Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Escuela Profesional de Ingeniería de Sistemas

Implementación de un modelo de predicción de contratación de tarjetas de crédito para una entidad financiera española

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

Cynthia Ursula BAUTISTA ALMEZA

ASESOR

Jose Alfredo HERRERA QUISPE

Lima, Perú

2021

Referencia bibliográfica

Bautista, C. (2021). *Implementación de un modelo de predicción de contratación de tarjetas de crédito para una entidad financiera española*. [Trabajo de suficiencia profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniería de Sistemas]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Cynthia Ursula Bautista Almeza
DNI	41893856
URL de ORCID	https://orcid.org/0000-0002-3468-4522
Datos de asesor	
Nombres y apellidos	Jose Alfredo Herrera Quispe
DNI	40362859
URL de ORCID	https://orcid.org/0000-0002-8207-9714
Datos de investigación	
Línea de investigación	No Aplica.
Grupo de investigación	ITDATA
Agencia de financiamiento	Financiamiento propio.
Ubicación geográfica de la investigación	España, Madrid, Tetuán 40°27'11.0"N 3°41'35.8"W 40.453006 -3.693282
Año o rango de años en que se realizó la investigación	Año 2021
URL de disciplinas OCDE	Ingeniería de Sistemas y Comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
Escuela Profesional de Ingeniería de Sistemas

Acta Virtual de Sustentación
del Trabajo de Suficiencia Profesional

Siendo las 18:00 horas del día 22 de julio del año 2021, se reunieron virtualmente los docentes designados como Miembros de Jurado del Trabajo de Suficiencia Profesional, presidido por el Ing. Bartra More Arturo Alejandro (Presidente), Mg. Machado Vicente Joel Fernando (Miembro) y el Dr. Herrera Quispe José (Miembro Asesor), usando la plataforma Meet (https://meet.google.com/eby-qduc-tyc), para la sustentación virtual del Trabajo de Suficiencia Profesional intitulado: ***"IMPLEMENTACIÓN DE UN MODELO DE PREDICCIÓN DE CONTRATACIÓN DE TARJETAS DE CRÉDITO PARA UNA ENTIDAD FINANCIERA ESPAÑOLA",*** *por la Bachiller* ***Bautista Almeza Cynthia Ursula****; para obtener el Título Profesional de Ingeniera de Sistemas.*

Acto seguido de la exposición del Trabajo de Suficiencia Profesional, el Presidente invitó a la Bachiller a dar las respuestas a las preguntas establecidas por los miembros del Jurado.

La Bachiller en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, la Bachiller obtuvo la nota de ***19 DIECINUEVE.***

A continuación el Presidente de Jurados el Ing. Bartra More Arturo Alejandro, declara a la Bachiller ***Ingeniera de Sistemas.***

Siendo las 19:00 horas, se levantó la sesión.

Presidente
Ing. Bartra More Arturo Alejandro

Miembro
Mg. Machado Vicente Joel Fernando

Miembro Asesor
Dr. Herrera Quispe José

FICHA CATALOGRÁFICA

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

AUTOR: CYNTHIA URSULA BAUTISTA ALMEZA

ASESOR: JOSE HERRERA QUISPE

LIMA-PERÚ, 2021

Título profesional: Ingeniero de Sistemas

Línea de Investigación: Tecnologías de ciencia de datos.

Pregrado: Escuela Profesional de Ingeniería de Sistemas - Facultad de Ingeniería de Sistemas e Informática, UNMSM

Formato 28 x 20 cm Página vi, 82

DEDICATORIA

Este trabajo se lo dedico a mis padres que siempre me hanmotivado a buscar nuevos retos.

Y a mi esposo, que me ha apoyado a obtener este nuevo hito en mi carrera profesional.

AGRADECIMIENTO

Agradezco a la empresaBluetab, empresa donde se desarrolló esta experiencia profesional.Gracias por darme la oportunidad de formarme en el área de Ciencia de Datos.

Finalmente quiero agradecer al profesor JoseHerrera Quispe que me ha guiado en la realización de este trabajo.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

IMPLEMENTACIÓN DE UN MODELO DE PREDICCIÓN DE CONTRATACIÓN DE TARJETAS DE CRÉDITO EN UNA ENTIDAD FINANCIERA ESPAÑOLA

Autor:	Bautista Almeza, Cynthia Ursula
Asesor:	Herrera Quispe, Jose
Título:	Trabajo de Suficiencia Profesional
Fecha:	Julio 2021

RESUMEN

El presente trabajo de suficiencia profesional describe la implementación de un modelo predictivo de contratación de tarjetas de crédito para una entidad financiera española desarrollado en el año 2019. El principal objetivo de la solución fue identificar los clientes propensos a contratar tarjetas de crédito de la entidad para enfocar las acciones comerciales a estos clientes y así conseguir un incremento en las ventas del producto, aumentar la efectividad de las campañas comerciales. Teniendo en cuenta el tipo de problema, se seleccionó la técnica de aprendizaje supervisado para la implementación del modelo de clasificación. Para la construcción del modelo se seleccionaron tres algoritmos: *Árbol de Decisión, Random Forest y GradientBoosting*. Posterior a la validación de los tres modelos, se seleccionó el algoritmo de *Random Forest.*

Como resultado de la aplicación del modelo predictivo en las acciones comerciales, se obtuvo una mejora en el 15% de la efectividad de las campañas comerciales y un incremento en las ventas de tarjetas de crédito de la entidad.

Palabras Clave: Machine Learning, CRISP-DM, Aprendizaje Supervisado, Clasificación. Árbol de Decisión, Random Forest, GradientBoosting, Tarjetas de Crédito, Campañas Comerciales.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

IMPLEMENTATION OF A CREDIT CARD HIRING PREDICTION MODEL IN A SPANISH FINANCIAL INSTITUTION

Author:	Bautista Almeza, Cynthia Ursula
Advisor:	Herrera Quispe, Jose
Title:	Professional SufficiencyWork
Date:	July 2021

ABSTRACT

Thisprofessionalexperiencereport describes theimplementationof a predictivemodelforcreditcardcontractsfor a spanishfinancialorganization. Themainobjectiveofthesolutionistoidentifycustomerspronetotakeoutcreditcardsofthee ntitytofocuscommercialactionsonthesecustomers and thusachieveanincrease in product sales, increasingtheeffectivenessofcommercialcampaigns. Takingintoaccountthetypeofsolutionto be implemented, thesupervisedlearningtechniquewasselectedfortheimplementationoftheclassificationm odel. Fortheconstructionofthemodel, threealgorithmswereselected: DecisionTree, Random Forest and GradientBoosting. After thevalidationofthethreemodels, theRandom Forest algorithmwasselected.

As a resultoftheapplicationofthepredictivemodel in commercialactions, therewas a 15% improvement in theeffectivenessofcommercialcampaigns and anincrease in theentity'screditcard sales.

Keywords: Machine Learning, CRISP-DM, SupervisedLearning, Classification, DecisionTree, Random Forest, GradientBoosting, CreditCards, CommercialCampaigns.

TABLA DE CONTENIDO

LISTADO DE FIGURAS

LISTADO DE TABLAS

INTRODUCCIÓN

El presente trabajo de suficiencia profesional describela implementación de un modelo de predicción de contratación de tarjetas de crédito para una entidad financiera española desarrollado en el año 2019.

En la actualidad existe un gran auge en los proyectos de *Machine Learning* en distintas organizaciones, sobre todo en las entidades financieras como es el caso de este proyecto. Lamentablemente no todos los proyectos de *Machine Learning* suelen desarrollarse correctamente y obtener los resultados esperados por la organización. Es por ello que la elección de una metodología especializada en proyectos de *Machine Learning*es primordial para el éxito de proyectos de esta índole. En el trabajo se detallarán todos los pasos que se han seguido para la implantación de la solución aplicando la metodología de CRISP-DM (*Cross Industry Standard Processfor Data Mining*), así como también los resultados obtenidos y el impacto que tuvo la solución dentro dela entidad.

Este trabajo se divide en cinco capítulos, los cuales se describen brevemente a continuación:

En elcapítulo I: Se detalla la trayectoria profesional del autor,que se centra principalmente en proyectos de consultoría. En este apartado se detallan funciones y experiencias adquiridas en cada cargo desempeñado. A nivel formativo se incluyen las formaciones, cursos y certificaciones realizadas por el autor.

En el capítulo II: Se resume el contexto en el que se desarrolla la experiencia profesional.Se describe la empresa donde se desarrolla la experiencia y las funciones que el autor desarrolla en ella.

Enelcapítulo III: Se detalla la situación problemática y la solución planteada. También se precisan las actividades desarrolladas en la implementación de la solución.

En el capítulo IV: Se describe la reflexión crítica en base a la experiencia del autor y sus aportes en la implementación de la solución y finalmente el trabajo termina con las conclusiones y recomendaciones propuestas por el autor.

CAPÍTULO I - TRAYECTORIA PROFESIONAL

Presentación Profesional

La autora de este trabajo de Experiencia Profesional es Bachiller en Ingeniería de Sistemas e Informática con dieciséis años de experiencia en el análisis y desarrollo de proyectos, con amplios conocimientos técnicos y funcionales. Su experiencia se centra mayormente en proyectos en el ámbito bancario. Persona proactiva orientada a resultados, con especial foco en la entrega y en la satisfacción del cliente final.

El objetivo de la autora consiste en asumir nuevos retos y responsabilidades aportando experiencia y capacidad de adaptación a nuevos entornos de desarrollo de nuevas soluciones.

Experiencia Profesional

Abril 2019 – Actualidad. **BluetabSolutions, Madrid (España)**

Cargo: *FunctionalTechnician*

Proyecto: *Global DataHub- Client Solutions*– Modelo de Datos GlobalOmnicanal

Funciones:

- Participación en un proyecto que tiene como objetivo diseñar un modelo de datos global con información multicanal. Este modelo permitirá construir casos de uso sobre el *journey* del cliente en base a diversas herramientas de marketing digital: *Adobe Analytics, Salesforce Marketing Cloud, Google Adwords*.
- Diseño de modelos de datos y posterior diseño físico en base de datos en *Big Data*.
- Comprensión detallada de las fuentes orígenes para su explotación analítica.
- Definición criterios de aceptación y validación de ingestas.
- Coordinación con diferentes departamentos ubicados en distintas geografías para la implantación de la solución.

