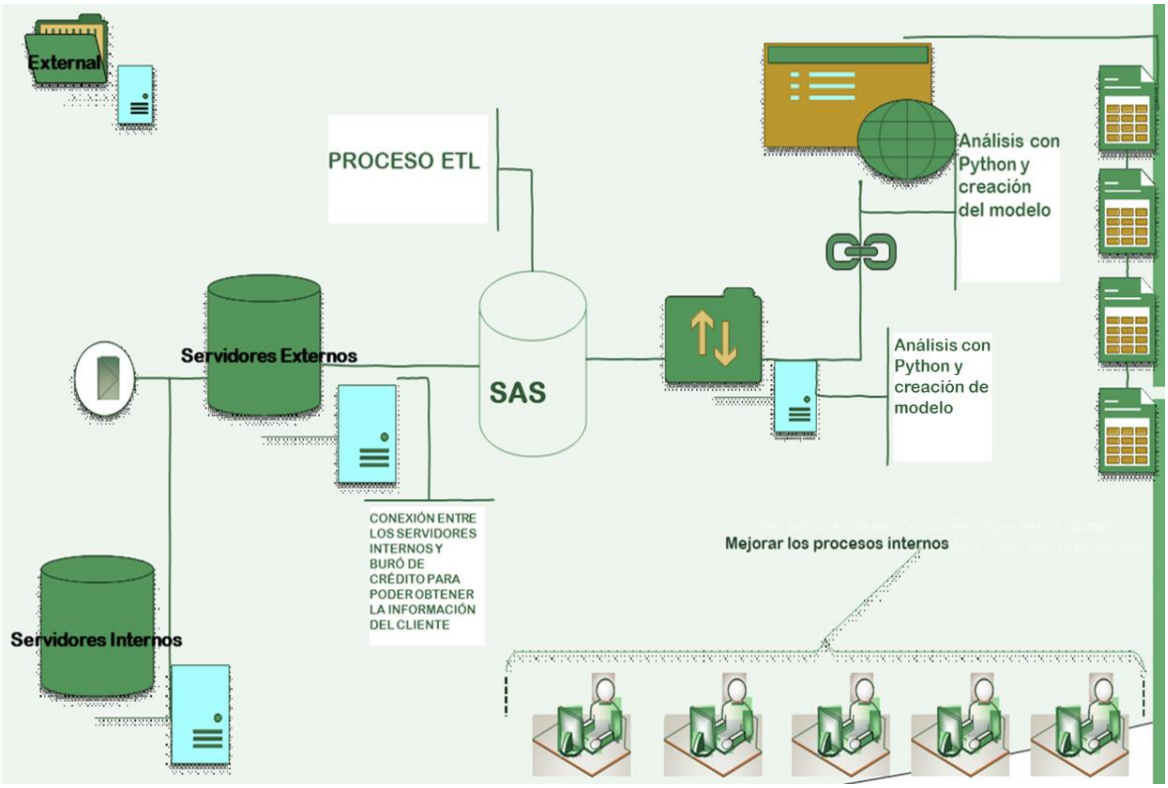


Figura 23. Arquitectura técnica del proyecto. Elaboración propia.

La metodología a usar para la creación del modelo es SEMMA esta es el acrónimo a cinco fases: (Sample, Explore, Modify, Model, Assess) La metodología es propuesta por SAS Institute Inc, la define como: "… proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones de negocios desconocidos.

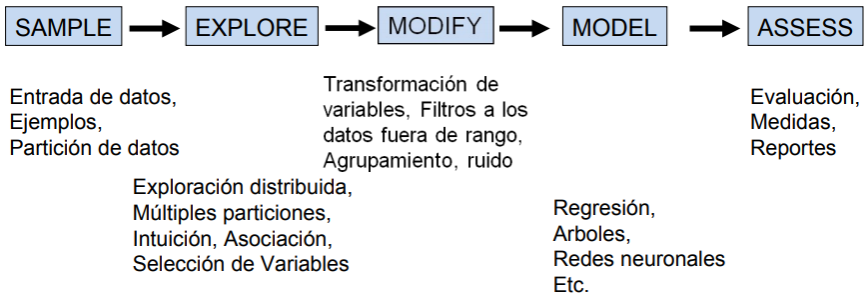


Figura 24. Modelo SEMMA.

Análisis de la BBDD

Para las primeras 3 etapas se encuentra el proceso ETL que consiste en la extracción tratamiento y carga de datos, en los anexos se encuentran el código en SAS con el que se extrajo la información, así como el layout de las variables con descripciones de cada una.

La base de datos consta de la información del primer y segundo préstamo que se han vendido a los diferentes clientes de BNP. Para poder explicar bien la base de datos la vamos a explicar tanto a nivel técnico como a nivel funcional.

Descripción técnica:

En cuanto a la estructura de la BBDD, ésta consta de 79 campos con 8.686 registros se compone de:

- 58 campos tipo Int
- 19 campos tipo Float
- 2 campos tipo Object

Por lo tanto, tenemos muy pocas variables categóricas que transformar, facilitando ese paso en el tratamiento de los datos.
Al analizar los campos, podemos ver que el campo llave es la PK de la tabla, habiendo sido enmascarada previamente para que la BBDD pueda ser objeto de este proyecto

Descripción funcional:

A nivel funcional, la base de datos se podría dividir en dos fuentes de datos principales;

- Datos del Buró de Crédito (BDC)
- Datos internos sobre los clientes

La información del Buró de Crédito se obtiene anterior a realizar la contratación del préstamo, y ya que en esta base de datos hay hasta 2 créditos por cliente, se obtiene en dos momentos temporales diferentes.

El dato más agregado sería el score que da al cliente, que se informa en estos dos campos:

Variable	Descripción
cotenewbdc	Score Buró de crédito 1 crédito
cotenewbdc_2	Score Buró de crédito 2 crédito

También de forma general, se tiene la información del número de cuentas en morosidad y la experiencia que tiene el Buró de Crédito sobre ese cliente:

Variable	Descripción
BDCBCRS13_2	Número de cuentas con morosidad actual 2 crédito
BDCBCRS13	Número de cuentas con morosidad actual
BC_AGE_OLD_OFF_ALL	Experiencia en Buró de Crédito
BC_AGE_OLD_OFF_ALL_2	Experiencia en Buró de Crédito 2 crédito

Para poder relativizar la información del Buró de Crédito, tenemos la información de cuántas veces se ha consultado la información sobre ese cliente en los últimos 30 y 60 días:

Variable	Descripción
BDCBCCONSULAM30D	Numero de consulta (excepto de Cetelem) hechas en BdC en los 60 últimos días para el titular.

BDCBCCONSULNUM30D	Numero de consulta (excepto de Cetelem) hechas en BdC en los 30 últimos días para el titular.
BDCBCCONSULNUM60D	Número de consultas a BDC.
BDCBCCONSULAM30D_2	Numero de consulta (excepto de Cetelem) hechas en BdC en los 60 últimos días para el titular. 2 crédito
BDCBCCONSULNUM30D_2	Numero de consulta (excepto de Cetelem) hechas en BdC en los 30 últimos días para el titular. 2 crédito
BDCBCCONSULNUM60D_2	Número de consultas a BDC. 2 crédito

Otra información muy relevante, es el "Month of payment" (MOP), que indica la antigüedad de los atrasos, pudiendo identificar quién lleva más meses debiendo dinero:

Variable	Descripción
BDCBCCL12MNEW	Peor MOP actual
BDCBCCL12MNEW_2	Peor MOP actual 2 crédito
MEXBCCL24M	Peor MOP histórico de los últimos 24 meses
MEXBCCL3M	Peor MOP histórico de los últimos 3 meses
MEXBCCLNR	Peor MOP actual comunicaciones/servicios
MEXBCCL24M_2	Peor MOP histórico de los últimos 24 meses 2 crédito
MEXBCCL3M_2	Peor MOP histórico de los últimos 3 meses 2 crédito
MEXBCCLNR_2	Peor MOP actual comunicaciones/servicios 2 crédito

Bajando un poco más al detalle, se puede observar el número de créditos que tiene el cliente en el momento de pedir los préstamos, así como el tipo de crédito y el estado del mismo:

Variable	Descripción
BDCBCACCLANUMNPA	El número de créditos activo tipo clásico mal pagado
BDCBCACCLANUMWPA	El número de créditos activo tipo clásico bien pagado
BDCBCACREVNUMNPA	El número de créditos activos tipo tarjeta mal pagado
BDCBCACREVNUMWPA	El número de créditos activos tipo tarjeta bien pagado
BDCBCCLADUENUM12M	El número de créditos tipo clásico vencido en los 12 últimos meses
BDCBCCOMNUM	El número de deuda de tipo comunicaciones
BDCBCINACREV	El número de créditos tipo tarjeta sin actividad
BDCBCLOSCLANUM12M	El número de créditos tipo clásico en pérdida total o parcial en los 12 últimos meses
BDCBCLOSREVNUM12M	El número de créditos tipo tarjeta en pérdida total o parcial en los 12 últimos meses
BDCBCREVDUENUM12M	El número de créditos tipo tarjeta vencido en los 12 últimos meses
BDCBCACCLANUMNPA_2	El número de créditos activo tipo clásico mal pagado 2 crédito
BDCBCACCLANUMWPA_2	El número de créditos activo tipo clásico bien pagado 2 crédito

BDCBCACREVNUMNPA_2	El número de créditos activos tipo tarjeta mal pagado 2 crédito
BDCBCACREVNUMWPA_2	El número de créditos activos tipo tarjeta bien pagado 2 crédito
BDCBCCLADUENUM12M_2	El número de créditos tipo clásico vencido en los 12 últimos meses 2 crédito
BDCBCCOMNUM_2	El número de deuda de tipo comunicaciones 2 crédito
BDCBCINACREV_2	El número de créditos tipo tarjeta sin actividad 2 crédito
BDCBCLOSCLANUM12M_2	El número de créditos tipo clásico en pérdida total o parcial en los 12 últimos meses 2 crédito
BDCBCLOSREVNUM12M_2	El número de créditos tipo tarjeta en pérdida total o parcial en los 12 últimos meses 2 crédito
BDCBCREVDUENUM12M_2	El número de créditos tipo tarjeta vencido en los 12 últimos meses 2 crédito

De todas formas, esta información de detalle estaría incompleta sin la información de los importes, ya que un cliente con una morosidad en un solo préstamo, pero con un importe muy alto, puede generar un daño muchísimo mayor a la entidad financiera:

Variable	Descripción
BDCBCACCLAAMNPA	El importe total pendiente créditos activos tipo clásico mal pagado
BDCBCACCLAAMWPA	El importe total pendiente créditos activos tipo clásico bien pagado
BDCBCACREVAMNPA	El importe total pendiente créditos activos tipo tarjeta mal pagado
BDCBCACREVAMWPA	El importe total pendiente créditos activos tipo tarjeta bien pagado
BDCBCCOMDUENUM	El importe total pendiente de deudas de tipo comunicaciones
BDCBCLOSCLAAM12M	El importe total de perdida de créditos tipo clásico en los 12 últimos meses
BDCBCACCLAAMNPA_2	El importe total pendiente créditos activos tipo clásico mal pagado 2 crédito
BDCBCACCLAAMWPA_2	El importe total pendiente créditos activos tipo clásico bien pagado 2 crédito
BDCBCACREVAMNPA_2	El importe total pendiente créditos activos tipo tarjeta mal pagado 2 crédito
BDCBCACREVAMWPA_2	El importe total pendiente créditos activos tipo tarjeta bien pagado 2 crédito
BDCBCCOMDUENUM_2	El importe total pendiente de deudas de tipo comunicaciones 2 crédito
BDCBCLOSCLAAM12M_2	El importe total de perdida de créditos tipo clásico en los 12 últimos meses 2 crédito

El total de los campos del Buró de crédito son 52, por lo que esta fuente de información aporta 2/3 de la base de datos del proyecto.

La información restante es la que tiene el Banco de los préstamos que ha realizado con estos clientes, teniendo como datos importantes más relevantes a nivel de cliente el scoring que se le dio en su momento, el precio de los autos comprados y el montante del crédito concedido:

Variable	Descripción
NEWCOTESCO	Score Interno 1 crédito
NEWCOTESCO_2	Score Interno 2 crédito
montofinal	montofinal
montofinal_2	montofinal_2
Precio_lista	Precio Vehículo 1 crédito
Precio_lista_2	Precio Vehículo 2 crédito

Además de esa información básica, tendríamos la información de los plazos de los créditos concedidos y la capacidad de endeudamiento del cliente:

Variable	Descripción
BC_NUM_SAT_OFF_A	Cuentas satisfactorias de Auto
BUDTITSALAIRE	Ingreso 1 Crédito
MAX_CREDIT_LIMIT	Línea de Crédito Máximo
PCTEND	Endeudamiento 1 crédito
PCTENG	Enganche 1 Crédito
TITANCIENNETEEMPL	Fecha de ingreso laboral
plazo	plazo
persona2	persona2
HIT	HIT BC 1 CREDITO
BC_NUM_SAT_OFF_A_2	Cuentas satisfactorias de Auto 2 crédito
BUDTITSALAIRE_2	Ingreso 2 Crédito
MAX_CREDIT_LIMIT_2	Línea de Crédito Máximo 2 crédito
PCTEND_2	Endeudamiento 2 crédito
PCTENG_2	Enganche 2 Crédito
TITANCIENNETEEMPL_2	Fecha de ingreso laboral 2 crédito
plazo_2	plazo_2
porc_liquid	porc_liquid

Para finalizar, tendremos la información de cuáles han sido morosos para el banco y cuáles no, detallando el número de cuotas que se han retrasado, que son los campos que nos van a determinar si un cliente es susceptible para marcarlo como posible candidato para diferentes acciones comerciales para BNP, o si por el contrario, finalizar toda relación una vez terminada las obligaciones contractuales:

Variable	Descripción
max_ret_1	Número de plazos del primer préstamo que el cliente se ha retrasado
max_ret_2	Número de plazos del segundo préstamo que el cliente se ha retrasado
Malo	Indicador que marca si un cliente se considera malo o no

Un aspecto importante para la creación del modelo es la definición de una marca de malo. Los datos incluyen 2 variables relativas a los máximos retrasos de los créditos, definiendo la variable de malo de la siguiente manera:

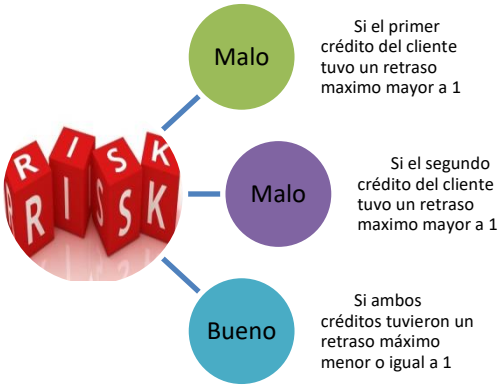


Figura 25. Clasificación de clientes por riesgos. Elaboración propia.

Nuestro proceso ETL nos arrojó 78 variables que pueden ayudar a entender a los clientes y generar un modelo que seleccione a los mejores clientes. De estas 78 variables tenemos que seleccionar las que predigan de mejor manera nuestra variable objetivo.

7. Solución tecnológica

Aprendizaje automático es aprender de los datos, es descubrir la estructura y los patrones que subyacen en ellos. El objetivo principal del aprendizaje automático es la extracción de la información contenida en un conjunto de datos con el fin de adquirir conocimiento que permita tomar decisiones sobre nuevos conjuntos de datos. Formalmente, se define como [Mitchell, 2006]:

Un sistema aprende de la experiencia E con respecto a un conjunto de tareas T y una medida de rendimiento P, si su rendimiento en T, medido según P, mejora con la experiencia E.