Cargo: *FunctionalTechnician*

Proyecto: *AdvanceAnalytics- Client Solutions* – Modelo Predictivo de Contratación

Funciones:

- Participación en un proyecto que tiene como objetivo construir un modelo de predicción para identificar los clientes propensos a la contratación de una tarjeta de crédito.
- Análisis y preparación de las fuentes orígenes para la construcción del modelo de clasificación de aprendizaje supervisado.

- Selección de algoritmo más óptimo para alcanzar los objetivos del proyecto.
- Entrenamiento del modelo e interpretación delos resultados.

Septiembre 2015 – Marzo 2019. **Everis, Madrid (España)**

Cargo:*Team Leader*

Proyecto: Fusión de Sociedad de Valores con Entidad Financiera

Funciones:

- Participación en el proyecto de fusión de la Sociedad de Valores con la Entidad Financiera y migración de módulos del sistema de *BackOffice* (de AS/400 a sistemas distribuidos).
- Diseño de modelo de datos para la migración de datos de sistemas en AS/400.
- Relación con el usuario para la toma de requisitos y coordinación de pruebas UAT.
- Diseño de procesos y definición de los planes de prueba de los nuevos procesos.
- Gestión de la relación entre el equipo proyecto y el laboratorio encargado de realizar los desarrollos.

Cargo: *Team Leader*

Proyecto: *SecuritiesCIB* -Sistema Informacional *Securities*

- Diseño y mantenimiento del modelo de datos del *Datawarehouse*.
- Diseño e implementación de procesos de ETL encargados de extraer los datos de los sistemas operacionales (*Middle y BackOffice*) para integrarlos en el *Datawarehouse*.
- Generación de informes a través de la herramienta *MicroStrategy*, explotados por los usuarios de negocio.
- Relación con el usuario para la toma de requisitos y coordinación de pruebas UAT.
- Gestión del proyecto aplicando *framework SCRUM*.

Cargo: *Team Leader*

Proyecto: *Corporate&InvestmentBanking* - Reforma Mercado de Valores

Funciones:

- Análisis, diseño y desarrollo de procesos *batch* encargados de generar reportes a entidades reguladoras como CNMV para la detección del abuso de mercado, BCE y SIF.
- Análisis, diseño y desarrollo de procesos encargados de realizar cuadres de las ejecuciones y titularidades existentes en los sistemas operacionales y los registrados

en el mercado (PTI).

- Participación en el proyecto de *Target2* (T2S) para la adaptación de interfaces de registro de titularidades (RT, TI), HTITUEA y HTITU03.

Enero 2010 – Agosto 2015. Tuyu Technology, Madrid (España)

Cargo: Analista

Proyecto: Subdirección de Declaraciones Fiscales e Informativas - Mantenimiento de Módulos de Renta

Funciones:

- Gestión de peticiones para los módulos de impuesto sobre la renta de las personas físicas, impuesto sobre la renta de no residentes, gestión de devoluciones e impuesto de sociedades.
- Responsable del análisis, diseño, implementación y pruebas de proyectos para los módulos de internet e intranet:
    - Sistema de gestión de declaraciones anuales y periódicas.
    - Sistema de gestión de solicitudes de borrador para IRPF.
    - Servicio de obtención online de borrador de declaración.
    - Sistema de devolución (adaptación mensajería SWIFT).

Mayo 2009 – Enero 2010. Hecaté Proyectos, Madrid (España)

Cargo: Analista Programador

Proyecto: Centro de Información -Mantenimiento de Módulos de Informes

Funciones:

- Optimización de procesos y mantenimiento de componentes batch para el Centro de Información.
- Construcción, pruebas e instalación de componentes para generación de informes periódicos para las áreas de finanzas y comercial.
- Seguimiento de evolución comercial y riesgo operacional.
- Evolución de depósitos confianza plus 4, 5 y 6; y del canal SUPERNET.
- Atención a usuarios y gestión de solicitudes.

Junio 2008 – Febrero 2009. Coritel, Barcelona (España)

Cargo: Programador Senior

Proyecto: Medios de Pago - Mantenimiento de Módulos Autorizador y Administrador de Tarjetas

Funciones:

- Implementación y pruebas de componentes para los módulos autorizador y administrador de tarjetas.
    - Administración de tarjetas para colectivos.
    - Gestión de incidencias de establecimientos a SERMEPA (chargebacks y representaciones).
    - Gestión de programa de puntos.
- Diseño de pantallas de terminal financiero para transacciones online.

Diciembre 2005 – Mayo 2008. **Stefanini IT Solutions, Lima (Perú)**

Cargo: Programador Senior

Proyecto: Factoría de Software

Funciones:

- Participación en la factoría de software de la Entidad Financiera y en el departamento de medios de pago.
- Codificación de componentes para proyectos del área de finanzas y contabilidad.
- Codificación de componentes para módulos de liquidación contabilidad, mantenimiento de cajeros y mantenimiento de tarjetas de crédito.

Formación Académica

2001 - 2006

Universidad Nacional Mayor de San Marcos, Lima (Perú)

Bachiller en Ingeniería de Sistemas e Informática.

Formación Académica Complementaria

ITIL Foundation V3 SunionGesfor Formación.
Project Management Professional Instituto de Formación Empresarial (IFE). Cámara Oficial de Comercio e Industria de Madrid.
J2EE - Desarrollo Web, JSF, SPRING SunionGesfor Formación.

Certificaciones

Professional Scrum Master I (02/2019) https://www.scrum.org/user/451455

CAPÍTULO II - CONTEXTO EN EL QUE SE DESARROLLÓ LA EXPERIENCIA

2.1. Empresa

Bluetab es una empresa tecnológicainternacional; con presencia en España, Perú, Colombia y México, dedicada a la asesoría empresarial y prestación de servicios basados en soluciones tecnológicas que emplean *Big Data*eInteligencia Artificial. Bluetab confía en los datos como herramienta de mejora de la rentabilidad, eficacia, progreso y crecimiento de las organizaciones que contratan sus servicios y, en su afán por mejorar su desempeño y contribuir al aumento de la satisfacción de todas sus partes interesadas, ha implantado un sistema de gestión de calidad conforme a la norma internacional ISO 9001:2015.

Soluciones que brinda la empresa:

- *Data Strategy*: Definir una estrategia de datos adecuada y seleccionar las tecnologías necesarias para su puesta en marcha requiere una visión independiente y experimentada.
    - *ArchitectureServices.*
    - *Data Governance& Data Quality.*
    - *Cloud Adoption& Security.*
- *Data Fabric:* Sabemos cómo poner en marcha el plan. Construir una plataforma de datos sobre entornos tecnológicos complejos requiere contar con profesionales experimentados
    - *Business Data Management.*
    - *BI & Modern BI.*
    - *Enterprise Reporting& Data VIZ.*
- *AumentedAnalytics:* La punta del *iceberg*. *Machine Learning* e Inteligencia Artifical con visión empresarial. Conocemos las técnicas y sabemos cómo utilizar la tecnología.
    - *Data Preparation.*
    - *Data Science.*
    - *AI & ML Applications*.

Los principales clientes de Bluetab España son:

Figura 1. Principales Clientes de Bluetab España (Fuente Intranet Bluetab)

2.2. Visión

Convertirse en socios de confianza de nuestros clientes, aportando soluciones con valor de negocio apalancadas en un uso de la tecnología de forma eficiente.

2.3. Misión.

Diseñar y desplegar soluciones de datos, desarrollando software para acelerar la implantación de nuestras soluciones y creando una cultura en la que se valora y se cuida al talento técnico.

2.4. Organización de la empresa

A continuación, se muestra el organigrama de la empresa Bluetab en el año 2021.

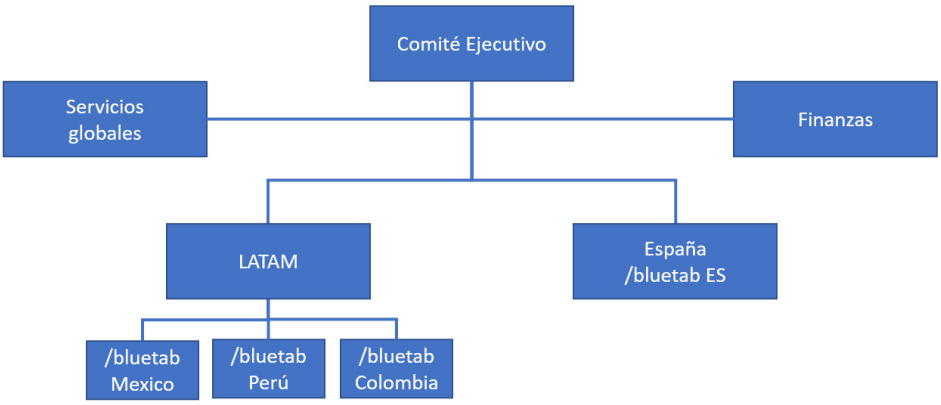


Figura 2. Organigrama Bluetab (Fuente Intranet Bluetab)

A continuación, se muestra el organigrama de Bluetab España en elaño 2021.

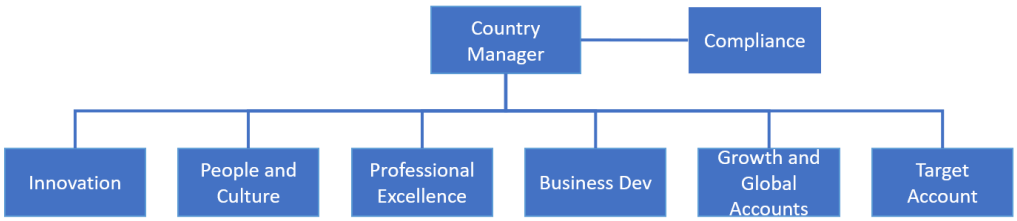


Figura 3. Organigrama Bluetab España (Fuente Intranet Bluetab)

2.5. Área, cargo y funciones desempeñadas

La autora de este Trabajo de Experiencia Profesional desempeña el cargo de *FunctionalTechnician* en la empresa Bluetab, desde abril del 2019 hasta la actualidad. Las funciones desempañadas son las siguientes:

a) Gestión de los usuarios manteniendo una correcta comunicación con el objetivo de comprender y cumplir los requisitos del cliente.
b) Asesoramiento funcional en el diseño de los modelos de datos y su posterior diseño físico en bases de datos Big Data.
c) Comprensión detallada de las fuentes orígenes para su explotación analítica.
d) *Feature&Engineering* o preparación de datos de entrada que tiene como objetivola explotación optimizada de los datos para crear un *dataset*que pueda ser consumido por el algoritmo.
e) Creación de soluciones *Machine Learning* con entornos de desarrollo *Python* y trabajando sobre grandes volúmenes de datos de diverso ámbito.
f) Analizar e interpretar los resultados del análisis y explicar las conclusiones al usuario.
g) Documentación de requerimientos y elaboración de documentos de análisis para su validación con el usuario.
h) Gestionar reuniones de trabajo y seguimiento con el usuario dentro del ámbito del proyecto.

2.6. Experiencia profesional realizada en la organización

Durante la experiencia profesional en Bluetab como *FunctionalTechnician*, la autora ha participado en la implementación del Modelo de Predicción de Contratación de Tarjetas de Crédito, cuya finalidad es incrementar la contratación de tarjetas de crédito.

CAPÍTULO III - ACTIVIDADES DESARROLLADAS

3.1. Situación problemática

3.1.1. Definición del problema

En el año 2019 la entidad financiera realizó un análisis de ventas del producto de tarjetas de crédito donde detectó una reducción en las ventas del producto. Tal como se observa en laFigura 4,la tendencia de efectividad en la venta de tarjetas de crédito es decreciente y en los últimos 4meses la efectividad de contratación había caído del 2,45% al 2,00%.La caída en las ventas supone una media mensual de 23.000 altas pérdidas cada mes, lo cual impactaría directamente en los objetivos anuales del 2019.

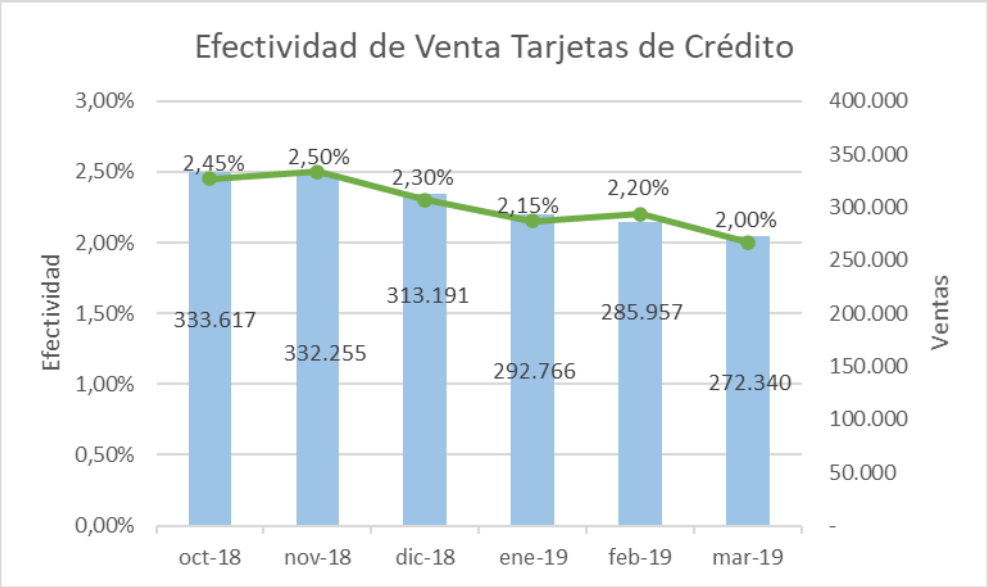


Figura 4. Efectividad de Campañas de Tarjetas de Crédito (Elaboración Propia)

La venta de tarjetas supone un 27% del total de ventas de productos de la entidad y tal como se ve en la Figura 5, es la tercera categoría con mayor cuota de mercado de la entidad.Es por ello que recuperar la venta de tarjetas de crédito constituye un objetivo estratégico para la entidad.

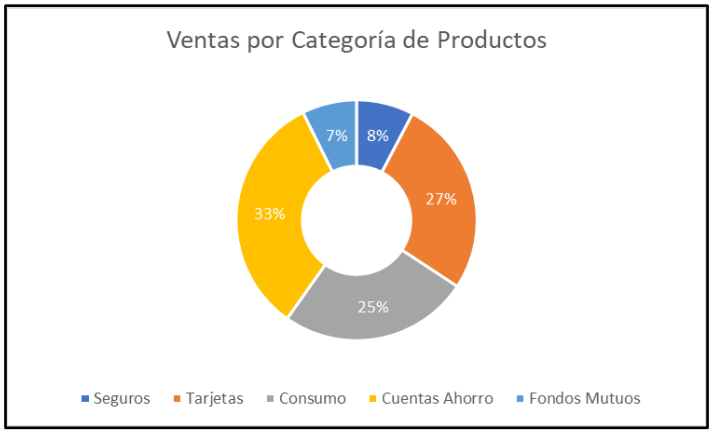


Figura 5. Ventas por Categoría de Productos 2019(Elaboración Propia)

Del análisis realizado se concluyó que se debe abordar el problema de decrecimiento de ventas para lograr los objetivos del año 2019 y recuperar el porcentaje de mercado que la entidad ha perdido.

3.2. Solución

Resulta primordial para la entidad recuperar la efectividad de las campañas de venta de tarjetas de crédito, en este sentido el área de *AdvanceAnalytics*recomendó implementar un modelo predictivo para determinar los clientes más propensos a la contratación de tarjetas de crédito.

El modelo predictivo realizará un análisis histórico de datos del cliente y con la selección del algoritmo adecuado permitirá conocer las relaciones entre las variables de entrada y salida, proporcionando a la entidad un mayor conocimiento del cliente y capacidad de predecir acciones futuras. Para la realización del proyecto será necesario que la entidad posea datos históricos de la contratación de productos de clientes, así como también información sobre características de los clientes (edad, género, ciudad, estado civil, etc.).

La aplicación del modelo entrenado sobre los datos permitirá identificar los clientes con mayor propensión a contratar tarjetas de crédito, este conocimiento servirá para direccionar estrategias comerciales a estos clientes, logrando incrementar las ventas e incrementando la satisfacción del cliente.

3.2.1. Objetivos

El principal objetivo de este proyecto es implementar un modelo predictivo para identificar los clientes potenciales para la contratación de tarjetas de crédito en los que debe enfocarse el área comercial para priorizar las estrategias de las campañas de marketing de la entidad.

El modelo predictivo será capaz de localizar patrones comunes de clientes propensos a la contratación de una tarjeta de crédito. La salida de este modelo será un *scoring* o probabilidad que permita ordenar a los clientes por su propensión a la contratación de una tarjeta de crédito.

3.2.1.1. Objetivos específicos

- Definir el diseño preliminar con el detalle a alto nivel de la solución a desarrollar.
- Determinar las variables clave que permitan identificar el perfil de los clientes más propensosa contrataruna tarjeta de crédito.
- Identificar las técnicas de los modelos de aprendizaje supervisados que más se adecuan para la solución.
- Validar el modelo en función a los resultados obtenidos de las técnicas de *Machine Learning*.
- Definir los criterios para la productivización de la solución en la entidad.

3.2.2. Enunciado del alcance del proyecto

El alcance del proyecto de implementación de un modelo de predicción de contratación de tarjetas de crédito se define con las actividades indicadas a continuación:

- Analizar las fuentes de datos disponibles para la implementación del modelo e identificar las variables que influyen de manera significativa en la probabilidad de adquisición de tarjetas de crédito.
- Desarrollar distintos modelos predictivos seleccionando los algoritmos que mejor se ajusten a la solución.
- Identificar, a partir de distintas métricas de comparación, el modelo con el mejor desempeño.

3.2.3. Etapas y metodología

La metodología utilizada en la implementación del modelo predictivo es la CRISP-DM (*Cross Industry Standard Processfor Data Mining*). La metodología CRISP-DM está compuesta por seis fases, las cuales dependen entre sí tanto en forma secuencial como cíclica pudiendo regresar a alguna de las fases anteriores para mejorar la aproximación obtenida, tal como se muestra en la siguiente figura.

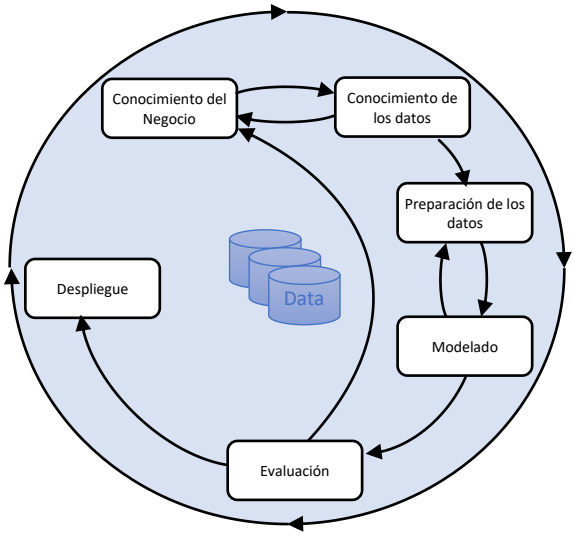


Figura 6. Fases de la Metodología CRISP-DM (Wirth & Hipp, 2000)

A continuación, se describe cada fase:

- **Fase 1 Comprensión del negocio**. Esta primera fase se centra en comprender los objetivos y requisitos del proyecto desde una perspectiva de los requerimientos funcionales para luego traducir ese conocimiento en la definición del problema de *Machine Learning* y trazar un plan para lograr los objetivos.

- **Fase 2 Comprensión de datos**. La fase de comprensión de datos comienza con la recopilación de las fuentes de datos origen y continúa con las actividades para comprender los datos origen, identificar problemas de calidad y revelar lascaracterísticas generales sobre los datos.