La solución tecnológica que se plantea se fundamenta en los algoritmos de Machine Learning para crear un modelo predictivo. Como se indicó más arriba, la selección de este modelo se realizará mediante la metodología SEMMA se SAS. Esta metodología conta de cinco fases, pero podemos agruparlas en dos grupos:

- Tratamiento de los datos (Sample, Explore, Modify)
- Elección del Modelo de predicción (Model, Assess)

Tratamiento de los datos:

En un primer análisis preliminar, se puede ver como los datos tienen algunas correlaciones, sobre todo en la parte de los importes de crédito bien y mal pagados:

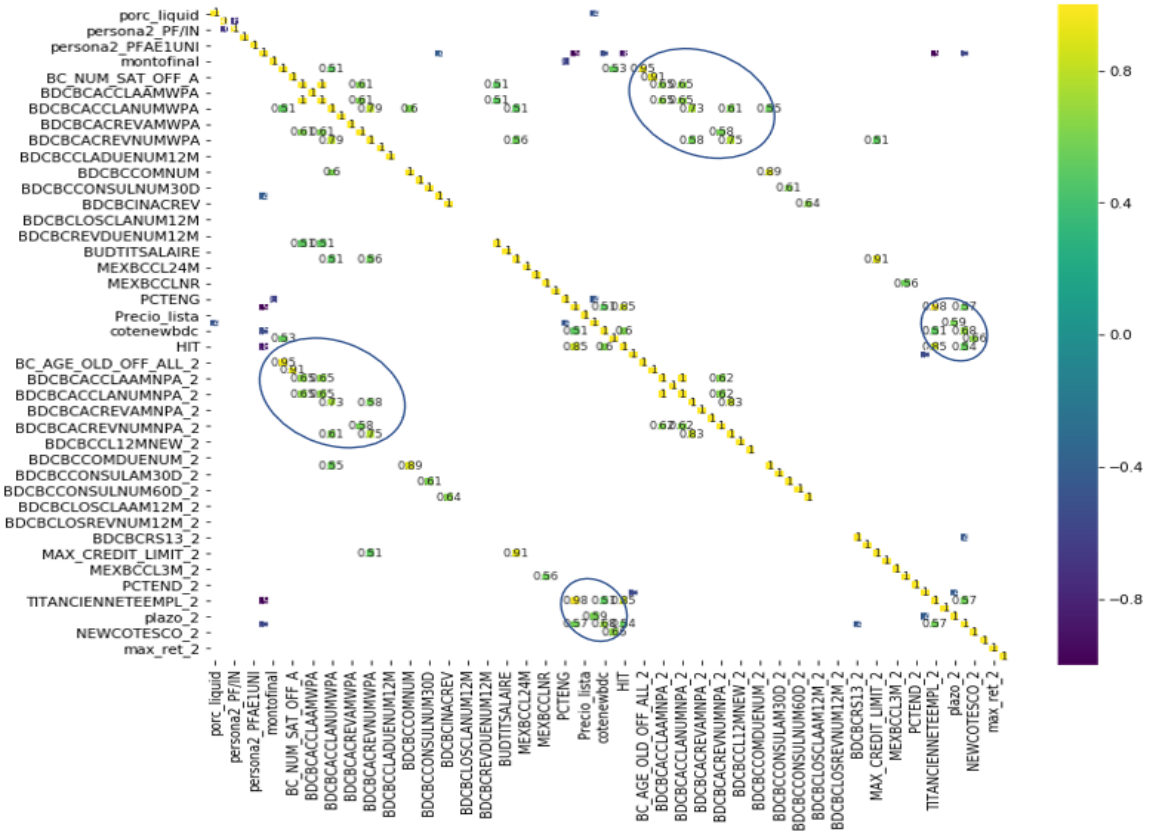


Figura 26. Matriz de correlación. Elaboración propia.

Si decidiéramos analizar la información sin tratar, el resultado que nos daría es que no existe ningún clúster claro, ya que el 99% de los datos se estarían en el mismo clúster:

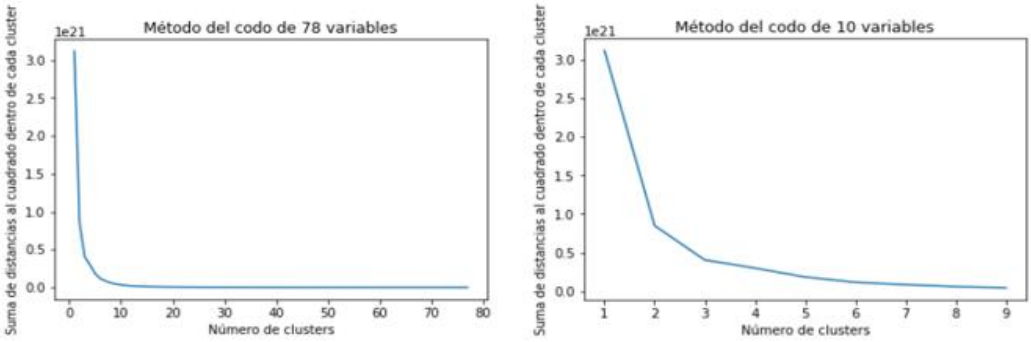


Figura 27. Método del codo.

Por lo tanto, la primera parte será realizar un análisis de componentes principales para poder eliminar algunas variables y quedarnos con aquellas que expliquen la mayor cantidad de varianza.

El análisis de componentes principales es un método estadístico multivariante que se clasifica como método de simplificación o de reducción de la dimensión. El ACP permite describir de forma sintética la estructura e interrelación de las variables originales.

El método tiene por objeto transformar un conjunto de variables en un nuevo conjunto denominado componentes principales. Los nuevos componentes tienen la característica de ser ortogonales (no correlacionados) y se ordenan de acuerdo a la cantidad de información (varianza) que llevan incorporada. Las componentes principales se expresan como una combinación lineal de las variables originales.

Por medio de la librería sklearn obtenemos el número de factores con los cuales se tiene la mayor cantidad de varianza que explique la base de datos (Código Anexo).

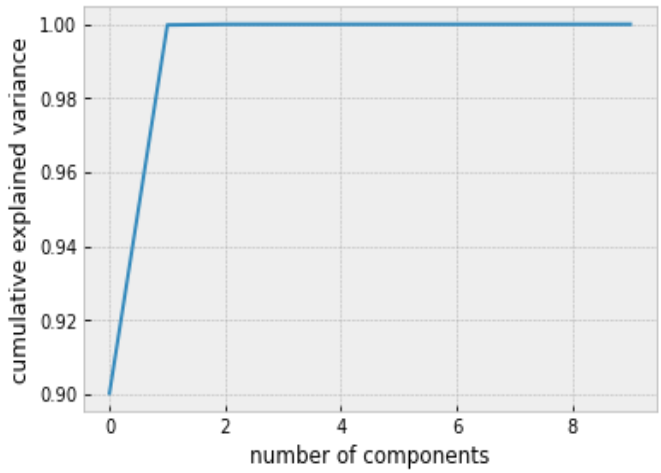


Figura 28. Componentes óptimos.

Con 2 componentes principales obtenemos el 99% de varianza de la data.

Se calcula la matriz rotada nos presenta la tributación que tiene cada variable al factor, por medio de la librería factor_analyzer, revisamos las variables que tributen más de 0.3 a algún factor y estas quedan dentro del modelo, al realizar este procedimiento nos quedamos con 31 variables de tipo numérico (Tabla Anexo).

El siguiente paso es revisar la correlación de las variables, esto se realizara por medio de clúster de variables numéricas en clústeres jerárquicos o disjuntos, con cada clúster se asocia una combinación lineal de las variables en el clúster, dicha combinación lineal puede ser la primera componente principal o la componente centroide, esto maximiza la varianza que es explicada por los componentes clúster, así el clúster de variables nos permite reducir la dimensión y la multicolinealidad de las variables.

Este análisis se realizará en Python por medio de la librería varclushi (Código Anexo), este proceso nos arroja un total de 21 clúster de variables, de los cuales tomaremos una variable por el indicador **RS_Ratio,** para aquellos clusters donde la diferencia sea mínima o exista un criterio de negocio fuerte se elegirá otra variable.

Estas 21 variables más la variable tipo de persona se meterán en el modelo de regresión para la obtención de un score que permita seleccionar a los mejores clientes.

Las variables resultantes son:

Nº Column	Clúster	Variable	RS_Own	RS_NC	RS_Ratio
0	0	TITANCIENNETEEMPL	88,3%	19,7%	14,5%
7	1	BDCBCACCLAAMNPA	77,5%	8,8%	24,7%
16	2	BDCBCACREVNUMWPA_2	81,2%	13,6%	21,7%
18	3	PCTENG	65,8%	24,6%	45,3%
21	4	BC_AGE_OLD_OFF_ALL	78,8%	3,9%	22,1%
28	5	BDCBCCONSULNUM30D_2	67,4%	13,0%	37,5%
30	6	max_ret_2	73,2%	1,2%	27,1%
33	7	BDCBCCL12MNEW	58,7%	2,3%	42,2%
37	8	PCTEND	72,6%	13,9%	31,8%
39	9	MEXBCCLNR	78,1%	1,1%	22,2%
42	10	BDCBCINACREV_2	81,6%	7,4%	19,8%
44	11	BDCBCCOMNUM_2	94,5%	18,6%	6,8%
46	12	BDCBCCLADUENUM12M	73,9%	2,9%	26,9%
48	13	NEWCOTESCO	83,0%	12,1%	19,4%
51	14	BUDTITSALAIRE_2	64,1%	0,3%	36,0%
52	15	BDCBCACREVAMNPA_2	51,1%	1,1%	49,4%
58	16	BDCBCACCLAAMWPA_2	73,4%	8,2%	29,0%
60	17	BDCBCACREVAMWPA_2	72,4%	14,1%	32,1%
61	18	porc_liquid	71,8%	3,8%	29,4%
63	19	BDCBCCONSULAM30D	52,6%	0,4%	47,6%
65	20	BC_NUM_SAT_OFF_A_2	68,2%	9,5%	35,1%
67	21	MEXBCCL3M	100,0%	1,2%	0,0%

Figura 29. Variables de persona en modelo de regresión.

Como siguiente paso es identificar qué modelo de predicción usaremos para poder tener el mejor criterio de clasificación.

Modelo de predicción

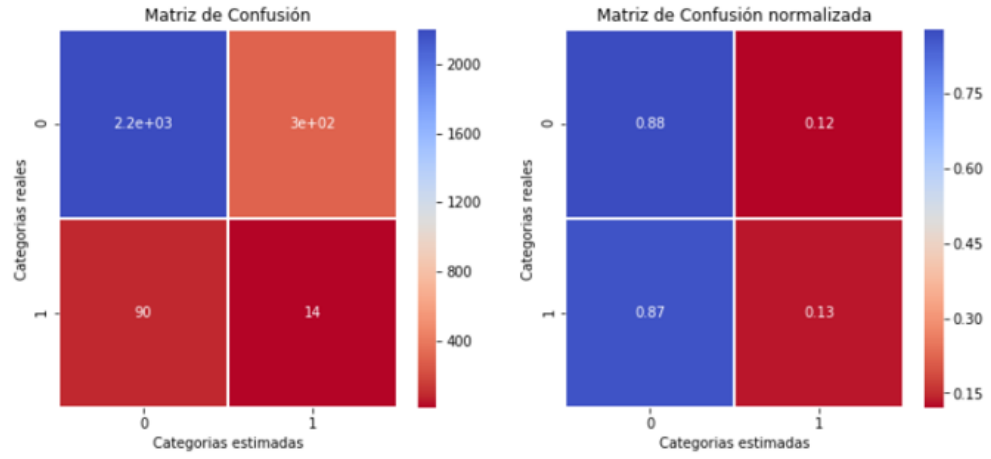
Una vez decididas las variables a utilizar, dividimos la BBDD en dos, quedando un conjunto entrenamiento del 30%.
Con el fin de identificar cual es el mejor modelo y cuál se ajusta de mejor manera a los datos hemos decidido aplicar 5 algoritmos, los cuales son:

- Linear SVM
- Random Forest
- Regresión logística
- Neural networks
- Árbol de decisiones.

Linear SVM:

Las máquinas de soporte vectorial (Support Vector Machines) son algoritmos de aprendizaje supervisado, relacionados con problemas de clasificación y regresión. Una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos más cercanos de las 2 clases, llamado vector soporte.

Al ajustar el modelo obtenemos la matriz de confusión y el área bajo la curva:




```
---Report de clasificación---
              precision    recall  f1-score   support

           0       0.96      0.88      0.92      2502
           1       0.04      0.13      0.07       104

   micro avg       0.85      0.85      0.85      2606
   macro avg       0.50      0.51      0.49      2606
weighted avg       0.92      0.85      0.88      2606
```

AUC: 0,5068

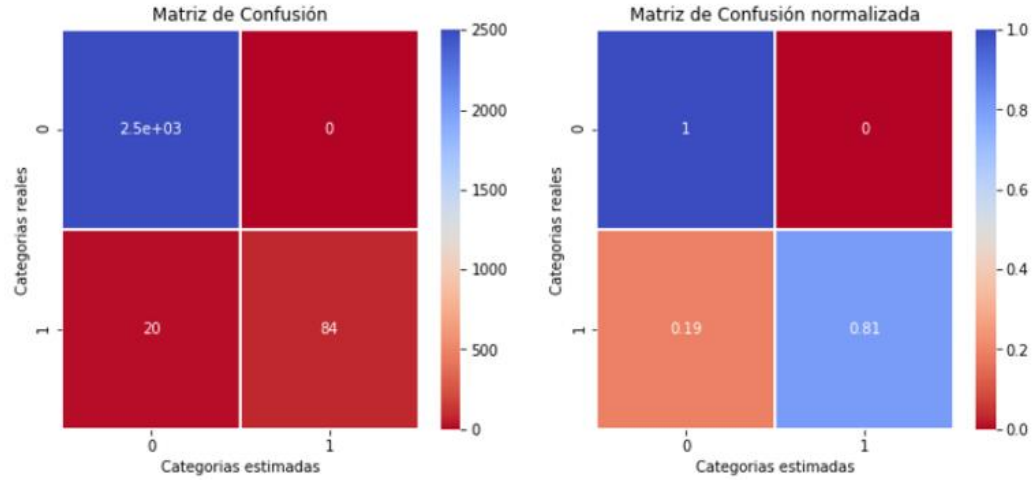
Figura 30. Matriz de confusión y AUC del modelo Linear SVM.

El resultado del AUC para este modelo es muy parecido a un aleatorio, por lo que su capacidad predictiva es muy baja. También se puede deducir que identificaría muy pocos fraudes (solo el 13%).

Random Forest:

Random forest (o random forests) es una combinación de árboles predictores, de forma que cada árbol depende de los valores de un vector aleatorio probado independientemente, y con la misma distribución para cada uno. Es muy popular y ampliamente utilizado, debido a que es más simple de entrenar y ajustar que otros algoritmos similares, como el de boosting o Bagging.