- **Fase 3 Preparación de datos.** La fase de preparación de datos incluye todas las actividades parala construcción del conjunto de datos de entrada para el modelo,a partir de los datos originales sin procesar.

- **Fase 4 Modelado.**En esta etapa, se seleccionan y aplican las técnicas de modelado que mejor se adapten al problema, ajustando los parámetros a los valores más óptimos. Existen diversas técnicas para el mismo tipo de problema de *Machine Learning*, algunas de las cuales tienen requisitos específicos de formato de datos. Como resultado, casi todos los proyectos vuelven a la etapa de preparación de datos.

- **Fase 5 Evaluación.**Al llegar a esta fase del proyecto ya se han construido uno o más modelos de alta calidad para la resolución del problema. Antes de continuar con la implementación final del modelo, es importante evaluar más a fondo el modelo y revisar los pasos de construcción del modelo para asegurarse de que los objetivos de negocio se cumplen correctamente. Al final de esta fase se debe decidir cómo utilizar los resultados del análisis de datos.

- **Fase 6 Despliegue.**Generalmente el proyecto no finaliza con la creación del modelo. Si bien el propósito del modelo es aumentar el conocimiento de los datos, el conocimiento resultante debe estructurarse y presentarse para su uso por parte del negocio. Dependiendo de sus necesidades, las etapas de desarrollo pueden ser tan simples como generar un informe o tan complejas como automatizar el análisis de datos dentro de la organización. En la mayoría de los casos, es el usuario de negocio es el que realiza estos pasos. De cualquier manera, es importante saber de antemano qué pasos debe seguir para beneficiarse del modelo creado.

Se estableció un periodo de 3 meses para la elaboración del modelo, distribuido en seis fases aplicando la metodología de CRISP-DM, tal como se muestra en la Figura 7. Las primeras dos fases tuvieron una duración de 3 semanas y consistieronen reuniones periódicas con el área de negocio para analizar las variables disponibles y determinar el histórico que se dispondría para la construcción del modelo.Además de comprender los resultados que esperaba el negocio. Las siguientes dos fasestuvieron una duración de 7 semanas y consistieron en preparar los datos, implementar y evaluar la solución. La última fase consistió en la implantación del proyecto y tuvo una duración de 2 semanas.

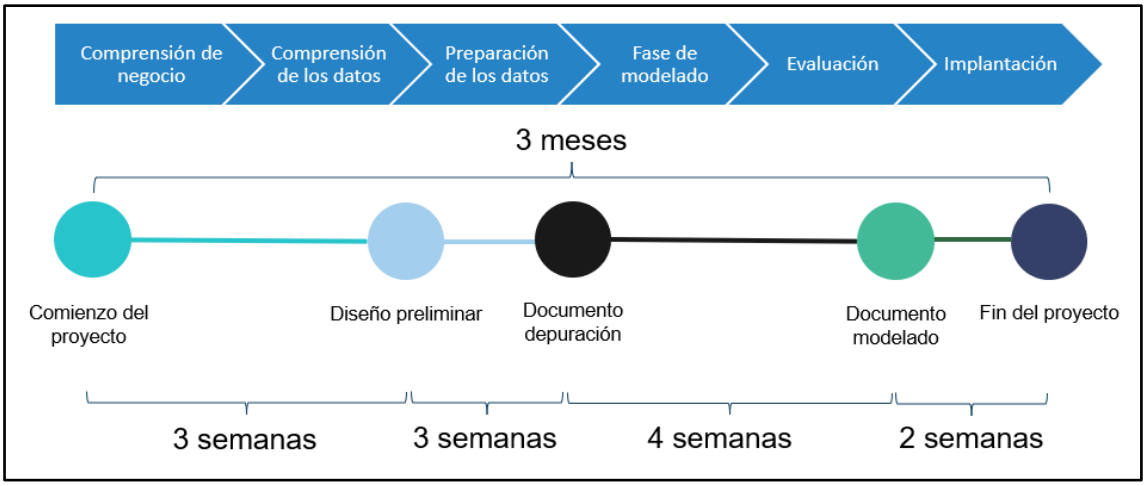


Figura 7. Planificación del Proyecto (Elaboración propia)

3.2.4. Fundamentos utilizados

3.2.4.1. Técnicas de Machine Learning

Machine Learning es una rama de la inteligencia artificial, se basa en la premisa de que programas computacionales puedan aprender de los datos sin intervención humana. El aprendizaje lo consiguen identificando patrones dentro de un gran número de datos. Este método de análisis de datos se basa en que un algoritmo sea capaz de predecir comportamientos futuros (Trujillo Fernandez, 2017).

Una de las definiciones más citadas que conceptualiza el término de Machine Learning es la aportada por Tom M. Mitchell.

> "Se dice que un programa informático aprende de una experiencia E con respecto a alguna clase de tarea T y una medición del rendimiento P, si su rendimiento en las tareas T, medido como P, mejora con la experiencia E" (Mitchell, 1990)

Los métodos de aprendizaje supervisado y el aprendizaje no supervisado, son los más usados dentro de los métodos de *Machine Learning. T*ambién existen otros métodos de *Machine learning* como el aprendizaje semi-supervisado y el aprendizaje con refuerzo. Nos centraremos en la definición de la técnica de aprendizaje supervisado, que es el método aplicado para la implementación de la solución descrita en estetrabajo.

Aprendizaje supervisado

El aprendizaje automático supervisado es la búsqueda de algoritmos que, a partir de instancias proporcionadas externamente, produzcan hipótesis de predicciones sobre instancias futuras. Los algoritmos aprenden a partir de datos etiquetados donde se conoce el resultado esperado. A partir de estos datos se compararán los resultados del algoritmo con los resultados esperados para encontrar errores y corregir el modelo. Por ejemplo, un conjunto de datos de entrada; cuyas observaciones están etiquetadas como "A" (aciertos) o "F" (fallos),el algoritmo será capaz de identificar qué observaciones del conjunto de entrada pertenecen a cada clase.

El aprendizaje supervisado es aplicado comúnmente en problemas donde los datos históricos son capaces de predecir eventos futuros. Por ejemplo, una aplicación es capazde anticiparse y puede predecirla probabilidad de que una transacción con tarjeta de crédito sea fraudulenta.

Existen fundamentalmente dos tipos de aprendizaje supervisado (Pallarés, 2019):

1. **Regresión.**En este caso la variable dependiente es continua, tiene como objetivo predecir valores continuos. Este tipo de aprendizaje es aplicable a distintos problemas para determinar un valor. Por ejemplo, cómo predecir el valor de un producto o determinar la cantidad de uso de un servicio por parte del cliente, etc.

2. **Clasificación.**En este caso la variable dependiente es discreta y categórica. A partir de un conjunto de datos de entrada, el algoritmo será capaz de predecir a qué clase pertenece cada una de las observaciones del conjunto. La clasificación puede ser binaria o multiclase, se denominará binaria cuando solo existan dos clases y multiclase cuando existan más de dos clases.

3.2.4.2. Algoritmos de Clasificación

Árboles de Decisión

Los árboles de decisión se ubican dentro de una rama del aprendizaje automático supervisado. Es una estructura de datos jerárquica que usa la estrategia de divide y vencerás. Crear reglas de clasificación en forma de árboles de decisión a partir de un conjunto de ejemplos dados. El árbol de decisión se construye de arriba hacia abajo utilizando la ganancia de información normalizada que resulta de elegir un atributo para dividir los datos. El atributo con la mayor

ganancia de información normalizada es el que se utiliza para tomar la decisión (Fernández, López, Galar, José del Jesus, & Herrera, 2013). Su estructura se convierte fácilmente en un conjunto de reglas simples.

Dentro de un árbol de decisión se pueden identificar los siguientes tipos de nodos:

- Primer nodo o nodo raíz: es en ese nodo donde se produce la primera división en función de la variable más importante (mayor ganancia).
- Nodos internos o intermedios: estos nodos se encuentran tras la primera división y vuelven a dividirse en función de las variables.
- Nodos terminales u hojas: indican la clasificación definitiva y se ubican en la parte inferior del árbol.

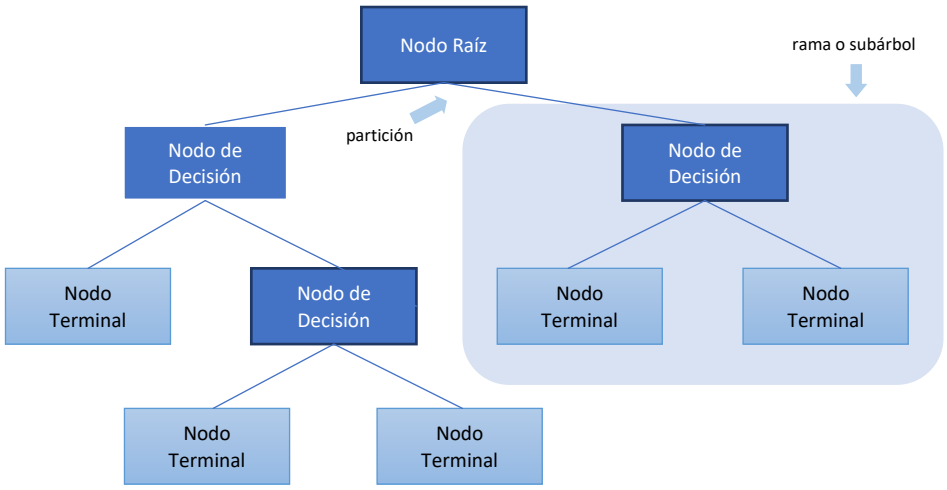


Figura 8. Estructura de un Árbol de Decisión (Elaboración propia)

Para la construcción de un árbol de decisión se aplica el algoritmo de Hunt, que se basa en la división óptima de un conjunto de datos en subconjuntos. La construcción inicia a partir de un conjunto de datos en un nodo, si todas las observaciones pertenecen a una misma clase se tratará de un nodo terminal, en caso contrario se dividirán los datos en función de una variable para formar subconjuntos más pequeños. La variable seleccionada para el particionamiento será la que mejor divida el conjunto de datos con respecto al criterio de partición. Este proceso se aplicará recursivamente a cada subconjunto hasta una condición de parada.