Al ajustar el modelo obtenemos la matriz de confusión y el área bajo la curva:




```
---Report de clasificación---
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      2502
           1       1.00      0.81      0.89       104

   micro avg       0.99      0.99      0.99      2606
   macro avg       1.00      0.90      0.94      2606
weighted avg       0.99      0.99      0.99      2606
```

AUC: 0,9038

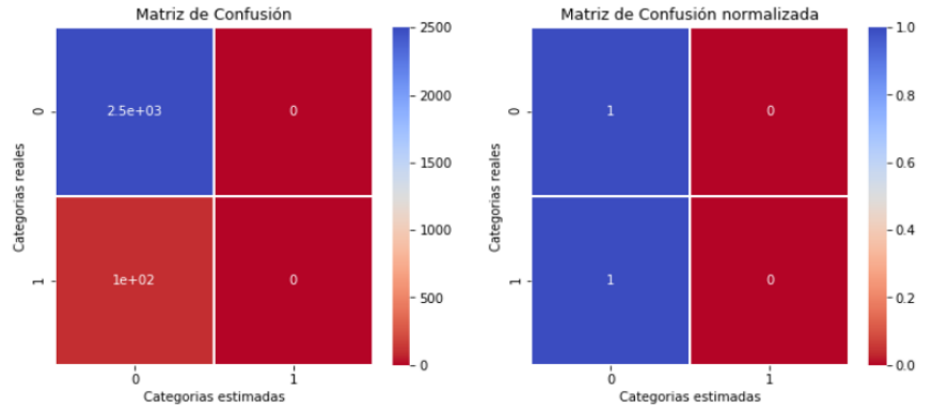
Figura 31. Matriz de confusión y AUC del modelo Random Forest.

Este modelo si tiene un AUC muy alto (0,9), y predice un alto porcentaje de morosidad, sin perder ninguna operación donde se gana dinero, por lo tanto, podría ser uno de los modelos elegidos.

Regresión logística:

Regresión logística es un tipo de análisis de regresión ampliamente utilizado para predecir el resultado de una variable categórica en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de ocurrencia de un evento como función de otros factores, mediante una función logística.
Este algoritmo fue utilizado con los siguientes parámetros:

Al ajustar el modelo obtenemos la matriz de confusión y el área bajo la curva:




```
---Report de clasificación---
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      2502
           1       0.00      0.00      0.00       104

   micro avg       0.96      0.96      0.96      2606
   macro avg       0.48      0.50      0.49      2606
weighted avg       0.92      0.96      0.94      2606
```

AUC: 0,5

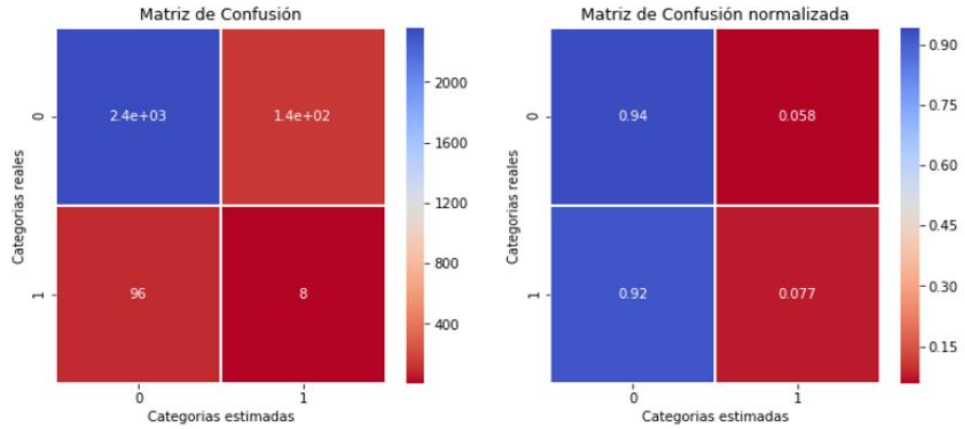
Figura 31. Matriz de confusión y AUC del modelo Regresión Logística.

Este modelo no está prediciendo, ya que estimaría todo como no fraude, por lo que quedaría descartado.

Neural Networks:

Las redes neuronales son modelos computacionales inspirados en el cerebro humano, para resolver los problemas de forma similar a este. Está conformada por un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales. La información de entrada atraviesa la red neuronal, sometiéndose a diversas operaciones produciendo unos valores de salida.

Al ajustar el modelo obtenemos la matriz de confusión y el área bajo la curva:




```
---Report de clasificación---
              precision    recall  f1-score   support

           0       0.96      0.94      0.95      2502
           1       0.05      0.08      0.06       104

   micro avg       0.91      0.91      0.91      2606
   macro avg       0.51      0.51      0.51      2606
weighted avg       0.92      0.91      0.92      2606
```

AUC: 0,5095

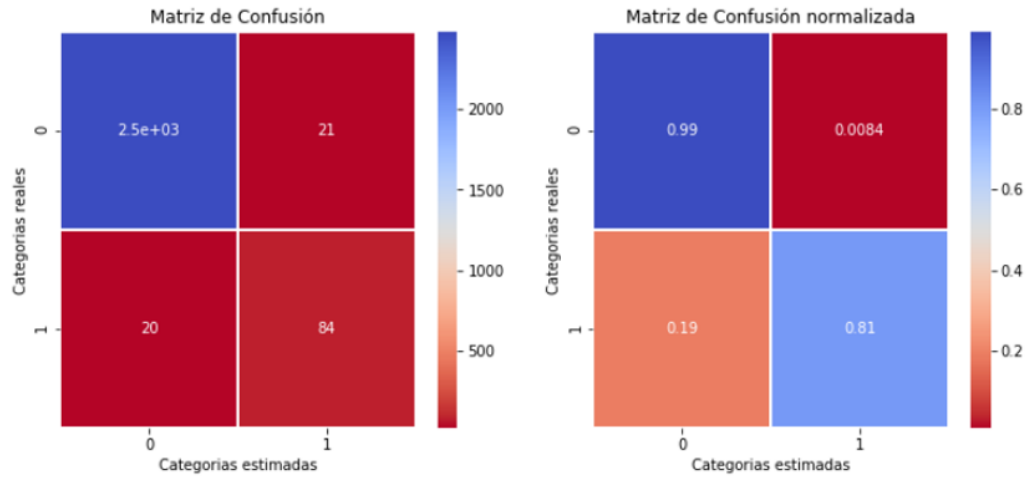
Figura 32. Matriz de confusión y AUC del modelo Neural Networks.

El modelo de neural networks tampoco consigue predecir demasiado, con un AUC muy parecido al Linear SVM. La diferencia entre los dos radica en que este predice peor los fraudes (solo un 8%), aunque acierta mejor en los no fraudes. De todas formas, también quedaría descartado.

Árboles de decisión:

Son modelos de predicción, en los que, dado un conjunto de datos, se fabrican diagramas de construcciones lógicas. Son muy similares a los sistemas de predicción basados en reglas, y sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Al ajustar el modelo obtenemos la matriz de confusión y el área bajo la curva:




```
---Report de clasificación---
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      2502
           1       0.80      0.81      0.80       104

   micro avg       0.98      0.98      0.98      2606
   macro avg       0.90      0.90      0.90      2606
weighted avg       0.98      0.98      0.98      2606
```

AUC: 0,8996

Figura 33. Matriz de confusión y AUC del modelo Árboles de Decisión.

Este modelo también predice muy bien la morosidad. La única diferencia con el Random Forest es que genera un 1% de falsos positivos, pero su AUC es casi idéntico.

Conclusiones

El resultado final de los algoritmos utilizados es el siguiente:

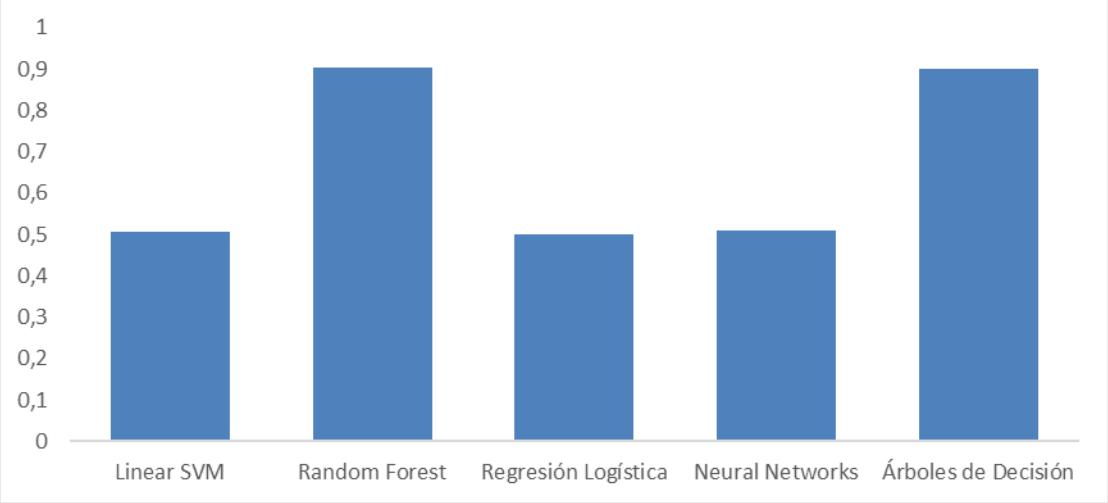


Figura 34. Cuadro resumen de algoritmos utilizados en el modelo.

Dado que Random Forest el método de árboles de decisión nos dieron un resultado similar decidimos escoger el primero ya que el mismo tiene las siguientes ventajas:

- Existen muy pocas suposiciones y por lo tanto la preparación de los datos es mínima
- Puede manejar hasta miles de variables de entrada e identificar las más significativas. Método de reducción de dimensionalidad.
- Una de las salidas del modelo es la importancia de variables.
- Incorpora métodos efectivos para estimar valores faltantes.
- No genera falsos positivos
- Los árboles de decisión son propensos al sobreajuste, especialmente cuando un árbol es particularmente profundo.

En resumen, se demuestra que el algoritmo de Random Forest constituye una excelente solución para este tipo de problemas debido a que su rendimiento y valores de resultado muestran robustez suficiente para generar confianza a la hora de fiar en el resultado, en este caso tenemos una cantidad significativa de datos y variables, muestra que Random Forest tiene muy buenos niveles de detección lo convierten en una opción sencilla y competitiva.

De todas formas, dada lo parejo del resultado, puede tener sentido monitorear los dos, y con el tiempo ir viendo cuál se comporta mejor. Por otro lado, hay que tener en cuenta que los árboles de decisión son fáciles de interpretar y de hacer visualizaciones.

8. Fases de proyecto

Para la correcta ejecución de un proyecto, es necesario dividirlo en diferentes fases e hitos. La metodología de proyectos ayuda a estandarizar qué pasos hay que seguir en un proyecto, y así minimizar las posibilidades de fracaso del mismo.

Aunque se siga una metodología estandarizada, los proyectos pueden ser de temáticas totalmente diferentes, haciendo que la complejidad y duración de cada fase no tengan que ver en un proyecto u otro. Esto hace que la estrategia a seguir y las habilidades necesarias para la gestión y ejecución de un proyecto serán diferentes en cada caso.

Los pasos que vamos a seguir en este proyecto serán los siguientes:

1. Entender el problema de negocio.
2. Definir los objetivos y el alcance del proyecto.
3. Seleccionar e identificar los datos.
4. Preparar los datos.
5. Analizar y transformar los datos.
6. Desarrollar y entrenar los modelos.
7. Validar los modelos.
8. Optimizar y medir el rendimiento.

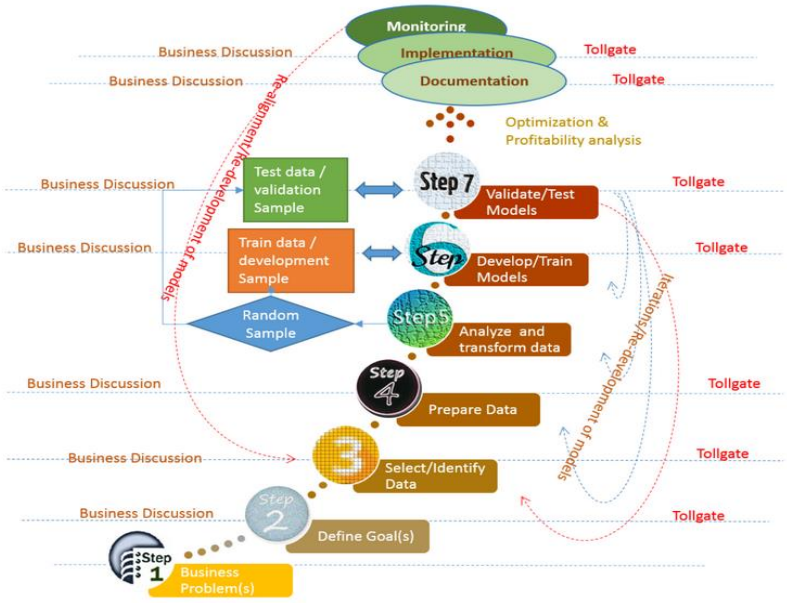


Figura 35. Fases del proyecto.

Paso 1: Entender el problema de negocio

Las organizaciones actuales están aumentando su uso de análisis avanzados y modelos predictivos. Los procesos de generación de modelos predictivos implican la preparación de datos, la verificación de la calidad de los datos, la reducción, el modelado, la predicción y el análisis de resultados. La generación de modelos predictivos de alta calidad es una actividad que requiere mucho tiempo debido al proceso de ajuste para

encontrar los parámetros óptimos del modelo y que a menudo se requiere para volver a desarrollar, ajustar o reutilizar los modelos en el futuro. Por eso es importante los siguientes puntos:

- Seguir las metodologías estándar y las mejores prácticas de la industria.
- Establecer gobernanza en la construcción de modelos.
- Mantener estándares de calidad en todos los ámbitos.
- Garantizar la reutilización de los códigos.
- Ahorrar tiempo y costes para el desarrollo futuro.
- Mejores prácticas.
- Cumplir con revisiones internas y externas.
- Requerimientos de auditoria.

Esta fase en nuestro proyecto se planificó en 3 días hábiles. Esta fase se ha estimado corta en tiempo debido a qué un integrante del equipo está muy familiarizado con el problema. El trabajo a realizar es confirmar con diferentes departamentos que necesidades tienen a las ya identificadas anteriormente.

Paso 2: Definir los objetivos y el alcance del proyecto

Después de las entrevistas con las diferentes áreas, reuniones de equipo y conociendo las limitaciones temporales del proyecto, se acota un alcance y unos objetivos.

Este paso es muy importante debido a que el éxito del proyecto se va a medir en función de cómo se consigan los objetivos definidos en este apartado.

En este proyecto se planificaron 3 días para ver cómo mediríamos el éxito del proyecto, y cuál sería el alcance del mismo.

Paso 3 Seleccionar y obtener los datos

Dado que el análisis predictivo se trata de utilizar grandes volúmenes de datos para obtener información sobre las tendencias y mantenerse a la vanguardia del juego, la fase de recopilación de datos es crucial para el éxito de la iniciativa. Lo más probable es que esto incluya información de múltiples fuentes, por lo que debe haber un enfoque unitario de los datos.

La mayoría de las veces, los datos se recopilarán en un lago de datos, que no debe confundirse con un almacén de datos, que tiene algunas diferencias estructurales significativas. Un lago de datos contiene información en estado sin procesar. Esto significa que puede ir desde estructurado (tablas) hasta semiestructurado, como XML o no estructurado (comentarios de redes sociales). Para el éxito del proyecto, es obligatorio comprender las diferencias y emplear las herramientas adecuadas.
En este proyecto las fuentes de datos han sido 2 principalmente:

- Base de originación de créditos:
- Bases de comportamiento crediticio

Dado que la información estaba estructurada en dos bases de datos actuales, la extracción de los datos fue relativamente sencilla. Más complejo fue analizar todas las variables y entenderlas. La duración total de estos dos pasos fue de 4 días.

Paso 4: Preparar los datos

En este paso, se pretende preparar los datos en el formato correcto para el análisis y la herramienta que desee utilizar. Para ello, es necesario:

- Realizar procesos de limpieza del dato.
- Definir variables y crear diccionario de datos.
- Unir / agregar múltiples conjuntos de datos.

Este punto fue crucial en el proyecto, ya que sin tener un input de datos en estado correcto no se habría podido hacer un correcto análisis de variables y modelado. Esta tarea duró 4 días hábiles.

Paso 5: Analizar y transformar variables. Muestreo aleatorio

Una vez que los datos están en forma y funcionales se realiza:

- Análisis univariante: para verificar la distribución de cada una de las variables y características.
- Análisis multivariados: para verificar las relaciones con otras variables y con variables dependientes.

Según el tipo de modelo que vaya a utilizar, es posible que necesite transformar las variables utilizando uno de los enfoques:

- Enfoque de unión: crear grupos distintos.
- Transformación:
    - Logarítmico, polinomial.
    - Raíz cuadrada, inversa, cuadrada, boxCox.
    - Tratamientos de valor extremo (atípicos).
- Reducción de las dimensiones.

En este proyecto se utilizaron procesos de:

- Eliminación de las variables categóricas.
- Normalización de datos.
- Análisis de correlación de datos.
- Generación de clústers y selección de las variables finales.

En cuanto al muestreo aleatorio, se hicieron varias pruebas con diferente porcentaje de tamaño del conjunto de entrenamiento, y al final nos pusimos de acuerdo de que el entrenamiento iba a ser del 70%.

Figura 36. Porcentaje de entrenamiento y validación según el tipo de variable.

Paso 6: Selección del modelo y desarrollo de modelos (capacitación)

Cuando se trata de modelar, a menudo es mejor usar las herramientas existentes. Hay innumerables bibliotecas, construidas en lenguajes de programación de código abierto como Python y R. No hay tiempo para reinventar la rueda, es más importante conocer las opciones disponibles y elegir la mejor para el trabajo. El objetivo final debería ser democratizar el modelado y ponerlo a disposición de los analistas de negocios, así como de los científicos de datos.

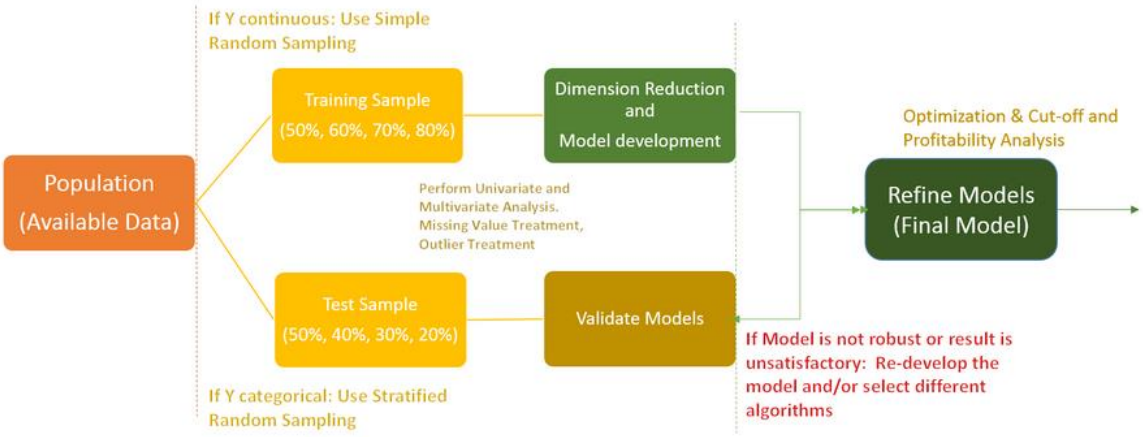


Figura 37. Proceso de validación de modelos

Para este proyecto, se han utilizado 5 modelos diferentes:

- Linear SVM
- Random Forest
- Regresión logística
- Neural networks
- Árbol de decisiones.

Esto nos ha permitido hacer muchas pruebas cambiando de variables y de tamaño de la muestra para buscar el mejor resultado. Esta fase ha sido la más extensa del proyecto, ocupándonos un 40% del tiempo del proyecto.

Paso 7: Validar modelos (pruebas), optimizar y rentabilidad

La realidad no es estática; tampoco lo son los datos. Un modelo puede ser válido durante un cierto período, mientras que las condiciones externas no cambian significativamente. Es una buena práctica volver a visitar los modelos periódicamente y probarlos con nuevos datos para asegurarse de que no hayan perdido su importancia.

Las preferencias de los clientes y las tendencias en los mercados de consumo a veces cambian tan rápido que las expectativas anteriores se convierten rápidamente en las noticias de ayer.

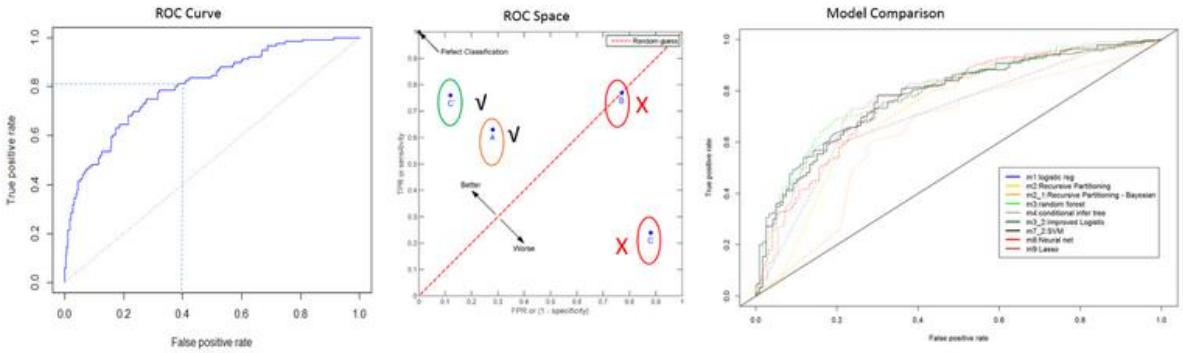


Figura 38. Método de la curva.

Este punto es muy importante debido a los siguientes puntos:

- Hay que estar seguro del mensaje que se va a transmitir es correcto.
- Hay que saber vender la utilidad del proyecto y qué beneficios aporta

En nuestro caso, al no existir un beneficio económico directo, será en siguientes proyectos donde se definirán estrategias comerciales para rentabilizar la segmentación de clientes.

Esta última fase tendrá una duración de 19 días, siendo su mayor parte la optimización de resultados y validación de los mismos.

Resumen de las fases

Resumiendo lo detallado en los puntos anteriores, e proyecto ha tenido una duración de 67 días laborables (en esfuerzo), de los cuáles se han repartido de la siguiente forma:

- 6 en la planificación (10%)
- 8 en la extracción y limpieza de los datos (10%)
- 34 en el análisis y modelado de los datos (50%)

- 19 en la validación de los datos (30%)

El detalle en tiempos sería el siguiente:

		Días
Planeamiento	Comprender el objetivo comercial	3
	Definir objetivos de modelado	3
Datos		
	Extracción de Datos de la base de datos	1
	Análisis de Datos	3
	Depuración y Normalizado de datos	4
Modelado		
	Analizar y transformar variables	4
	Muestreo Aleatorio	3
	Selección del modelo	7
	Desarrollo de Modelo	20
Validación		
	Realizar pruebas	5
	Optimizar resultados	5
	Validar modelos	7
	Análisis de rentabilidad	2
	Total	67

Figura 39. Cuadro resumen de los tiempos empleados en cada fase del proyecto.

Y este sería el diagrama de Gantt:

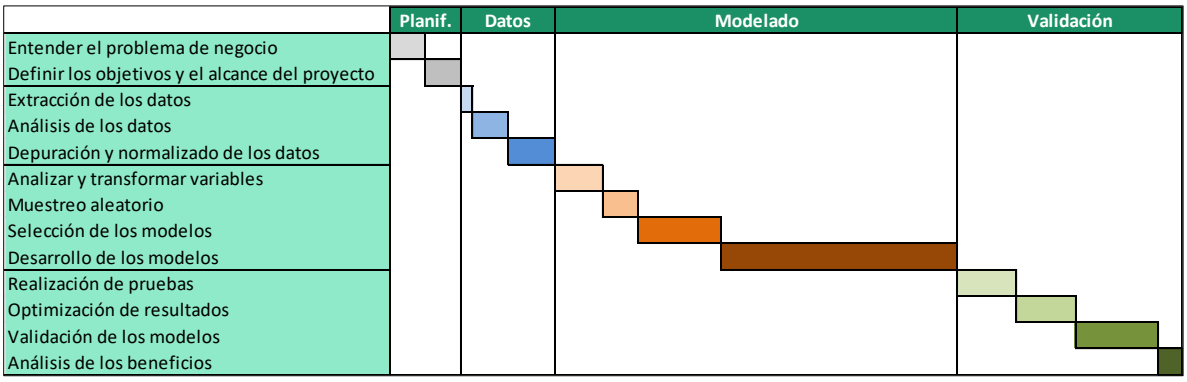


Figura 40. Diagrama de Gantt. Elaboración propia.

9. Beneficios esperados

Los beneficios que esperamos obtener a nivel de proyecto son:

- Obtener una Base de Datos de clientes depurada con aquellos clientes que sean de vital importancia para la entidad.
- Identificar las bases de las principales estrategias de marketing a llevar a cabo para fidelizar a los clientes de la compañía.
- Propuestas de mejora y eficiencia para atraer a nuevos clientes.
- Establecer un modelo de control y seguimiento del cliente, para poder anticiparnos a sus necesidades.

Para desglosar estos beneficios, se van a clasificar en tangibles, intangibles y estratégicos.

Beneficios tangibles

Los beneficios tangibles son aquellos que se pueden medir de alguna forma.

- Aumento de ingresos: No estimamos que este proyecto vaya a generar un aumento de ingresos. En las siguientes fases del proyecto si se espera generar acciones de marketing que:
  - Se adquieran nuevos clientes y se conviertan más mediante el uso de la segmentación.
  - Se reduzca la tasa de abandono, incrementar la fidelidad de los clientes que interesen.
  - Aumento de la satisfacción de los clientes.
- Ahorro en costes: Si se va a ahorrar costes debido a:
  - Liberación de cargas de trabajo
  - Aumento la productividad del departamento comercial.
  - Mejor seguimiento y control de determinados clientes
  - Reducción de la morosidad identificando a tiempo los clientes que van a ser potencialmente deudores

Beneficios Intangibles:

Los beneficios intangibles son aquellos que, aunque no se pueden medir, si tienen un

impacto en el negocio importante.

- Dotar a la información de clientes.
- Entender mejor las variables que son causantes de una mayor probabilidad de impago.
- Estar actualizados en tendencias de modelos y scoring.
- Ser más avanzado tecnológicamente, implica mayor adaptabilidad a los cambios de mercado.

Beneficios estratégicos:

Los beneficios estratégicos son los que nos facilitan la generación y/o consecución de las líneas estratégicas marcadas por la dirección.

- Tener un sistema de segmentación de clientes, que sea la base de mejorar las campañas de marketing y la satisfacción del cliente
- Debido a esa segmentación que abre la puerta a nuevas líneas de productos os.
- Mejora del posicionamiento de la marca debido a las dos medidas anteriores.

10. Análisis financiero

En el siguiente análisis mostraremos los cálculos y resultados de VAN y el ROI del proyecto.

Análisis de rentabilidad de proyecto				
	Año 1	Año 2	Año 3	Año 4
Inversión				
Equipos	€ 4,000.00			
Hardware	€ 30,000.00			
Software	€ 2,500.00	€ 2,500.00	€ 2,500.00	€ 2,500.00
Total de inversiones	**€ 36,500.00**	**€ 2,500.00**	**€ 2,500.00**	**€ 2,500.00**

	Año 1	Año 2	Año 3	Año 4
Ingresos / Beneficios				
Ingresos por créditos	€ 720,000.00	€ 760,000.00	€ 800,000.00	€ 840,000.00
Mejora de productividad y competitividad	€ 2,000.00	€ 2,500.00	€ 3,000.00	€ 3,500.00
Total de Ingresos	**€ 722,000.00**	**€ 762,500.00**	**€ 803,000.00**	**€ 843,500.00**

	Año 1	Año 2	Año 3	Año 4
Gastos				
Consultoría y auditoria	€ 50,000.00	€ 50,000.00	€ 50,000.00	€ 50,000.00
Costo de contacto	€ 4,000.00	€ 3,500.00	€ 3,000.00	€ 2,500.00
Total de gastos	**€ 54,000.00**	**€ 53,500.00**	**€ 53,000.00**	**€ 52,500.00**

Flujo de caja operativo	**€ 668,000.00**	**€ 709,000.00**	**€ 750,000.00**	**€ 791,000.00**

VAN	**$2,296,972.88**
Tir	**1524%**

Figura 41. Análisis Económico - Financiero. Elaboración propia.

El rubro de equipo se compone de los siguientes rubros:

- Equipos: se estima que vamos a requerir un científico de datos y un ingeniero de datos este rubro es para ejecutar las compras de los equipos.
- Hardware: estimamos que se requiere invertir en una solución de base de datos y en un servidor que nos funcione para generar ambientes de pruebas.
- Software: estimamos que se requiere licenciar una herramienta de visualización como Power Bi y también una suscripción a Cloudera con el fin de entrenar modelos o cuando se requiera más recursos de los que se tienen en sitio.

Respecto a los ingresos se constituye de la siguiente manera:

Actualmente se realizan aproximadamente 4000 contactos mensuales, de los cuales menos del 10% se logra firmar con un contrato prendario.

Estimamos que con nuestro modelo lograremos identificar con más exactitud estos clientes. Por lo tanto, estamos estimando que para el primer año lograremos fidelizar el

9.5 de los clientes.

Para el segundo año la proyección es generar contratos prendarios con al menos 10 % de los clientes. En el caso del tercer y cuarto año con el apoyo de los entrenamientos del modelo y depuración de los datos, esperamos mejorar 0.5 por año.
Con respecto a los contactos, se tiene identificado que los costos por contacto son de 1 euro por contacto. Esto nos lleva a que conforme el modelo se entrene y sea más eficiente, los contactos a realizar van a ser menores, porque se atacara la población meta de manera más exacta.

Respecto a los gastos se estima que vamos a requerir contar con 2 personas que nos apoyen en desarrollo y mejoras del proyecto.

Después de analizar los montos del VAN y TIR se demuestran resultados muy optimistas, Hacemos énfasis que este análisis no se consideran todas las variables del banco por lo que los cálculos realizados son estimaciones.

11. Optimización de los resultados (Sprint IV)

El proyecto realizado para BNP Paribas estableció unos objetivos de segmentación para reducir principalmente su tasa de abandono, a través de estrategias de fidelización que les permitan atraer nuevos clientes y clientes recurrentes.

Según la Ley de Pareto el 80% de los ingresos son generados por el 20% de los clientes. Lo que nos lleva a tener que realizar un mayor esfuerzo comercial y de marketing sobre ese 20% al ser los clientes que aportan la mayor parte del negocio.