La medida de la impureza de un nodo es el criterio aplicado para ladivisión o partición del conjunto de datos. Esta impureza se calcula aplicando el Índice Gini o Entropía.

El índice de Gini mide el grado de pureza de un nodo, esto quiere decir que mide la probabilidad de que en un nodo no existan dos observaciones de la misma clase.Aplicando este criterio se seleccionará la variable con menor Gini ponderado, porque mientras mayor sea el índice de Gini menor pureza. Frecuentemente divide en un nodo una clase mayoritaria y el resto de clases los clasifica en otros nodos, generándose divisiones desbalanceadas.

El índice de Gini se define como:

Dado un conjunto de casos T, con varias clases $c_i, i = 1,2,\dots,n$.

$$G(T) = 1 - \sum_{i=1}^{n} (P_i)^2$$

Donde P_i es la probabilidad de que una observación sea de la clase i.

La entropía mide el grado de incertidumbre de una muestra. La entropía será cero si el nodo es puro, esto quiere decir que las observaciones del nodo son de la misma clase. En caso contrario alcanzará el valor máximo de 1. La entropía suele construir nodos balanceados en el número de observaciones.

La entropía se define como:

Dado un conjunto de casos T, con múltiples clases c_i, $i = 1,2,\dots,n$.

$$E(T) = -\sum_{i=1}^{n} P_i * \log_2 P_i$$

Donde P_i es la proporción de casos en T que pertenecen a la clase i.

Relacionado con la entropía se define la Ganancia de Información que busca la división con menor incertidumbre, esto quiere decir con menor entropía ponderada de la variable.

El número de información necesaria para clasificar un caso del conjunto de casos T se define en la función $E(T)$. Dado la variable A que divide el conjunto T en los subconjuntos $E_j, j = 1,2,\dots,m$, la entropía total del sistema de subconjuntos se calculará como:

$$E(T, A) = \sum_{j=1}^{m} P(T_j) E(T_j)$$

donde $P(Tj)$ es la proporción de casos en T que pertenecen a Tj y $E(Tj)$ es la entropía del subconjunto Tj.

Y la ganancia de información que aporta dividir el conjunto Trespecto a la variable A, se define como:

$$Ganancia(T, A) = E(T) - E(T, A)$$

donde $E(T)$ es el valor de la entropía antes de realizar la división del conjunto de datosy$E(T, A)$ es el valor de la entropía del sistema de subconjuntos generados por la subdivisión a partir de la variable A.

Las principales ventajas algoritmo de Árbol de Decisión son:

- Los árboles son fáciles de interpretar y visualizar porque tiene representación gráficamente.
- Requieren menor limpieza y preprocesamiento de los datos a diferencia de otros algoritmos.
- Puede obtener una predicción utilizando las observaciones que pertenecen al último nodo alcanzado, cuando el valor de la variable de predicción no esté disponible para una observación. La precisión de la predicción se reduce, pero aun así se puede lograr.
- Permite identificar de manera rápida y eficiente las variables con el mayor poder predictivo.
- Son aplicables a problemas de regresión y clasificación.

Por el contrario, sus principales desventajas son las siguientes:

- El poder de predicción de los modelos construidos con un único árbol es bastante inferior a la conseguida con otros modelos que combinan varios árboles. Debido a su tendencia al *overfitting* y alta varianza.
- Alta sensibilidad a datos de entrenamiento desbalanceados.
- Una pequeña variación en los datos puede dar lugar a un árbol de decisión diferente.
- Al categorizar predictores continuos, pierden parte de su información en la división de los nodos.

Random Forest

El algoritmo *Random Forest*es una técnica de aprendizaje supervisado basada en el método de *ensemble* o ensamblado, que consiste en el entrenamiento de varios modelos para resolver un mismo problema. Dentro del método *ensemble*, el *algoritmo Random Forest* aplica la técnica de*Bagging*, en la que el entrenamiento de los clasificadores es en paralelo (Breiman, 2001).

El algoritmo generará múltiples árboles de decisión a partir de un conjunto de datos de entrenamiento y posteriormente combina los resultados. La técnica utilizada se denomina *BoostrapAggregating*, que consiste en generar una serie de muestras de variables o predictores para generar un conjunto de modelos. Las predicciones de los modelos se combinan aplicando una media aritmética, tal como se muestra en la siguiente figura.

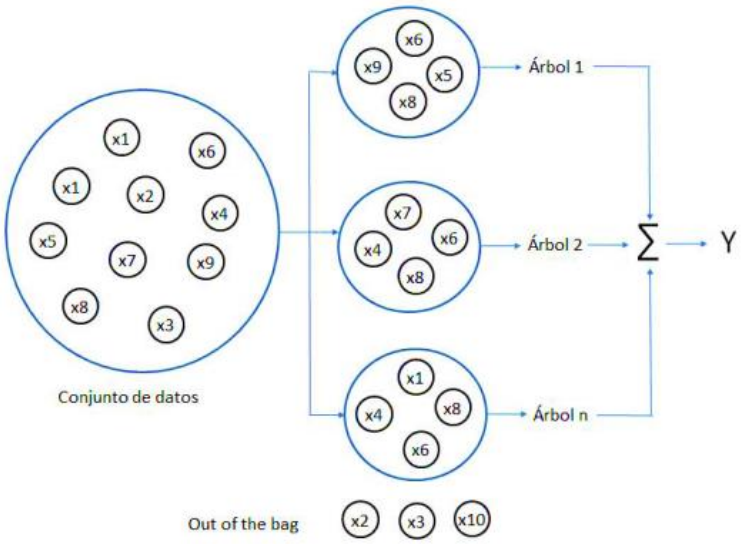


Figura 9. Algoritmo Random Forest (Espinosa-Zúñiga, 2020)

El algoritmo de *RandomForest* consta de los siguientes pasos (Trujillo Fernandez, 2017):

1. Se crean aleatoriamente n subconjunto a partir del conjunto de datos de entrenamiento. Cada árbol utilizará estos subconjuntos como conjuntos de entrenamiento. Debido a que la selección de los datos es aleatoria, no todos los datos de la muestra original estarán en los subconjuntos generados. Los datos no incluidos se denominan "*outof bag*".
2. Cada árbol generado contiene una muestra aleatoria de predictores$n$,donde$n < N$ (N = total de predictores).
3. Se crea cada árbol con la máxima profundidad posible.
4. Para los problemas de clasificación, el algoritmodará como resultado la votación mayoritaria de los árboles.

Como se mencionó anteriormente, el método de *Random Forest* se basa en la construcción de una serie de árboles de decisión. Para este método una muestra del conjunto de datos entra en un árbol y se realizan una serie de pruebas binarias en cada nodo (llamadas "*split*") hasta llegar a su hoja donde se encuentra la respuesta. Esta técnica se utiliza para dividir un problema complejo en una serie de problemas sencillos.

En esta fase de entrenamiento, el algoritmo tiene como objetivo optimizar los parámetros de las funciones de prueba binarias (llamadas *"split")* a partir del conjunto datos de entrenamiento.

$$\theta_k^* = argmax_{j \in_j} I_j$$

La función de ganancia de información es la utilizada para este proceso:

$$I_j = H(j) - \sum_{i \in 1,2} \frac{|S_j^i|}{|S_j|} H(S_j^i)$$

Donde S representa la muestra que hay en el nodo por dividir y S^i son los dos conjuntos que se crean producto de la división. La función mide la entropía del conjunto, y esta depende del tipo de problema a resolver (Breiman, 2001).

El algoritmo de *Random Forest* se define como un clasificador que está formado por un conjunto de clasificadores con estructura de árbol $\{h(x, \Theta_k), k = 1, \dots\}$, donde el $\{\Theta_k\}$ son los vectores aleatorios distribuidos idénticamente y cada árbol participa en la selección de la clase más popular(Medina Merino & Ñique Chacón, 2017).

La ganancia de información representada en la función $I(.)$, dado un conjunto de clasificadores $h_1(x), h_2(x), \dots, h_k(x)$que se seleccionarán al azar con la distribución del vector aleatorio Y, X. La clasificación será más confiable, mientras mayor sea el margen. La función de margen se define como:

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

La generalización de error de clasificación se define como:

$$PE^* = P_{X,Y}(mg(X, Y) < 0)$$

Donde $P_{X,Y}$ indica la probabilidad está sobre el espacio X, Y. Dado un gran número de árboles se muestra una convergencia para el errory la misma estructura del árbol. Tal como se indica en el siguiente teorema:

> A medida que aumenta el número de árboles, es muy probable que todas las secuencias $\Theta 1, \Theta 2, \ldots ; PE *$convergen a:
>
> $$P_{X,Y}(P_\theta(H(X,\theta) = Y) - max_{j \neq Y} P_\theta(h(X,\theta) = j) < 0)$$

Esto demuestra porqueel algoritmo*Random Forest* no sobreajusta aunque se aumente el número de árboles, pero generar un valor límite de la generalización de error.

Las principales ventajas del algoritmo de *Random Forest* son:

- Son aplicables a problemas de regresión y clasificación.
- El algoritmo admite variables discretas y continuas.
- Estima la importancia de cada variable en la clasificación y selecciona las variables más importantes.
- Puede ser utilizado sobre una gran cantidad de datos manteniendo un buen desempeño. Genera una clasificación muy certera en caso de tener una muestra suficientemente grande.

Por el contrario, sus principales desventajas son las siguientes:

- Al contrario que los árboles de decisión, el *Random Forest* es un modelo difícil de interpretar gráficamente.
- Para ajustar el modelo, puede ser necesario un tratamiento de los datos de entrada.
- En el caso de que exista un elevado ruido en los datos, puede tender a sobreajustarpara algunos problemas de clasificación.
- Se tiene poco control sobre lo que hace el modelo.