Adicionalmente, los usuarios suelen recurrir en mayor medida a aquellas organizaciones a las que son fieles y, sin embargo, su satisfacción por una marca será transmitida en menor medida en la que lo haría un cliente descontento. Lo que nos muestra la importancia de gestionar y medir un plan de fidelización.

KPIs del plan de fidelización

Para el correcto desarrollo de esta estrategia, es necesario definir previamente los kpi`s que permitan medir el éxito de nuestro plan de fidelización.

- **CPS (Customer Profiability Score)**

Este indicador nos ayudará a medir la rentabilidad de un cliente en un período concreto. El CPS resulta especialmente interesante para poder establecer rankings de clientes en función de la rentabilidad obtenida para la compañía, y que resulta de gran ayuda a la hora de seleccionar qué acción comercial o campaña de marketing aplicar en un momento determinado sobre cada cliente.
Su cálculo se realiza de la siguiente manera: CPS = Ingresos totales cliente – Gastos totales cliente / Gastos totales cliente

- **LTV (Life Time Value)**

De esta métrica obtenemos la rentabilidad del cliente a lo largo del tiempo, lo cual nos es muy útil si lo comparamos con el CAC (coste de adquisición de un cliente) que deberá ser menor que el LTV, para evitar que la rentabilidad obtenida sea menor que su coste.
El conocer este ratio nos permitirá determinar posibles ventajas en la financiación como tipos de interés preferentes, comisión de apertura y cancelación, etc, que podamos realizar a nuestros clientes con el fin de mantenerlos y sabiendo sobre qué margen trabajar.
Se mide de la siguiente manera: LTV = Valor venta media x recurrencia (mes o anual) x vida media del cliente

- **Tasa de retención y tasa de deserción**

Para el caso que nos ocupa el cálculo de estas dos tasas es absolutamente necesario. En primer lugar, se debe calcular la tasa de deserción, que son aquellos clientes que nos abandonan cada año por cada 100. Conociendo la tasa de deserción, y si la entidad no fuese capaz de adquirir nuevos clientes, podríamos saber el tiempo que durarán los clientes a ese ritmo de abandono. De ahí la importancia de fidelizar a nuestros clientes, el ampliar la línea de negocio que nos permita romper la absoluta dependencia que se

tiene del sector automovilístico, en concreto de KIA. El hecho de tener un único cliente, los puede llevar a perder poder negociador y quedar supeditado a la exigencias y directrices del mismo.
En cuanto a la tasa de retención nos permite saber el porcentaje de clientes que repiten con nosotros año tras año, lo que nos permitirá establecer unos objetivos en nuestro plan de fidelización para años consecutivos.
Para el cálculo de la tasa de retención, simplemente habrá que restar 100 a la tasa de deserción.

- **NPS (Net Promoter Score)**

Esta métrica mide algo tan importante, para la fidelización de clientes como la lealtad del mismo, pudiendo prever su comportamiento cuando se realiza una acción determinada. Para ello, se deberá realizar una encuesta a los clientes, en la que se les preguntará si recomendarían la empresa a algún amigo o familiar, otorgando una puntuación de 0 a 10. Según las respuestas obtenidas podemos clasificar a los clientes en 3 grupos:

- **Promotores** para aquellos que hayan respondido 9 y 10. Estos son clientes a los que podemos considerar como leales a la marca, siendo seguramente embajadores y prescriptores potenciales.
- **Pasivos** para los que hayan respondido 7 y 8. Estos podemos considerarlos como clientes satisfechos, pero que no han llegado a un nivel de entusiasmo tal que podamos considerarlos como posibles prescriptores de la marca. Estos tienen probabilidades de que en algún momento lleguen a ser infieles y se vayan con tu competencia.
- **Detractores** para el resto de puntuaciones. Representan un peligro para la compañía ya que se trata de clientes insatisfechos que pueden dañar la imagen de la entidad (redes sociales, foros, etc.). Adicionalmente, las acciones comerciales que habría que realizar sobre este tipo de clientes para convertirlos en promotores sería muy costosa y con escaso éxito de conversión.

Para el cálculo del NPS deberemos restar al porcentaje de Promotores, el porcentaje de Detractores, y cuyo resultado estará entre (-100) y 100. Si el resultado es un número positivo, podremos considerar el nivel de lealtad o fidelidad como bueno. Y si es negativo, tienes un problema con el grado de lealtad de tus clientes y habrá que mejorar la estrategia de fidelización.

- **Tasa de conversión**

Este indicador nos permitirá conocer cuántos clientes potenciales se han convertido finalmente en clientes. El resultado obtenido nos ayudará a identificar y comprender mejor las necesidades de los clientes e ir adaptando nuestra estrategia a sus inquietudes. La tasa de conversión se obtendrá dividiendo los objetivos conseguidos entre el total de las interacciones realizadas con los usuarios.

- **Tasa de cancelación de clientes**

Tasa que mide el grado de fidelización de los usuarios con la marca a través de la web de la empresa.
Esta métrica muestra aquellos clientes que han dejado de tener una actividad habitual en nuestra web, lo que nos permitirá conocer el abandono de los clientes y los motivos de dicho abandono. Su cálculo se seleccionará el total de clientes que no han realizado

ninguna actividad u operación durante un periodo de tiempo específico, y se dividirá entre el total de usuarios.

- **Índice de satisfacción del cliente**

Resulta fundamental conocer el grado de satisfacción del cliente con respecto a la marca, con el fin de adecuar la estrategia de fidelización a los objetivos marcados. Para conocer este dato, es necesario hacer una breve encuesta tras cualquier interacción realizada con la compañía, en la que se ponderarán las respuestas a distintas preguntas sobre la satisfacción general de un producto concreto o de la marca. El resultado de la encuesta estará comprendido entre 0 y 100, y las posibles respuestas serán: Nada satisfecho, poco satisfecho, satisfecho, muy satisfecho o sin respuesta.
Este índice no sólo permitirá conocer el propio nivel de satisfacción de sus clientes, sino que podrá compararse con el del mercado, en aquellos informes que salgan al respecto, y poder obtener así cómo le perciben sus clientes frente a la competencia. El índice de satisfacción del cliente complementa al NPS, ya que diversos estudios señalan que para medir la fidelidad de los clientes es más preciso y fiable utilizar índices basados en medias de indicadores asociados a preguntas con escalas de satisfacción.

Una vez definidas las métricas a aplicar, estas deberán ir respaldadas por una serie de estrategias que ayuden en la consecución de los objetivos marcados para BNP Paribas.

Estrategias competitivas

Una vez analizada la competencia adoptaremos las siguientes ventajas competitivas que nos permitirá desmarcarnos de nuestros principales competidores:

- Diferenciación de productos o servicios

  Es una estrategia competitiva cuyo principal objetivo es que el consumidor perciba de forma diferente el producto o servicio ofrecido por una empresa, dotándolo de una cualidad única que será valorada de forma positiva por los consumidores. Se hace más énfasis en la calidad que en el precio. En nuestro caso, esa cualidad diferenciadora será un trato personalizado según los segmentos de clientes identificados, oferta de servicios complementarios e inmediatez de servicio.

- Segmentación de clientes

  Esta estrategia está enfocada hacia la venta y las conversiones. Concretamente, se encarga de la clasificación de los clientes que ya posee la empresa. Esta segmentación de clientes se encargará de optimizar el rendimiento del negocio mediante el uso de acciones específicas dirigidas a segmentos de clientes conocidos, con rasgos clave identificables. Al aplicar esta estrategia, la empresa dividirá los segmentos comerciales en los grupos de clientes que sean de su interés (tarjeta), ofreciéndoles aquellos productos bancarios que satisfagan sus necesidades y anticipándonos a las mismas. Esta estrategia nos ayudará a mejorar nuestra tasa de abandono o Churn Rate.

- Desarrollo de nuevos productos

  En esta estrategia el objetivo clave es incorporar productos nuevos al mercado en el que se compite. Si estos productos incorporan atributos que cubren las

necesidades del consumidor, mientras que los productos actuales no lo hacen, se producirá un aumento del número de clientes, nutriendo a la empresa de nuevos clientes y disminuyendo el impacto ocasionado por la tasa de abandono.

Estrategias funcionales

- En el ámbito web y de redes sociales:

  Se hacen imprescindibles, estrategias de marketing online tanto el posicionamiento SEO como SEM, entre otras estrategias más novedosas:

  - Estrategia SEO (optimización para motores de búsqueda) hace referencia a una variedad de técnicas y estrategias para mejorar el posicionamiento y la visibilidad en los resultados orgánicos de los diferentes buscadores (es decir, que no se pagará a la empresa del buscador para que nuestra página esté mejor posicionada).

    En este apartado, la entidad tendrá que trabajar sobre estos dos factores clave en el posicionamiento de la página web:

    1) **La relevancia**. El buscador reconocerá que una web es relevante cuanto más precisa sea la información contenida en la página para una búsqueda concreta, es decir, cuando responde a la duda o la pregunta que ha formulado el usuario. Para mejorar la relevancia de nuestra web, la organización deberá utilizar técnicas de SEO on site (factores internos de la propia página web que podemos cambiar): optimización de palabras clave y URL, enlace interno, estructura web, tiempos de carga más rápidos, etc
    2) **La autoridad.** Un sitio web tiene autoridad cuando es popular, y esta popularidad se mide según el número de enlaces que apuntan hacia ella (backlinks). Para mejorar la autoridad de una web se usan técnicas de SEO off site, entre las que destaca el linkbuilding.

    Debido a la constante evolución de la tecnología, el posicionamiento SEO se encuentra ante un nuevo escenario al que deberá adaptarse, ya que las búsquedas por voz han evolucionado durante los últimos años y ganado cada vez más adeptos, por el uso de búsquedas realizadas a través de dispositivos móviles. Al realizar una consulta por voz, no se leerá el primer resultado, sino que la elección del asistente dependerá de la precisión en el contenido de la página web, la búsqueda será más compleja siendo de tipo long tail en la que el lenguaje es más personal y humanizado. Y con el fin de que la empresa no se quedé atrás deberá comenzar a trabajar en este tipo de estrategias.

  - Estrategia SEM (marketing en motores de búsqueda). Hace referencia a las técnicas empleadas para mejorar el posicionamiento de nuestra web mediante el uso de anuncios pagados que aparecen en los buscadores para determinadas palabras clave. Al igual que ocurría con SEO, con la llegada de las búsquedas por voz surgen interesantes oportunidades para la publicidad por voz, ya que el modelo PPC (pay per click) tendrá que evolucionar para evitar quedar obsoleto, debido a que los usuarios no verán los anuncios pop up mostrados en pantalla. Como en cualquier

estrategia, es de vital importancia anticiparse al cambio y comenzar una estrategia de optimización adaptada a las búsquedas por voz en las campañas actuales para poder obtener una ventaja competitiva y poder ser más eficientes, evitando sufrir un fuerte impacto en el futuro.

- ROPO (Research Online Purchase Offline). Como consecuencia de los nuevos hábitos de los usuarios surgen nuevas oportunidades para que las empresas se den a conocer o fidelicen a sus clientes. Los consumidores buscan en la web, antes de realizar su proceso de compra, acudiendo a los comparadores online, que dan información de manera objetiva pudiendo comparar ofertas de diferentes empresas, ayudando en la toma de decisiones de los usuarios. Inclusive, los propios clientes de la entidad pueden acudir a este tipo de comparadores para valorar si la oferta de la compañía es lo suficientemente buena para aparecer en el ranking, en cuyo caso, el cliente tendrá la percepción de haber elegido correctamente entre las alternativas del mercado.

  La empresa deberá, por tanto, intentar aparecer en las mejores posiciones de los buscadores online para tantos productos como les resulte posible tener presencia, ya que las ventajas derivadas son muchas:

  1) **Ratios de conversión altos**: si el comparador está especializado en un sector en particular, su base de usuarios estará muy cualificada para esos productos, por lo que habrá un grado de interés alto de los usuarios para contratar.
  2) **Canal de captación de clientes eficiente** y más rentable que otros canales de marketing como SEM, Ferias, Patrocinios…
  3) **Difundimos la marca de la empresa**: mayor visibilidad de sus ofertas y productos para usuarios cualificados.
  4) **Acceso a información**: las empresas se pueden beneficiar de la información proporcionada por los comparadores para entender lo competente que pueden resultar sus propios productos, comparándolos frente a los del resto de competidores.

- Implantación de un Chatbot**:** tecnología por la cual un usuario puede mantener una conversación con un programa informático, a través de la página web o mediante el uso de una app. Esto llevará al cliente a estar atendido las 24 horas del día, y debido a los nuevos hábitos de los usuarios, ayudará a mejorar la relación con el cliente.

- Inbound Marketing**,** estrategia que combina las tendencias de posicionamiento SEO, en redes sociales, blogging, content marketing y marketing automation para atraer al consumidor ofreciéndole contenidos y experiencias de valor en un espacio libre de publicidad. Una de las estrategias más importantes dentro del performance marketing para sector bancario, son los contenidos.

  1) **Redes sociales**, nos permiten obtener un feedback inmediato de nuestro clientes y potenciales, sin estar sujetos a horarios. La empresa deberá mantener una escucha activa ante las necesidades de sus clientes y público objetivo, a través de su presencia en redes sociales.

2) **Creación y optimización de un blog**, fundamental para lograr un buen posicionamiento SEO debido al desarrollo de contenidos.
3) **Drip marketing**, es una potente herramienta para el sector bancario, que realiza envío de mails de información relevante para nuestros suscriptores intentando llegar a ellos mediante la utilización de ofertas y paquetes personalizados que ayudarán a aumentar las conversiones. Así mismo, también son muy efectivas para cerrar transacciones que quedaron incompletas.

- En el ámbito interno de la empresa.

  - CRM (Gestión de las relaciones con clientes) tarta de mejorar tres áreas básicas: gestión comercial, marketing y servicio de atención al cliente. Se debe, por tanto, a una estrategia orientada al cliente que busca mejorar las relaciones con clientes y potenciales. Como consecuencia de la utilización de herramientas CRM obtendremos una gestión comercial estructurada, potenciando la productividad en las ventas, al mismo tiempo que extraemos un conocimiento profundo del cliente, útil para plantear campañas de marketing más efectivas. Dicha estrategia potencia la fidelización y satisfacción de los clientes para conducirles a ventas cruzadas y recurrentes.

    Adicionalmente, a las herramientas de CRM en el ámbito interno de la empresa se realizarán encuestas de calidad o satisfacción del cliente (que no sólo medirá la percepción de la calidad de los servicios recibidos, sino que nos permitirá conocer las fortalezas y debilidades que presentan nuestros productos o servicios) mencionadas junto al indicador NPS en el apartado de métricas.

Estrategias corporativas

- Estrategia Global.

  - Imagen de marca: En un intento por mejorar la imagen que se tiene de las entidades financieras, un tercio del sector bancario global ha apostado por los principios de una Banca Responsable y Sostenible, iniciativa liderada por las Naciones Unidas y que representa un impulso masivo para la acción climática y la sostenibilidad.
    Los bancos se comprometieron a alinear estratégicamente sus negocios con las metas del Acuerdo de París sobre Cambio Climático y los Objetivos de Desarrollo Sostenible, y ampliar masivamente su contribución al logro de ambos pactos. A esta iniciativa se ha sumado BNP Paribas, que con este mensaje, transmite una preocupación por la sociedad y por sus clientes. Esta medida se traducirá en una mejor imagen de la marca y la generación de empatía hacia la misma.

  - Conseguir una buena experiencia del cliente: El sector financiero está viviendo un profundo cambio estructural impulsado por la irrupción de las nuevas tecnologías y la proliferación de nuevos canales de comunicación, esto ha originado la creación de un nuevo perfil de cliente que apoyado en la tecnología, adquiere nuevos hábitos, aumentando su nivel de exigencia al demandar experiencias personalizadas y únicas con las que

lograr la máxima satisfacción posible, cambiando la manera en que se relaciona e interactúa con la empresa. Dicha relación es más compleja y difícil de entender, y menos controlable y predecible por parte de la empresa, por lo que deberán desarrollarse nuevas estrategias y modelos de marketing.

En los últimos años, los clientes se ven cada vez más afectados por información y contenidos no producidos, ni controlados, por las empresas, es decir, por la influencia social, obteniendo información de los comentarios y opiniones de otros consumidores (foros, blogs, redes sociales…)

Por tanto, una experiencia personalizada, consistente y satisfactoria se ha convertido en un factor diferencial, para poder asegurar el crecimiento empresarial y alcanzar una ventaja competitiva sostenible, que permitirá no sólo fidelizar a nuestros clientes sino captar nuevos clientes. Debido a que la mayoría de los usuarios que obtienen una buena experiencia son más propensos a recomprar y recomendar la compañía, la no adopción e implementación del marketing de experiencias amenaza la supervivencia de cualquier negocio, puesto que debilita la lealtad del cliente. Aún son pocas las compañías que personalizan las experiencias de los consumidores basadas en el conocimiento que tienen del mismo por comunicaciones y relaciones anteriores y lo ajustan a sus necesidades y preferencias, por tanto, ser de los primeros en adoptar esta estrategia nos permitirá diferenciarnos de la competencia.

12. Anexos

Código ETL (Extract, transform & Load)

```
options COMPRESS=BINARY;

DATA _NULL_;
    CALL SYMPUT("mes", PUT(YEAR(intnx("Month", Today(), -1, "B")), 4.) ||
                PUT(MONTH(intnx("Month", Today(), -1, "B")), Z2.)) ;
    CALL SYMPUT("mes_ant", PUT(YEAR(intnx("Month", Today(), -2, "B")), 4.) ||
               PUT(MONTH(intnx("Month", Today(), -2, "B")), Z2.)) ;
RUN;

%put _user_;
/**********CLIENTE REPETITIVO**************/

libname cierre "\\dataserversad\Info\SAD\Cierre";
libname risk "\\dataser11versad\Info\SAD\Risk";
libname bases   "\\dataserver\SAD\Estrategia 2018\Cliente Repetitivo\Bases";
libname   rva
"\\datariesgo\DATARIESGO\RiskPlaning\Jorge\Calificacion_CNBV\Roll_Rates\Cartera\I
FRS\Base_histórica\2019\&mes.";
libname inst    "\\dataserversad\Info\MIS DB\Historica\Instancias";
libname pivot  "\\dataserversad\Info\MIS DB\Historica\Pivote";
libname cck "\\dataserversad\Info\MIS DB\Historica\CCK";

/* ordenar por fecha renombrando CCBINST*/
PROC sort data=cierre.CCBINST_MIS_&mes. out=ccbinstMJ_&mes. (drop=a rename=
(CVEMES_FINMIS=cvemes_fin)); by cve_mes; run;

/* Tomar los primeros 10 componentes del RFC para tomarlos sin homoclave */
/* 2015 fecha en la que se encuentra cotejada la variable RFC*/

data  dat_fin;
set   ccbinstMJ_&mes.;
rfc_key=substr(TITNUMERORN,1,10) ;
where cvemes_fin>=201501 and ejer=1;run;

proc sort data=dat_fin;by rfc_key;run;
proc freq data=  dat_fin  ; table   cvemes_fin /nocol norow nopercent; run;

 /*numero de veces que el cliente tuvo creditos (por rfc)*/

proc sql;
create table conteo as select rfc_key, count(rfc_key) as veces
from dat_fin group by rfc_key
order by veces desc;
quit;

proc freq data=conteo; table veces/ missing nocol norow; run;

/*agrupaciones de 'veces' al ver que los que regresan más de cuatro veces son muy
pocos*/
```

```
data conteo_1;
set conteo;
if veces=1 then num=1;
if veces=2 then num=2;
if veces=3 then num=3;
if veces>=4 then num=4;
run;

proc sort data=conteo_1;by rfc_key;run;

/* unión de tablas*/

data filtro_4;
merge dat_fin (in= a where= (rfc_key not in ("")) ) conteo_1 (in=b);
by rfc_key;
if a ;
if b then flag=1;else flag=0;
run;


proc freq data=  filtro_4  ; table   flag /nocol norow nopercent; run;


proc sort data=filtro_4;by rfc_key cvemes_fin; run;

data filtro_4_1 ;
set filtro_4;
by rfc_key;
if first.rfc_key then first_cred=1; else first_cred=0;

if veces=1 then cliente_conocido=0;
else if (veces ge 2 and first_cred ne 1) then cliente_conocido=1; * Cliente repetitivo
segundo crédito o más;
else if (veces ge 2 and first_cred eq 1) then cliente_conocido=0; * Cliente repetitivo
segundo crédito o más;

run;
proc freq data=   filtro_4_1 ; table    cliente_conocido/nocol norow ; run;


data filtro_5 (keep= erdos plazo marca_fin hit marca_cck cosecha ejer montofinal
numauto rfc_key cvemes_fin fechant_fin cve_mes fecha_ant monto_cck monto_ant
pcteng enganche_ant pctend endeud_ant
                                    NEWCOTESCO scint_ant cotenewbdc scbdc_ant
agrupado agrup_ant status stat_ant BUDTITSALAIRE salario_ant
                                    Precio_lista precio_ant veces num MARCA_UNION
persona2 numauto1);


set filtro_4;
by rfc_key;

retain tmp_fechafin;
if first.rfc_key then do;
```

```
tmp_fechafin=.;

end;
fechant_fin=tmp_fechafin;
tmp_fechafin=cvemes_fin;

run;


proc sort data=filtro_5;by rfc_key cvemes_fin; run;

 proc freq data=   filtro_5 ; table cvemes_fin   /nocol norow nopercent; run;


/*se utilizó un retain para generar las variables anteriores en relación a cada una de las
variables importantes
 así veremos a aquellos clientes que hayan regresado a solicitar un crédito, siendo
aquellos los que tengan esta
 información rellenada en sus respectivas columnas*/


data filtro_6 (keep= erdos plazo marca_fin hit marca_cck cosecha ejer montofinal
numauto rfc_key cvemes_fin
                                        fechant_fin cve_mes fecha_ant monto_cck
monto_ant pcteng enganche_ant pctend endeud_ant
                                    NEWCOTESCO scint_ant cotenewbdc scbdc_ant
agrupado agrup_ant status stat_ant BUDTITSALAIRE
                                        salario_ant Precio_lista precio_ant veces num
MARCA_UNION hit_ant persona2 numauto1);

set filtro_5;
by rfc_key;

format tmp_stat stat_ant agrup_ant tmp_agrup status $100.;


retain   tmp_monto tmp_eng tmp_end tmp_scint tmp_scbdc tmp_agrup tmp_stat
tmp_salario tmp_precio tmp_hit;
if first.rfc_key then do;
        tmp_hit=.;
        tmp_monto=.;
        tmp_eng=.;
        tmp_end=.;
        tmp_scint=.;
        tmp_scbdc=.;
        tmp_agrup="";
        tmp_stat="";
        tmp_salario=.;
        tmp_precio=.;
end;

hit_ant=tmp_hit;
tmp_hit=hit;

monto_ant=tmp_monto;
```

```
tmp_monto=monto_cck;

enganche_ant=tmp_eng;
tmp_eng=pcteng;

endeud_ant=tmp_end;
tmp_end=pctend;

scint_ant=tmp_scint;
tmp_scint=NEWCOTESCO;

scbdc_ant=tmp_scbdc;
tmp_scbdc=cotenewbdc;

agrup_ant=tmp_agrup;
tmp_agrup=agrupado;

stat_ant=tmp_stat;
tmp_stat=status;

salario_ant=tmp_salario;
tmp_salario=BUDTITSALAIRE;

precio_ant=tmp_precio;
tmp_precio=Precio_lista;


run;

/************ DIFERENCIAS ************/

/*diferencia de montos (número), enganche, endeudamiento y montos en porcentaje*/

data difs;
set filtro_6;
if monto_ant=. then dif_monto=.;
else dif_monto=monto_ant - monto_cck;

if enganche_ant=. then dif_eng=.;
else dif_eng= (PCTENG-enganche_ant)/100;

if endeud_ant=. then dif_end=.;
else dif_end= (PCTEND-endeud_ant)/100;

run;

data difs_1;
set difs;
if monto_ant=. then porc_monto=.;
else porc_monto=(100*monto_cck/monto_ant)-100;
run;

/*SEGMENTACIÓN DE COTAS*/

DATA cote_scorei;
```

```
set difs_1;
format NEWCOTESCO cote_int $15.;
if NEWCOTESCO=0 then cote_int="0";
else if NEWCOTESCO=1 then cote_int="1";
else if NEWCOTESCO=2 or NEWCOTESCO=3 or NEWCOTESCO=4  then
cote_int="2-4";
else if NEWCOTESCO=5 or NEWCOTESCO=6 then cote_int="5-6";
else if NEWCOTESCO=7 or NEWCOTESCO=8 or NEWCOTESCO=9 then cote_int="7-
9";


format scint_ant cote_int_ant $5.;
if scint_ant=. then cote_int_ant="";
else if scint_ant=0 then cote_int_ant="0";
else if scint_ant=1 then cote_int_ant="1";
else if scint_ant=2 or scint_ant=3 or scint_ant=4  then cote_int_ant="2-4";
else if scint_ant=5 or scint_ant=6 then cote_int_ant="5-6";
else if scint_ant=7 or scint_ant=8 or scint_ant=9 then cote_int_ant="7-9";

run;

/*SCORE INT*/

data dif_score_int;
set cote_scorei;

format scint_ant cote_int_ant dif_scint  $15.;

if cote_int_ant="" then dif_scint="";

else if cote_int_ant="0" and cote_int="0" then dif_scint="0->0";
else if cote_int_ant="0" and cote_int="1" then dif_scint="0->1";
else if cote_int_ant="0" and cote_int="2-4" then dif_scint="0->2-4";
else if cote_int_ant="0" and cote_int="5-6" then dif_scint="0->5-6";
else if cote_int_ant="0" and cote_int="7-9" then dif_scint="0->7-9";

else if cote_int_ant="1" and cote_int="0" then dif_scint="1->0";
else if cote_int_ant="1" and cote_int="1" then dif_scint="1->1";
else if cote_int_ant="1" and cote_int="2-4" then dif_scint="1->2-4";
else if cote_int_ant="1" and cote_int="5-6" then dif_scint="1->5-6";
else if cote_int_ant="1" and cote_int="7-9" then dif_scint="1->7-9";

else if cote_int_ant="2-4" and cote_int="0" then dif_scint="2-4->0";
else if cote_int_ant="2-4" and cote_int="1" then dif_scint="2-4->1";
else if cote_int_ant="2-4" and cote_int="2-4" then dif_scint="2-4->2-4";
else if cote_int_ant="2-4" and cote_int="5-6" then dif_scint="2-4->5-6";
else if cote_int_ant="2-4" and cote_int="7-9" then dif_scint="2-4->7-9";

else if cote_int_ant="5-6" and cote_int="0" then dif_scint="5-6->0";
else if cote_int_ant="5-6" and cote_int="1" then dif_scint="5-6->1";
else if cote_int_ant="5-6" and cote_int="2-4" then dif_scint="5-6->2-4";
else if cote_int_ant="5-6" and cote_int="5-6" then dif_scint="5-6->5-6";
else if cote_int_ant="5-6" and cote_int="7-9" then dif_scint="5-6->7-9";

else if cote_int_ant="7-9" and cote_int="0" then dif_scint="7-9->0";
```

```
else if cote_int_ant="7-9" and cote_int="1" then dif_scint="7-9->1";
else if cote_int_ant="7-9" and cote_int="2-4" then dif_scint="7-9->2-4";
else if cote_int_ant="7-9" and cote_int="5-6" then dif_scint="7-9->5-6";
else if cote_int_ant="7-9" and cote_int="7-9" then dif_scint="7-9->7-9";
run;

/*DIFERENCIA DE FECHAS*/

data  fecha_format_fin;

set dif_score_int  ;

format fecha_fin fechante_fin date9.;

fecha_fin=MDY(substr(put(cvemes_fin,$6.),5,2),1,substr(put(cvemes_fin,$6.),1,4));
fechante_fin=MDY(substr(put(fechant_fin, $6.),5,2),1,substr(put(fechant_fin,$6.),1,4));


run;

data difs_3;
set fecha_format_fin;
dif_fecha_fin=intck('month',fechante_fin, fecha_fin);
run;

data enganche;
set difs_3;
enganche=PCTENG/100;
RUN;

data info;
set enganche;
if 0<=enganche<10 then PCTENG_1=10;
else if enganche=>10 and enganche<20 then PCTENG_1=20;
else if enganche=>20 and enganche<30 then PCTENG_1=30;
else if enganche=>30 and enganche<40 then PCTENG_1=40;
else if enganche=>40 and enganche<50 then PCTENG_1=50;
else if enganche=>50 then PCTENG_1=100;
run;

proc sort data= rva.DETALLE_OFICIAL out=DETALLE_OFICIAL ; by  erdos fecha
;run;

data  last_erdos (keep= erdos fecha cartera rename=fecha=last_fecha) ;
set  DETALLE_OFICIAL   ;
by erdos;
if last.erdos;
run;

proc sort data= last_erdos ; by  erdos;run;

data dec1 ;
set info;
where veces>1;
run;
```

```
proc freq data= dec1   ; table veces*num   /nocol norow nopercent; run;

proc sort data= dec1 ; by rfc_key fecha_fin  ;run;

data   dec1_first ;
set    dec1 ;
by rfc_key;
if first.rfc_key;
run;

proc contents data=  dec1  ;run;
proc sort data= dec1_first ; by  erdos ;run;

data     base_last;
merge   dec1_first (IN=a )  last_erdos (in=b) ;
by   erdos ;
if a;
if b then flag=1;else flag=0;
run;

proc sort data=info out=info_2;by rfc_key cvemes_fin; where num>=2; run;

data dec2 (keep= BC_AGE_OLD_OFF_ALL_2
BC_NUM_SAT_OFF_A_2
BDCBCACCLAAMNPA_2
BDCBCACCLAAMWPA_2
BDCBCACCLANUMNPA_2
BDCBCACCLANUMWPA_2
BDCBCACREVAMNPA_2
BDCBCACREVAMWPA_2
BDCBCACREVNUMNPA_2
BDCBCACREVNUMWPA_2
BDCBCCL12MNEW_2
BDCBCCLADUENUM12M_2
BDCBCCOMDUENUM_2
BDCBCCOMNUM_2
BDCBCCONSULAM30D_2
BDCBCCONSULNUM30D_2
BDCBCCONSULNUM60D_2
BDCBCINACREV_2
BDCBCLOSCLAAM12M_2
BDCBCLOSCLANUM12M_2
BDCBCLOSREVNUM12M_2
BDCBCREVDUENUM12M_2
BDCBCRS13_2
MAX_CREDIT_LIMIT_2
MEXBCCL24M_2
MEXBCCL3M_2
MEXBCCLNR_2
TITANCIENNETEEMPL_2
erdos_2 rfc_key status_2 NEWCOTESCO_2 PCTENG_2 cotenewbdc_2
BUDTITSALAIRE_2 Precio_lista_2 montofinal_2 plazo_2 PCTEND_2);
set info_2;
by rfc_key;
if last.rfc_key;
```

```
rename erdos=erdos_2 status=status_2 NEWCOTESCO=NEWCOTESCO_2
                    PCTENG=PCTENG_2 PCTEND=PCTEND_2
cotenewbdc=cotenewbdc_2 BUDTITSALAIRE=BUDTITSALAIRE_2
Precio_lista=Precio_lista_2
montofinal=montofinal_2 plazo=plazo_2 marca_fin=marca_fin_2
BC_AGE_OLD_OFF_ALL=BC_AGE_OLD_OFF_ALL_2
BC_NUM_SAT_OFF_A=BC_NUM_SAT_OFF_A_2
BDCBCACCLAAMNPA=BDCBCACCLAAMNPA_2
BDCBCACCLAAMWPA=BDCBCACCLAAMWPA_2
BDCBCACCLANUMNPA=BDCBCACCLANUMNPA_2
BDCBCACCLANUMWPA=BDCBCACCLANUMWPA_2
BDCBCACREVAMNPA=BDCBCACREVAMNPA_2
BDCBCACREVAMWPA=BDCBCACREVAMWPA_2
BDCBCACREVNUMNPA=BDCBCACREVNUMNPA_2
BDCBCACREVNUMWPA=BDCBCACREVNUMWPA_2
BDCBCCL12MNEW=BDCBCCL12MNEW_2
BDCBCCLADUENUM12M=BDCBCCLADUENUM12M_2
BDCBCCOMDUENUM=BDCBCCOMDUENUM_2
BDCBCCOMNUM=BDCBCCOMNUM_2
BDCBCCONSULAM30D=BDCBCCONSULAM30D_2
BDCBCCONSULNUM30D=BDCBCCONSULNUM30D_2
BDCBCCONSULNUM60D=BDCBCCONSULNUM60D_2
BDCBCINACREV=BDCBCINACREV_2
BDCBCLOSCLAAM12M=BDCBCLOSCLAAM12M_2
BDCBCLOSCLANUM12M=BDCBCLOSCLANUM12M_2
BDCBCLOSREVNUM12M=BDCBCLOSREVNUM12M_2
BDCBCREVDUENUM12M=BDCBCREVDUENUM12M_2
BDCBCRS13=BDCBCRS13_2
MAX_CREDIT_LIMIT=MAX_CREDIT_LIMIT_2
MEXBCCL24M=MEXBCCL24M_2
MEXBCCL3M=MEXBCCL3M_2
MEXBCCLNR=MEXBCCLNR_2
TITANCIENNETEEMPL=TITANCIENNETEEMPL_2
;
run;

proc sort data=dec2;by rfc_key ; run;
proc sort data=base_last;by rfc_key ; run;

data decision;
merge base_last (in= a) dec2 (in=b);
by rfc_key;
if a and b;
run;

/* Porcentaje de liquidación del primer crédito a la financiacion de su segundo crédito*/

data liquidados;
set decision;
porc_liquid=((montofinal-cartera)/montofinal)*100;
run;


proc freq data=liquidados   ; table  porc_liquid ;run;
```

```
data liquidados_2;
set liquidados;
format porc_liquid_0 $10.;
if porc_liquid<10 then porc_liquid_0="'0-10";
        else if porc_liquid=>10 and porc_liquid<20 then porc_liquid_0="'10-20";
        else if porc_liquid=>20 and porc_liquid<30 then porc_liquid_0="'20-30";
        else if porc_liquid=>30 and porc_liquid<40 then porc_liquid_0="'30-40";
        else if porc_liquid=>40 and porc_liquid<60 then porc_liquid_0="'40-60";
        else if porc_liquid=>60 then porc_liquid_0="'60-100";
        run;

data   tiempo ;
set    liquidados_2 ;
format fecha_segundo_credito fecha_last date9.;
fecha_segundo_credito=MDY(substr(put(cve_mes_2,$6.),5,2),1,substr(put(cve_mes_2,
$6.),1,4));
fecha_last=MDY(substr(put(last_fecha,$6.),5,2),1,substr(put(last_fecha,$6.),1,4));
run;

data DIF_FECHA ;
set    tiempo ;
if fecha_segundo_credito>=fecha_last then
TIEMPO_REGRESO=intck("month",fecha_last,fecha_segundo_credito);
else TIEMPO_TRANSCURRIDO=intck("month",fecha_fin,fecha_segundo_credito);
RUN;

proc freq data= DIF_FECHA  ; table TIEMPO_REGRESO*porc_liquid_0 ;run;
data DIF_FECHAS;
set DIF_FECHA;
format TIEMPO_REGRESO_0 $10.;
if 0<=TIEMPO_REGRESO<=6 then TIEMPO_REGRESO_0="'0-6";
        else if TIEMPO_REGRESO>6 and TIEMPO_REGRESO<=12 then
TIEMPO_REGRESO_0="'6-12";
        else if TIEMPO_REGRESO>12 and TIEMPO_REGRESO<=24 then
TIEMPO_REGRESO_0="'12-24";
        else if TIEMPO_REGRESO>24 and TIEMPO_REGRESO<=36 then
TIEMPO_REGRESO_0="'24-36";
        else if TIEMPO_REGRESO>36 then TIEMPO_REGRESO_0="'>36";
        run;

data BASE_DECISION;
set DIF_FECHAS;
format TIEMPO_TRANSCURRIDO_0 $10.;
if 0<=TIEMPO_TRANSCURRIDO<=6 then TIEMPO_TRANSCURRIDO_0="'0-6";
        else if TIEMPO_TRANSCURRIDO>6 and TIEMPO_TRANSCURRIDO<=12
then TIEMPO_TRANSCURRIDO_0="'6-12";
        else if TIEMPO_TRANSCURRIDO>12 and TIEMPO_TRANSCURRIDO<=24
then TIEMPO_TRANSCURRIDO_0="'12-24";
        else if TIEMPO_TRANSCURRIDO>24 and TIEMPO_TRANSCURRIDO<=36
then TIEMPO_TRANSCURRIDO_0="'24-36";
        else if TIEMPO_TRANSCURRIDO>36 then
TIEMPO_TRANSCURRIDO_0="'>36";
        run;
```

```
/*Cruce Risk*/

proc sort data=risk.RISK_MIS_&mes. out=risk_1 (keep=erdos max_ret rename=
max_ret=max_ret_1); by erdos; run;
proc sort data=risk.RISK_MIS_&mes. out=risk_2 (keep=erdos max_ret rename=
(max_ret=max_ret_2 erdos=erdos_2)); by erdos; run;

Proc sort data=BASE_DECISION  ; by erdos ; run;

data BASE_DECISION_1;
merge BASE_DECISION  (in=a) risk_1 (in=b);
by erdos;
if a and b;
run;

proc sort data= BASE_DECISION_1 ; by erdos_2  ;run;

data BASE_DECISION_2;
merge BASE_DECISION_1  (in=a) risk_2 (in=b);
by erdos_2;
if a and b;
run;

proc contents data=  BASE_DECISION_2  ;run;

/*Enmascarar Datos*/

proc sort data= BASE_DECISION_2 ; by rfc_key  ;run;

data   base_final (keep= BC_AGE_OLD_OFF_ALL_2
BC_NUM_SAT_OFF_A_2
BDCBCACCLAAMNPA_2
BDCBCACCLAAMWPA_2
BDCBCACCLANUMNPA_2
BDCBCACCLANUMWPA_2
BDCBCACREVAMNPA_2
BDCBCACREVAMWPA_2
BDCBCACREVNUMNPA_2
BDCBCACREVNUMWPA_2
BDCBCCL12MNEW_2
BDCBCCLADUENUM12M_2
BDCBCCOMDUENUM_2
BDCBCCOMNUM_2
BDCBCCONSULAM30D_2
BDCBCCONSULNUM30D_2
BDCBCCONSULNUM60D_2
BDCBCINACREV_2
BDCBCLOSCLAAM12M_2
BDCBCLOSCLANUM12M_2
BDCBCLOSREVNUM12M_2
BDCBCREVDUENUM12M_2
BDCBCRS13_2
MAX_CREDIT_LIMIT_2
MEXBCCL24M_2
MEXBCCL3M_2
```