GradientBoosting

El algoritmo *GradientBoosting* es una técnica de aprendizaje supervisado basada en el método de *ensemble* o ensamblado, que consiste en el entrenamiento de varios modelos para resolver un mismo problema al igual que el algoritmo de *Random Forest*. Aunque, difiere en que los clasificadores se entrenan de forma secuencial y cada clasificador sucesivo se construye teniendo en cuenta los errores de la predicción del clasificador anterior. De esta manera, cuantos más árboles se construyan, el residuo es cada vez más pequeño, aunque no se debe

abusar del número de árboles, porque puede dar lugar a sobreajuste. En la Figura 10. se representa gráficamente la combinación de árboles de decisión individuales.

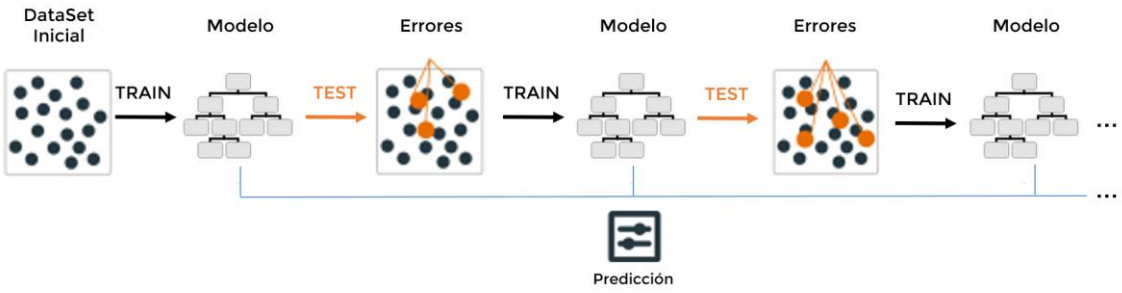


Figura 10. Algoritmo GradientBoosting(Boehmke & Greenwell, 2020)

Para construir el modelo final, el algoritmo sigue los siguientes pasos(Trujillo Fernandez, 2017):

1. La construcción del algoritmo inicia con asignarle el mismo peso a las observaciones de la muestra del conjunto de datos de entrenamiento con los que se va a crear el primer clasificador. El peso está definido como $w_i = \frac{1}{n}$, con $i = 1,2 \dots n$, donde n es el tamaño de la muestra.
2. El primer clasificador se entrena utilizando la muestra de entrenamiento y se mide el error cometido por el modelo y la ponderación α_1 asociada al modelo $g_1(x)$. Se incrementan los pesos en los casos de entrenamiento en los que el modelo calcula erróneamente.
3. Estafase se repite hasta que la función del error disminuye (criterio de convergencia) o hasta que se llega al número de iteraciones marcadasoriginalmente.
4. Elresultado del modelo será un promedio ponderado de los resultados obtenidos.

Para definir matemáticamente el algoritmo de *GradientBoosting*,se parte de un modelo de predicción imperfecto f_m. Para mejorarlo se añade un estimador h de manera que $f_{m+1} = f_m + p$. Para calcular $p$se parte de la premisa de que $p$no tiene errores, cumpliendo $f_{m+1} = f_m + p = y$. En consiguiente$p = y - f_m$(Cimarra Muñoz, 2018).

Donde $y - f_m$ es la función residual.El objetivo es minimizar la función residual o reducir una función de pérdida $p(y, f) \in \mathbb{R}$, donde f es la función de predicción. Se busca la evaluación de una función de predicción óptima f^*:

$$f^* \coloneqq argmin_f \mathbb{E}_{Y,X}[p(y, f(x^T))]$$

La expectativa $\mathbb{E}_{Y,X}$comúnmente no es conocida porque el algoritmo no trata de reducir esta expectativa, sino que trata de reducir el "riesgo empírico" o valor medio observado $R \coloneqq \sum_{i=1}^{n}(y, f(x_i^T)$.

Las principales ventajas del algoritmo de *GradientBoosting*son:

- Es uno de los algoritmos de clasificación más precisos.
- Aunque la cantidad de observaciones sea pequeña, soportauna gran cantidad de variables.
- Es capaz de estimar datos incompletos.
- No es sensible frente a valores atípicos.
- Al utilizar múltiples árboles se reduce considerablemente el riesgo de *overfiting*.
- Aplicable a problemas de clasificación y también de regresión.

Por el contrario, sus principales desventajas son las siguientes:

- Es difícil de interpretar debido a que combina múltiples árboles a diferencia de los modelos construidos a partir de un árbol único.
- El algoritmo pierde parte dela información para predictores continuos porquelas categoriza al dividir los nodos.
- No puede extrapolar más allá del rango de los predictores observados en el conjunto de datos de entrenamiento.
- Es muy sensible a la presencia de ruido y valores atípicos en la muestra de entrenamiento.

3.2.4.3. Métricas de evaluación de modelos

Accuracy

La métrica *Accuracy*(exactitud) es la métrica más básica para evaluar un modelo de Machine Learning. Mide el porcentaje de observaciones en la que el modelo haacertado y se calcula a partir de la matriz de confusión:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Donde:

TP= *True Positives* (Verdaderos positivos)

TN= *True Negatives*(Verdaderos Negativos)

FP= *False Positives*(Falsos Positivos)

FN= *False Negatives*(Falsos Negativos)

Para un conjunto de datos desbalanceados, si el modelo predice siempre la clase mayoritaria, el su nivel de Accuracy sería muy alto. Es por ello que se sugiere aplicar esta métrica de validación junto con otras, como la que se menciona a continuación.

Curva ROC – AUC

El área bajo la curva ROC o AUC es una de las métricas más eficaces para la validación del poder predictivo de un modelo de clasificación binaria. La curva ROC se representa gráficamente trazando la tasa de sensibilidad (verdaderos positivos) y la tasa de 1-especificidad (falsos positivos). El AUC oÁrea Bajo la Curva ROC(Bradley, 1997)es la medida más eficaz para la validación de clasificadores. Esta métrica es la más recomendada para la selección de un modelo teniendo en cuenta su interpretación, tal como se muestra a continuación:

Por ejemplo, en la Figura 11, la gráfica representa un modelo con valor diagnóstico perfecto, el modelo es capaz de distinguir el 100 % entre clase positiva y la clase negativa.

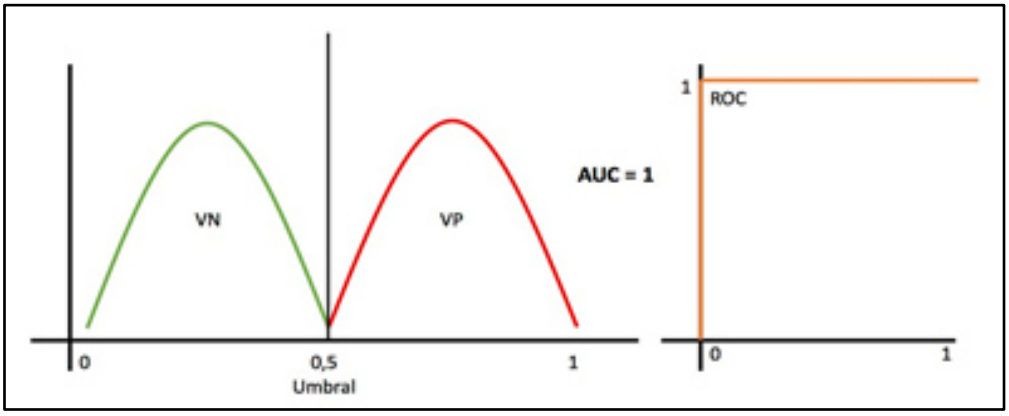


Figura 11. Curva ROC-Valor Diagnóstico Perfecto(aprendeIA, 2021)

La Figura 12 muestra un modelo con valor diagnóstico medio, muestra una situación común donde las distribuciones se superponen e introducen errores de FN (Falsos Negativos) y FP

(Falsos Positivos), en este caso tenemos un AUC 0,7 que quiere decir que hay un 70% de probabilidad de que el modelo pueda distinguir entre una clase positiva y negativa.

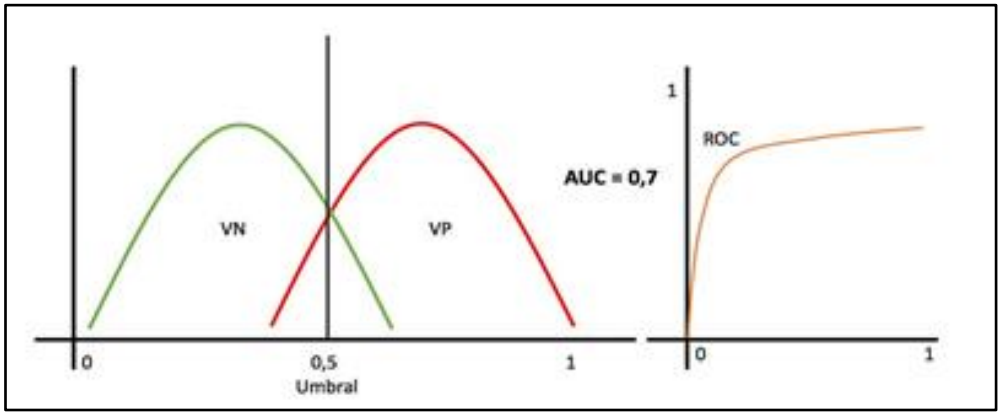


Figura 12. Curva ROC-Valor Diagnóstico Medio(aprendeIA, 2021)

Y en la Figura 13, donde el valor AUC es 0,5, es la peor situación porque el modelo no tiene poder de discriminación.