MEXBCCLNR_2
TITANCIENNETEEMPL_2
BC_AGE_OLD_OFF_ALL
BC_NUM_SAT_OFF_A
BDCBCACCLAAMNPA
BDCBCACCLAAMWPA
BDCBCACCLANUMNPA
BDCBCACCLANUMWPA
BDCBCACREVAMNPA
BDCBCACREVAMWPA
BDCBCACREVNUMNPA
BDCBCACREVNUMWPA
BDCBCCL12MNEW
BDCBCCLADUENUM12M
BDCBCCOMDUENUM
BDCBCCOMNUM
BDCBCCONSULAM30D
BDCBCCONSULNUM30D
BDCBCCONSULNUM60D
BDCBCINACREV
BDCBCLOSCLAAM12M
BDCBCLOSCLANUM12M
BDCBCLOSREVNUM12M
BDCBCREVDUENUM12M
BDCBCRS13
MAX_CREDIT_LIMIT
MEXBCCL24M
MEXBCCL3M
MEXBCCLNR
TITANCIENNETEEMPL
BUDTITSALAIRE
BUDTITSALAIRE_2
HIT
NEWCOTESCO
NEWCOTESCO_2
PCTEND
PCTEND_2
PCTENG
PCTENG_2
Precio_lista
Precio_lista_2
cotenewbdc
cotenewbdc_2
llave
max_ret_1
max_ret_2
montofinal
montofinal_2
persona2
plazo
plazo_2
porc_liquid
rfc_key

);

```
set   BASE_DECISION_2  ;
llave + 1;
  if first.rfc_key then llave = 1;
run;

proc contents data=  base_final  ;run;
```

Diccionario de variables

Variable	Descripción
BC_AGE_OLD_OFF_ALL	Experiencia en Buró de Crédito
BC_AGE_OLD_OFF_ALL_2	Experiencia en Buró de Crédito 2 credito
BC_NUM_SAT_OFF_A	Cuentas satisfactorias de Auto
BC_NUM_SAT_OFF_A_2	Cuentas satisfactorias de Auto 2 credito
BDCBCACCLAAMNPA	El importe total pendiente créditos activos tipo clasico mal pagado
BDCBCACCLAAMNPA_2	El importe total pendiente créditos activos tipo clasico mal pagado 2 credito
BDCBCACCLAAMWPA	El importe total pendiente créditos activos tipo clasico bien pagado
BDCBCACCLAAMWPA_2	El importe total pendiente créditos activos tipo clasico bien pagado 2 credito
BDCBCACCLANUMNPA	El numero de créditos activo tipo clasico mal pagado
BDCBCACCLANUMNPA_2	El numero de créditos activo tipo clasico mal pagado 2 credito
BDCBCACCLANUMWPA	El numero de créditos activo tipo clasico bien pagado
BDCBCACCLANUMWPA_2	El numero de créditos activo tipo clasico bien pagado 2 credito
BDCBCACREVAMNPA	El importe total pendiente créditos activos tipo tarjeta mal pagado
BDCBCACREVAMNPA_2	El importe total pendiente créditos activos tipo tarjeta mal pagado 2 credito
BDCBCACREVAMWPA	El importe total pendiente créditos activos tipo tarjeta bien pagado
BDCBCACREVAMWPA_2	El importe total pendiente créditos activos tipo tarjeta bien pagado 2 credito
BDCBCACREVNUMNPA	El numero de créditos activos tipo tarjeta mal pagado
BDCBCACREVNUMNPA_2	El numero de créditos activos tipo tarjeta mal pagado 2 credito
BDCBCACREVNUMWPA	El numero de créditos activos tipo tarjeta bien pagado
BDCBCACREVNUMWPA_2	El numero de créditos activos tipo tarjeta bien pagado 2 credito
BDCBCCL12MNEW	Peor MOP actual
BDCBCCL12MNEW_2	Peor MOP actual 2 credito

BDCBCCLADUENUM12M	El numero de créditos tipo clasico vencido en los 12 ultimos meses
BDCBCCLADUENUM12M_2	El numero de créditos tipo clasico vencido en los 12 ultimos meses 2 credito
BDCBCCOMDUENUM	El importe total pendiente de deudas de tipo comunicaciones
BDCBCCOMDUENUM_2	El importe total pendiente de deudas de tipo comunicaciones 2 credito
BDCBCCOMNUM	El numero de deuda de tipo comunicaciones
BDCBCCOMNUM_2	El numero de deuda de tipo comunicaciones 2 credito
BDCBCCONSULAM30D	Numero de consulta (excepto de Cetelem) hechas en BdC en los 60 últimos días para el titular.
BDCBCCONSULAM30D_2	Numero de consulta (excepto de Cetelem) hechas en BdC en los 60 últimos días para el titular. 2 credito
BDCBCCONSULNUM30D	Numero de consulta (excepto de Cetelem) hechas en BdC en los 30 últimos días para el titular.
BDCBCCONSULNUM30D_2	Numero de consulta (excepto de Cetelem) hechas en BdC en los 30 últimos días para el titular. 2 credito
BDCBCCONSULNUM60D	Número de consultas a BDC.
BDCBCCONSULNUM60D_2	Número de consultas a BDC. 2 credito
BDCBCINACREV	El numero de créditos tipo tarjeta sin actividad
BDCBCINACREV_2	El numero de créditos tipo tarjeta sin actividad 2 credito
BDCBCLOSCLAAM12M	El importe total de perdida de creditos tipo clasico en los 12 ultimos meses
BDCBCLOSCLAAM12M_2	El importe total de perdida de creditos tipo clasico en los 12 ultimos meses 2 credito
BDCBCLOSCLANUM12M	El numero de créditos tipo clasico en perdida total o parcial en los 12 ultimos meses
BDCBCLOSCLANUM12M_2	El numero de créditos tipo clasico en perdida total o parcial en los 12 ultimos meses 2 credito
BDCBCLOSREVNUM12M	El numero de créditos tipo tarjeta en perdida total o parcial en los 12 ultimos meses
BDCBCLOSREVNUM12M_2	El numero de créditos tipo tarjeta en perdida total o parcial en los 12 ultimos meses 2 credito
BDCBCREVDUENUM12M	El numero de créditos tipo tarjeta vencido en los 12 ultimos meses
BDCBCREVDUENUM12M_2	El numero de créditos tipo tarjeta vencido en los 12 ultimos meses 2 credito
BDCBCRS13	Número de cuentas con morosidad actual
BDCBCRS13_2	Número de cuentas con morosidad actual 2 credito
BUDTITSALAIRE	Ingreso 1 Credito
BUDTITSALAIRE_2	Ingreso 2 Credito
cotenewbdc	Score Buró de crédito 1 credito
cotenewbdc_2	Score Buró de crédito 2 credito