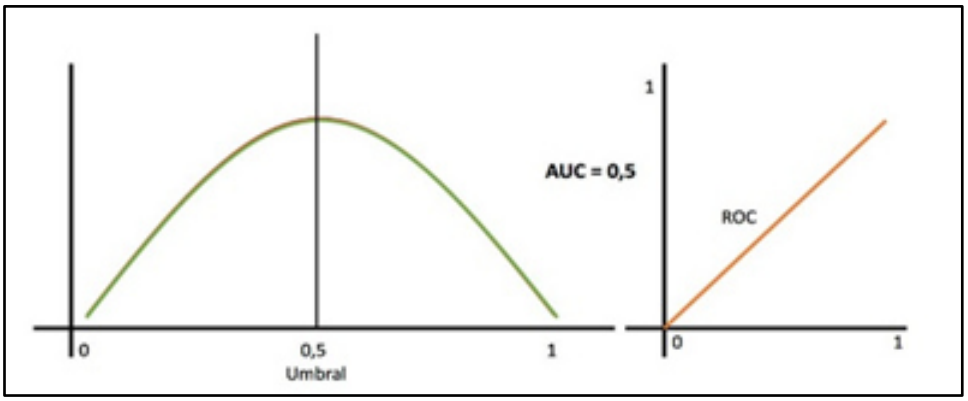


Figura 13. Curva ROC-Sin Valor Diagnóstico (aprendeIA, 2021)

De acuerdo al poder predictivo de un modelo, los valores del AUC se pueden encontrar entre los siguientes intervalos:

- [0,5]: Sin capacidad de predicción.
- [0.5, 0.6): Modelo malo.
- [0.6, 0.75): Modelo regular.
- [0.75, 0.9): Modelo bueno.
- [0.9, 0.97): Modelo muy bueno.
- [0.97, 1): Modelo excelente.

3.2.5. Implementación de las áreas de procesos y sus buenas prácticas

En esta sección se detallará la implementación del modelo de aprendizaje supervisado en el entorno de la entidad bancaria y aplicando las fases de la metodología CRISP-DM

3.2.5.1. Fase 1. Comprensión del negocio

- ***Objetivos del negocio***

  Implementar una estrategia comercial basada en una solución *Machine Learning* con visión del cliente, mejorará la tasa de éxito de las campañas comerciales de tarjetas de crédito y aumentará tanto los ingresos globales como el compromiso/satisfacción del cliente.

- ***Objetivos del proyecto***

  Implementar un modelo predictivo que permita determinar las principales variables que identifiquen a un cliente de la entidad con mayor propensión a contratar una tarjeta de crédito.

- ***Resultados esperados.***

  El modelo predictivo identificará patrones comunes de clientes propensos a la contratación de una tarjeta de crédito dos meses después a la ejecución del algoritmo, teniendo en cuenta que la ejecución del algoritmo se realizará los últimos días del mes. La salida de este algoritmo será un *scoring* que permita ordenar a los clientes por su propensión a la contratación de una tarjeta de crédito. Este resultado permitirá que el área comercial enfoque las campañas comerciales a estos clientes y así reducir costes y aumentar su efectividad.

3.2.5.2. Fase 2. Comprensión de los datos

- ***Comprensión a alto nivel de los datos.*** Elárea de negocio selecciona la fuente de datos para la implementación del modelo:
  - **Fuente de datos origen:** Vista de clientes-productos mensual (datos del cliente, cartera de productos contratados).
  - **Histórico:** La entidad pone a disposición del proyecto 17 fotos de la vista de clientes-productos para construir del modelo.

- ***Comprensión detallada de los datos.***

  La vista de clientes-productosdisponible para la construcción del modelo está formada por 48 variables y la volumetría es de 13.619.575 registros. Cada registro contendrá la foto de las características del cliente y la cartera de productos de ese cliente en la fecha de extracción de los datos.

	fec_dato	cod_cliente	xti_empleado	nom_pais_resid	xti_género	num_edad	fec_alta_cli	xti_nuevo	num_antiguedad	xti_estado
0	2018-01-27	1375586	N	ES	H	35.0	2018-01-11	0.0	6	1.0
1	2018-01-27	1050611	N	ES	V	23.0	2015-08-10	0.0	35	1.0
2	2018-01-27	1050612	N	ES	V	23.0	2015-08-10	0.0	35	1.0
3	2018-01-27	1050613	N	ES	H	22.0	2015-08-10	0.0	35	1.0
4	2018-01-27	1050614	N	ES	V	23.0	2015-08-10	0.0	35	1.0

Figura 14. Muestra de Datos Entrada (Elaboración Propia)

En la siguiente tabla se muestra el listado de variables en los datos de entrada y la tipología de cada variable.

Tabla 1. Listado de Variables en Datos de Entrada (Elaboración Propia)

Nombre de Variable	Descripción Variable	Tipo de Dato	Tipo de Variable
fec_dato	Fecha de dato	date	Numérica
cod_cliente	Código de cliente	long	Categórica
xti_empleado	Indicador de empleado	string	Categórica
	Nombre de país de residencia del cliente	string	Categórica
xti_genero	Indicador de género	string	Categórica
num_edad	Edad del cliente	int	Numérica
fec_alta_cli	Fecha de alta del cliente	date	Numérica
xti_nuevo	Indicador de nuevo del cliente	int	Categórica
num_antiguedad	Antigüedad en meses del cliente	int	Numérica
xti_estado	Indicador de estado de cliente titular	int	Categórica
fec_baja_cli	Fecha de baja del cliente	date	Numérica
xti_tipo_cliente	Tipo de cliente	int	Categórica
xti_relacion	Relación del cliente con la entidad	string	Categórica
xti_residencia	Indicador de residencia en el país de la entidad.	string	Categórica
xti_extranjero	Indicador de residencia en el extranjero	string	Categórica
xti_conyugue	Indicador de conyugue empleado	int	Categórica
nom_canal	Nombre del canal asociado al cliente	string	Categórica
xti_fallecimiento	Indicador de fallecimiento del cliente	string	Categórica
xti_domicilio	Indicador de domicilio informado	int	Categórica
cod_provincia	Código de provincia del domicilio	int	Categórica
nom_provincia	Nombre de provincia del domicilio	string	Categórica
xti_activo	Indicador cliente activo	int	Categórica
imp_renta	Importe de renta anual del cliente	float	Numérica

cod_segmento	Código de segmento del cliente	string	Categórica
xti_ahor_fin_ult1	Indicador de cuenta de ahorro	int	Numérica
xti_aval_fin_ult1	Indicador de cuenta de garantías	int	Numérica
xti_cco_fin_ult1	Indicador de cuenta corriente	int	Numérica
xti_cder_fin_ult1	Indicador de cuenta derivada	int	Numérica
xti_cno_fin_ult1	Indicador de cuenta nómina	int	Numérica
xti_ctju_fin_ult1	Indicador de cuenta junior	int	Numérica
xti_ctma_fin_ult1	Indicador de cuenta más particular	int	Numérica
xti_ctop_fin_ult1	Indicador de cuenta particular	int	Numérica
xti_ctpp_fin_ult1	Indicador de cuenta de particular plus	int	Numérica
xti_deco_fin_ult1	Indicador de depósitos a corto plazo	int	Numérica
xti_deme_fin_ult1	Indicador de depósitos a medio plazo	int	Numérica
xti_dela_fin_ult1	Indicador de depósitos a largo plazo	int	Numérica
xti_ecue_fin_ult1	Indicador de cuenta online	int	Numérica
xti_fond_fin_ult1	Indicador de fondos	int	Numérica
xti_hip_fin_ult1	Indicador de hipoteca	int	Numérica
xti_plan_fin_ult1	Indicador de pensiones	int	Numérica
xti_pres_fin_ult1	Indicador de préstamos	int	Numérica
xti_reca_fin_ult1	Indicador de impuestos	int	Numérica
xti_tjcr_fin_ult1	Indicador de tarjeta de crédito	int	Numérica
xti_valo_fin_ult1	Indicador de valores	int	Numérica
xti_viv_fin_ult1	Indicador de cuenta de vivienda	int	Numérica
xti_nomina_ult1	Indicador de nómina	int	Numérica
xti_nom_pens_ult1	Indicador de pensiones	int	Numérica
xti_recibo_ult1	Indicador de débito directo	int	Numérica

Como primera fase para la comprensión de las variables que forman parte del *dataset*de entrada, se realiza un análisis univariante en el que se incluye el comportamiento de cada variable independiente.Con este análisis podemos identificar la distribución de las variables y la existencia de valores nulos u *outliers* (*Anexo 1: Análisis Univariante)*.

- ***Público objetivo.***

  Se entiende por público objetivo la población sobre la que se va a ejecutar este modelo. Es importante que esta población esté definida previamente porque es la misma con la que se debe entrenar el algoritmo.

  En el caso del trabajo descrito, el público objetivo lo forman los clientes activos que no tengan contratada una tarjeta de crédito y que tengan una antigüedad en la entidad mayor o igual a 6 meses.

- ***Definición de target.***

  El target o variable objetivo en los modelos de este tipo se construye como un indicador binario en el que se marca como 1 o positivo a aquellos clientes que cumplen el evento a predecir y como 0 o negativo los que no lo cumplen.

  - Se consideran **casos positivos** en la fecha de referencia, aquellos clientes que, sin haber contratado una tarjeta de crédito en el mes de ejecución, son propensos a contratar una tarjeta dos meses después de la ejecución, sin haberla contratado en el mes anterior.

  - Se consideran **casos negativos** en la fecha de referencia, aquellos clientes que no son propensos a contratar una tarjeta en cualquiera de los meses del periodo de entrenamiento.

  Aunque el objetivo del modelo es predecir qué clientes contratarán una tarjeta de crédito dos meses después a la ejecución del modelo, desde un punto de vista de entrenamiento del modelo es necesario incluir un mes intermedio adicional ya que los datos de cierre de mes no se obtienen en los últimos días del mes.

- ***Horizonte de predicción.***

  Se entiende por horizonte de predicción o meses ciegos a la diferencia en tiempo entre la última actualización de datos y la fecha para la que el modelo está entrenado para predecir. Es muy importante ser cuidadoso en la definición de este horizonte porque es el que nos permite no llegar tarde con las acciones que se pueda derivar de los resultados de este análisis.

  Se plantea utilizar como horizonte de predicción 1 mes. Esto significa que con los datos de cierre Julio 2019 se predecirán los clientes que realizar contrataciones de tarjetas de crédito en el mes de Septiembre.