HIT	HIT BC 1 CREDITO
llave	llave
MAX_CREDIT_LIMIT	Linea de Crédito Maximo
MAX_CREDIT_LIMIT_2	Linea de Crédito Maximo 2 credito
max_ret_1	max_ret_1
max_ret_2	max_ret_2
MEXBCCL24M	Peor MOP histórico de los últimos 24 meses
MEXBCCL24M_2	Peor MOP histórico de los últimos 24 meses 2 credito
MEXBCCL3M	Peor MOP histórico de los últimos 3 meses
MEXBCCL3M_2	Peor MOP histórico de los últimos 3 meses 2 credito
MEXBCCLNR	Peor MOP actual comunicaciones/servicios
MEXBCCLNR_2	Peor MOP actual comunicaciones/servicios 2 credito
montofinal	montofinal
montofinal_2	montofinal_2
NEWCOTESCO	Score Interno 1 credito
NEWCOTESCO_2	Score Interno 2 credito
PCTEND	Endeudamiento 1 credito
PCTEND_2	Endeudamiento 2 credito
PCTENG	Enganche 1 Credito
PCTENG_2	Enganche 2 Credito
persona2	persona2
plazo	plazo
plazo_2	plazo_2
porc_liquid	porc_liquid
Precio_lista	Precio Vehiculo 1 credito
Precio_lista_2	Precio Vehiculo 2 credito
rfc_key	RFC
TITANCIENNETEEMPL	Fecha de ingreso laboral
TITANCIENNETEEMPL_2	Fecha de ingreso laboral 2 credito

Análisis de componentes principales

```
# Se dan de alta las librerias necesarias para poder desarrollar el proyecto

import pandas as pd
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import SGDClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA as sklearnPCA

%matplotlib inline

RANDOM_SEED = 42
```

```
n_dim = 64

plt.style.use('bmh')

a = pd.read_csv("Base_eoi.csv")
print(a)

num_a = a.select_dtypes(include=['int64','float64']).copy()
print(num_a)

num_a.dropna(inplace=True)

print(num_a)


import numpy as np
from sklearn.decomposition import PCA

pca = PCA(n_components=10)

pca.fit(num_a)

numa_pca=pca.transform(num_a)

print("shape of numa_pca", numa_pca.shape)
expl = pca.explained_variance_ratio_
print(expl)
print('suma:',sum(expl[0:2]))

plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xlabel('number of components')
plt.ylabel('cumulative explained variance')
plt.show()

num_a.drop([ 'Malo','llave', 'max_ret_1','max_ret_2'], axis=1)

from factor_analyzer import FactorAnalyzer

fa = FactorAnalyzer()
fa.analyze(num_a, 2, rotation="varimax")

print(fa.loadings)

fa.loadings.to_excel('factors.xlsx', sheet_name='factors')
```

Selección de Variables

Variables	Factor1	Factor2	Selección Variables
montofinal	-13%	11%	0
BC_AGE_OLD_OFF_ALL	22%	-4%	0
BC_NUM_SAT_OFF_A	20%	-2%	0
BDCBCACCLAAMNPA	0%	76%	1
BDCBCACCLAAMWPA	28%	23%	0

BDCBCACCLANUMNPA	0%	76%	1
BDCBCACCLANUMWPA	68%	32%	1
BDCBCACREVAMNPA	-2%	11%	0
BDCBCACREVAMWPA	29%	17%	0
BDCBCACREVNUMNPA	4%	50%	1
BDCBCACREVNUMWPA	66%	17%	1
BDCBCCL12MNEW	-4%	17%	0
BDCBCCLADUENUM12M	-5%	16%	0
BDCBCCOMNUM	31%	37%	1
BDCBCCONSULAM30D	3%	2%	0
BDCBCCONSULNUM30D	28%	38%	1
BDCBCCONSULNUM60D	27%	27%	0
BDCBCINACREV	29%	5%	0
BDCBCRS13	-4%	21%	0
BUDTITSALAIRE	-7%	3%	0
MAX_CREDIT_LIMIT	24%	-4%	0
MEXBCCL24M	2%	34%	1
MEXBCCL3M	1%	11%	0
MEXBCCLNR	-8%	14%	0
PCTEND	31%	-1%	1
PCTENG	49%	-5%	1
TITANCIENNETEEMPL	72%	17%	1
Precio_lista	0%	-1%	0
plazo	2%	6%	0
cotenewbdc	70%	-25%	1
NEWCOTESCO	51%	-21%	1
HIT	73%	20%	1
montofinal_2	-12%	12%	0
BC_AGE_OLD_OFF_ALL_2	49%	-2%	1
BC_NUM_SAT_OFF_A_2	34%	9%	1
BDCBCACCLAAMNPA_2	-1%	81%	1
BDCBCACCLAAMWPA_2	18%	24%	0
BDCBCACCLANUMNPA_2	-1%	81%	1
BDCBCACCLANUMWPA_2	53%	39%	1

BDCBCACREVAMNPA_2	-1%	15%	0
BDCBCACREVAMWPA_2	28%	20%	0
BDCBCACREVNUMNPA_2	3%	54%	1
BDCBCACREVNUMWPA_2	54%	25%	1
BDCBCCLADUENUM12M_2	-5%	18%	0
BDCBCCOMNUM_2	31%	39%	1
BDCBCCONSULAM30D_2	4%	4%	0
BDCBCCONSULNUM30D_2	22%	40%	1
BDCBCCONSULNUM60D_2	25%	29%	0
BDCBCINACREV_2	27%	5%	0
BDCBCRS13_2	-6%	36%	1
BUDTITSALAIRE_2	-3%	2%	0
MAX_CREDIT_LIMIT_2	46%	-3%	1
MEXBCCL24M_2	1%	39%	1
MEXBCCL3M_2	-1%	14%	0
MEXBCCLNR_2	-8%	17%	0
PCTEND_2	30%	-3%	1
PCTENG_2	51%	-9%	1
TITANCIENNETEEMPL_2	72%	17%	1
Precio_lista_2	3%	0%	0
plazo_2	0%	6%	0
cotenewbdc_2	70%	-30%	1
NEWCOTESCO_2	56%	-19%	1
porc_liquid	-4%	-2%	0
max_ret_1	-5%	3%	0
max_ret_2	-8%	5%	0
Malo	-9%	9%	0
llave	2%	-1%	0

Codigo VARCLUS

```
from varclushi import VarClusHi
vc = VarClusHi(df1,maxeigval2=1,maxclus=None)
vc.varclus()

vc.rsquare.to_excel('varclus.xlsx', sheet_name='varclus')
```

Resultados de clúster de Variables

Cluster	Variable	RS_Own	RS_NC	RS_Ratio

0	TITANCIENNETEEMPL	83%	20%	21%
0	TITANCIENNETEEMPL_2	83%	20%	21%
0	HIT	76%	15%	28%
0	cotenewbdc_2	56%	23%	57%
0	PCTENG_2	44%	11%	63%
0	PCTENG	42%	8%	63%
0	cotenewbdc	52%	27%	65%
1	BDCBCACCLAAMNPA	78%	11%	25%
1	BDCBCACCLANUMNPA	78%	11%	25%
1	BDCBCACCLAAMNPA_2	78%	16%	27%
1	BDCBCACCLANUMNPA_2	78%	16%	27%
1	BDCBCACREVNUMNPA	49%	5%	54%
1	BDCBCACREVNUMNPA_2	49%	10%	56%
2	BDCBCACREVNUMWPA_2	81%	13%	22%
2	BDCBCACREVNUMWPA	77%	18%	27%
2	BDCBCACCLANUMWPA_2	78%	24%	29%
2	BDCBCACCLANUMWPA	77%	35%	35%
3	BDCBCCOMNUM_2	94%	19%	7%
3	BDCBCCOMNUM	94%	19%	7%
4	MAX_CREDIT_LIMIT_2	69%	10%	34%
4	BC_AGE_OLD_OFF_ALL_2	70%	16%	36%
4	BC_NUM_SAT_OFF_A_2	49%	8%	56%
5	BDCBCCONSULNUM30D	81%	13%	22%
5	BDCBCCONSULNUM30D_2	81%	13%	22%
6	MEXBCCL24M_2	63%	9%	41%
6	MEXBCCL24M	49%	8%	55%
6	BDCBCRS13_2	37%	8%	69%
7	PCTEND	73%	11%	31%
7	PCTEND_2	73%	14%	32%
8	NEWCOTESCO	83%	13%	20%
8	NEWCOTESCO_2	83%	21%	22%

Validación de Modelos

```
import sklearn as sk
from sklearn import svm
from sklearn.linear_model import LogisticRegression
import pandas as pd
import os

os.chdir('/Users/johannygonzalez.5')
heart = pd.read_csv('Base_eoi_0.csv', sep=',',header=0, low_memory=False)

heart.info()
y = heart.iloc[:,9]
X = heart.iloc[:,:9]
```

Regresion logistica

```
LR = LogisticRegression(random_state=0, solver='lbfgs', multi_class='ovr').fit(X, y)
LR.predict(X.iloc[500:,:])
round(LR.score(X,y), 4)
```

Resultado

0.0.9838

Random Forest

```
import sklearn as sk
from sklearn.ensemble import RandomForestClassifier
```

```
RF = RandomForestClassifier(n_estimators=100, max_depth=2, random_state=0)

RF.fit(X, y)

RF.predict(X.iloc[500:,:])

round(RF.score(X,y), 4)
```

Resultado

1

Neural Networks

```
import sklearn as sk

from sklearn.neural_network import MLPClassifier

NN = MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(5, 2), random_state=1)

NN.fit(X, y)

NN.predict(X.iloc[500:,:])

round(NN.score(X,y), 4)
```

Resultado
0.9838

Information Value

```
# import packages
import pandas as pd
import numpy as np
import pandas.core.algorithms as algos
from pandas import Series
import scipy.stats.stats as stats
import re
import traceback
import string
```

```
max_bin = 20
force_bin = 3

# define a binning function
def mono_bin(Y, X, n = max_bin):

    df1 = pd.DataFrame({"X": X, "Y": Y})
    justmiss = df1[['X','Y']][df1.X.isnull()]
    notmiss = df1[['X','Y']][df1.X.notnull()]
    r = 0
    while np.abs(r) < 1:
        try:
            d1 = pd.DataFrame({"X": notmiss.X, "Y": notmiss.Y, "Bucket": pd.qcut(notmiss.X,
n)})
            d2 = d1.groupby('Bucket', as_index=True)
            r, p = stats.spearmanr(d2.mean().X, d2.mean().Y)
            n = n - 1
        except Exception as e:
            n = n - 1

    if len(d2) == 1:
        n = force_bin
        bins = algos.quantile(notmiss.X, np.linspace(0, 1, n))
        if len(np.unique(bins)) == 2:
            bins = np.insert(bins, 0, 1)
            bins[1] = bins[1]-(bins[1]/2)
        d1 = pd.DataFrame({"X": notmiss.X, "Y": notmiss.Y, "Bucket": pd.cut(notmiss.X,
np.unique(bins),include_lowest=True)})
        d2 = d1.groupby('Bucket', as_index=True)

    d3 = pd.DataFrame({},index=[])
    d3["MIN_VALUE"] = d2.min().X
    d3["MAX_VALUE"] = d2.max().X
    d3["COUNT"] = d2.count().Y
    d3["EVENT"] = d2.sum().Y
    d3["NONEVENT"] = d2.count().Y - d2.sum().Y
    d3=d3.reset_index(drop=True)

    if len(justmiss.index) > 0:
        d4 = pd.DataFrame({'MIN_VALUE':np.nan},index=[0])
        d4["MAX_VALUE"] = np.nan
        d4["COUNT"] = justmiss.count().Y
        d4["EVENT"] = justmiss.sum().Y
        d4["NONEVENT"] = justmiss.count().Y - justmiss.sum().Y
        d3 = d3.append(d4,ignore_index=True)

    d3["EVENT_RATE"] = d3.EVENT/d3.COUNT
    d3["NON_EVENT_RATE"] = d3.NONEVENT/d3.COUNT
    d3["DIST_EVENT"] = d3.EVENT/d3.sum().EVENT
    d3["DIST_NON_EVENT"] = d3.NONEVENT/d3.sum().NONEVENT
    d3["WOE"] = np.log(d3.DIST_EVENT/d3.DIST_NON_EVENT)
    d3["IV"] = (d3.DIST_EVENT-
d3.DIST_NON_EVENT)*np.log(d3.DIST_EVENT/d3.DIST_NON_EVENT)
    d3["VAR_NAME"] = "VAR"
```

```
    d3 = d3[['VAR_NAME','MIN_VALUE', 'MAX_VALUE', 'COUNT', 'EVENT',
'EVENT_RATE', 'NONEVENT', 'NON_EVENT_RATE',
'DIST_EVENT','DIST_NON_EVENT','WOE', 'IV']]
    d3 = d3.replace([np.inf, -np.inf], 0)
    d3.IV = d3.IV.sum()

    return(d3)

def char_bin(Y, X):

    df1 = pd.DataFrame({"X": X, "Y": Y})
    justmiss = df1[['X','Y']][df1.X.isnull()]
    notmiss = df1[['X','Y']][df1.X.notnull()]
    df2 = notmiss.groupby('X',as_index=True)

    d3 = pd.DataFrame({},index=[])
    d3["COUNT"] = df2.count().Y
    d3["MIN_VALUE"] = df2.sum().Y.index
    d3["MAX_VALUE"] = d3["MIN_VALUE"]
    d3["EVENT"] = df2.sum().Y
    d3["NONEVENT"] = df2.count().Y - df2.sum().Y

    if len(justmiss.index) > 0:
        d4 = pd.DataFrame({'MIN_VALUE':np.nan},index=[0])
        d4["MAX_VALUE"] = np.nan
        d4["COUNT"] = justmiss.count().Y
        d4["EVENT"] = justmiss.sum().Y
        d4["NONEVENT"] = justmiss.count().Y - justmiss.sum().Y
        d3 = d3.append(d4,ignore_index=True)

    d3["EVENT_RATE"] = d3.EVENT/d3.COUNT
    d3["NON_EVENT_RATE"] = d3.NONEVENT/d3.COUNT
    d3["DIST_EVENT"] = d3.EVENT/d3.sum().EVENT
    d3["DIST_NON_EVENT"] = d3.NONEVENT/d3.sum().NONEVENT
    d3["WOE"] = np.log(d3.DIST_EVENT/d3.DIST_NON_EVENT)
    d3["IV"] = (d3.DIST_EVENT-
d3.DIST_NON_EVENT)*np.log(d3.DIST_EVENT/d3.DIST_NON_EVENT)
    d3["VAR_NAME"] = "VAR"
    d3 = d3[['VAR_NAME','MIN_VALUE', 'MAX_VALUE', 'COUNT', 'EVENT',
'EVENT_RATE', 'NONEVENT', 'NON_EVENT_RATE',
'DIST_EVENT','DIST_NON_EVENT','WOE', 'IV']]
    d3 = d3.replace([np.inf, -np.inf], 0)
    d3.IV = d3.IV.sum()
    d3 = d3.reset_index(drop=True)

    return(d3)

def data_vars(df1, target):

    stack = traceback.extract_stack()
    filename, lineno, function_name, code = stack[-2]
    vars_name = re.compile(r'\((.*?)\).*$').search(code).groups()[0]
    final = (re.findall(r"[\w']+", vars_name))[-1]

    x = df1.dtypes.index
```

```
    count = -1

    for i in x:
        if i.upper() not in (final.upper()):
            if np.issubdtype(df1[i], np.number) and len(Series.unique(df1[i])) > 2:
                conv = mono_bin(target, df1[i])
                conv["VAR_NAME"] = i
                count = count + 1
            else:
                conv = char_bin(target, df1[i])
                conv["VAR_NAME"] = i
                count = count + 1

            if count == 0:
                iv_df = conv
            else:
                iv_df = iv_df.append(conv,ignore_index=True)

    iv = pd.DataFrame({'IV':iv_df.groupby('VAR_NAME').IV.max()})
    iv = iv.reset_index()
    return(iv_df,iv)



# Information value de las variables

final_iv, IV = data_vars(variables , variables.Malo)

final_iv

IV.sort_values('IV').to_excel('IV.xlsx', sheet_name='IV')
IV.sort_values('IV')
```