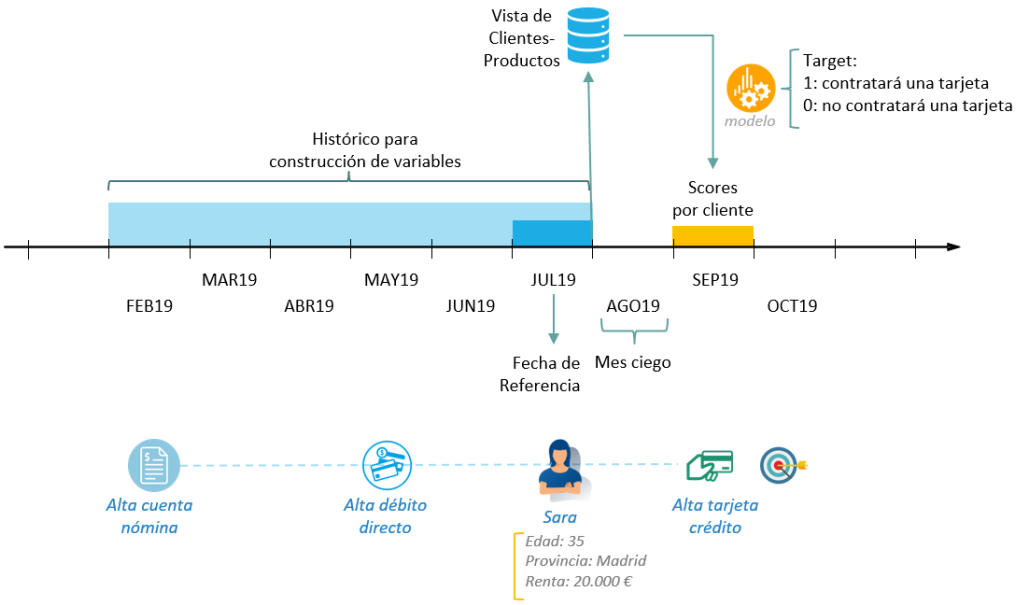


Figura 15. Definición de Horizonte de Predicción (Elaboración Propia)

Producto de las dos primeras fases se genera el primer entregable que es el Diseño Preliminar (*Anexo 2. Diseño Preliminar*), este documento es primordial para una correcta ejecución del proyecto y sirve para garantizar que los requerimientos del negocio se sido entendidos claramente y que los resultados son los esperados.

3.2.5.3. Fase 3. Preparación de los datos

- ***Construcción de Predictores Inteligentes***

  Una variable predictora o sintética es una variable que se construye a partir de variables originales, de esta forma se resume la información original. Las variables predictoras suelen ser habitualmente promedios, conteos, sumatorios, ratios o incrementos de varios meses con el objetivo de medir las evoluciones de métricas como saldos o número de movimientos.

  Las variables predictoras definidas son a partir de los datos originales son:

- num_prods_total: Contador del total de productos. Calculado como:

$$(xti\_ahor\_fin\_ult1 + xti\_aval\_fin\_ult1 + xti\_cco\_fin\_ult1 + xti\_cder\_fin\_ult1 + xti\_cno\_fin\_ult1 + xti\_ctju\_fin\_ult1 + xti\_ctma\_fin\_ult1 + xti\_ctop\_fin\_ult1 + xti\_ctpp\_fin\_ult1 + xti\_deco\_fin\_ult1 + xti\_deme\_fin\_ult1 + xti\_dela\_fin\_ult1 + xti\_ecue\_fin\_ult1 + xti\_fond\_fin\_ult1 + xti\_hip\_fin\_ult1 + xti\_plan\_fin\_ult1 + xti\_pres\_fin\_ult1 + xti\_reca\_fin\_ult1 + xti\_tjcr\_fin\_ult1 + xti\_valo\_fin\_ult1 + xti\_viv\_fin\_ult1 + xti\_nomina\_ult1 + xti\_nom\_pens\_ult1 + xti\_recibo\_ult1)$$

- inc_total_productos_0_3m: Incremento del total de productos desde hace 3 meses. Calculado como:

$$\frac{(num\_prods\_total\_min3 - num\_prods\_total)}{num\_prods\_total\_min3}$$

- inc_total_productos_0_6m: Incremento del total de productos desde hace 6 meses. Calculado como:

$$\frac{(num\_prods\_total\_min6 - num\_prods\_total)}{num\_prods\_total\_min6}$$

- flag_cancelacion_total_0_3m: Indicador de cancelación de productos totales en los últimos 3 meses. Posibles valores: 1/0.

$$(num\_prods\_total\_min3 - num\_prods\_total), Si > 1 \rightarrow 1, Sino \rightarrow 0$$

- flag_cancelacion_total_0_6m: Indicador de cancelación de productos totales en los últimos 6 meses. Posibles valores: 1/0.

$$(num\_prods\_total\_min6 - num\_prods\_total), Si > 1 \rightarrow 1, Sino \rightarrow 0$$

- flag_contratacion_total_0_3m: Indicador de contratación de productos totales en los últimos 3 meses. Posibles valores: 1/0.

$$(num\_prods\_total - num\_prods\_total\_min3), Si > 1 \rightarrow 1, Sino \rightarrow 0$$

- flag_contratacion_total_0_6m: Indicador de contratación de productos totales en los últimos 6 meses. Posibles valores: 1/0.

$$(num\_prods\_total - num\_prods\_total\_min6), Si > 1 \rightarrow 1, Sino \rightarrow 0$$

- ***Depuración de los datos.***

  La mayoría de los algoritmos predictivos obtienen las estimaciones de probabilidad en base a ajustes de las medias de las variables numéricas. Esto hace que sean muy sensibles a los outliers (valores extremos o atípicos), debido a queun único valor muy lejano al rango de valores habituales puede distorsionar las medias y generar estimaciones que no tendrán tanto acierto para la población en global.

Por ello, es necesario realizar labores de depuración en los datos. Estas labores consisten en eliminar o sustituir ciertos registros del conjunto de entrenamiento, incluso sabiendo que son datos reales y correctos, pero incluirlos como son originalmente tendría efectos negativos en el acierto del modelo.

A continuación, se detallan los criterios de depuración que se han tenido en cuenta para la construcción del conjunto de entrenamiento.

- ***Tratamiento de valores ausentes***

  La presencia de valores ausentes o *missingvalues* es muy común cuando se trabaja con gran cantidad de datos. Para el modelo es importante tener en cuenta estos valores nulos porque ignorarlos o definir un criterio de tratamiento puede tener gran impacto en los modelos, tales como la pérdida de precisión o la presencia de sesgos importantes.

  Existen varias soluciones para tratar los valores ausentes. La opción por defecto es usar exclusivamente los datos informados, pero esta opción no siempre es adecuada porque puede excluir clases necesarias y perder poder de analítico. Otra alternativa es rellenar los valores nulos dependiendo del tipo de variable.

  - ***Categóricas:*** Para este tipo de variables, los valores nulos se sustituirán por un valor por defecto "NA".
  - ***Numéricas:*** Para este tipo de variables, los valores nulos se sustituirán siguiendo el siguiente criterio:
    - *Importe de renta.* En este caso se asume que la falta de este dato es por un error en la fuente origen y se calculará siguiendo el siguiente criterio:
      - Mediana de renta por provincia. Este criterio se aplicará para los clientes con provincia informada, representa el 42.19% de los registros. La renta se calculará teniendo en cuenta la mediana renta de la provincia en la que se encuentra el domicilio del cliente.
      - Mediana de renta global. Este criterio se aplicará para los clientes que no tienen informada la provincia, representa el 2.37% de los registros. La renta se

calculará teniendo en cuenta la mediana renta del total de clientes.

- ***Tratamiento de Outliers***

Debido a que el número de observaciones que se tiene en nuestro conjunto tiene un tamaño muy elevado, podemos sin perder generalidad y representatividad, eliminar o modificar los registros considerados outliers.

Una vez detectada la presencia de algunos valores que necesitan depuración hay que establecer un criterio para definir cuáles exactamente son los que se van a depurar. Hay dos opciones:

- Si se tiene el conocimiento de los límites que debe tener la distribución de la variable correcta se aplican estos límites y los valores que estén fuera se eliminan o sustituyen.
- Si no se conocen los límites de la distribución, se puede utilizar cualquiera de las técnicas estadísticas para la limpieza de *outliers*.

En el *dataset* de entrada se detectó la existencia de *outliers* para la variable edad, para la sustitución de los *outliers* se realizó un análisis de la distribución de las variables tal como se muestra en la siguiente figura:

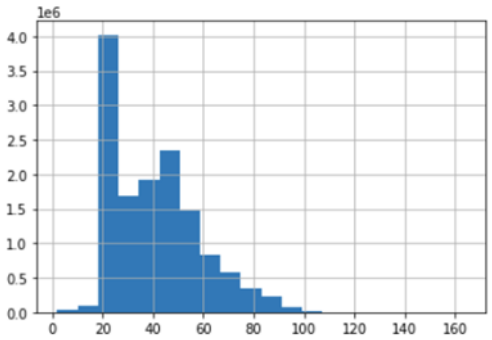


Figura 16. Histograma de Variable Edad (Elaboración Propia)

Posterior al análisis realizado se decide aplicar la siguiente sustitución de valores mayores a 100 por el valor de 100 y en el caso en caso de valores menores a 18 se sustituirán por 18.

Producto de esta fase se genera el segundo entregable que es el Documento de depuración (*Anexo 3. Documento de Depuración*), este documento contendrá el detalle de las decisiones tomadas para el tratamiento de valores nulos y outliers indicado anteriormente. Este documento también contendrá la selección de variables más relevantes para entrenar el modelo.

3.2.5.4. Fase 4. Modelado

- ***Construcción de Tablón de Estudio***

  El tablón de estudio será utilizado en todas las fases de la creación del modelo: entrenamiento, test y validación. Este tablón será producto de la uniónde los casos positivos y negativos. El tablón deberá contener:

  - Fecha de referencia.
  - Identificador del cliente.
  - Variables seleccionadas.
  - Variables predictoras.
  - Target o variable objetivo.

  Estos tablones de datos contendrán una única fila por cliente. Las fechas de referencia asociadas a cada cliente se calcularán de la siguiente forma:

  - **Para los casos positivos,** es decir los clientes que contratarán una tarjeta de crédito, se tomará como fecha de referencia dos meses antes al mes en el que ocurre la contratación. En el caso de clientes que en algún periodo dieron de alta una tarjeta en varios periodos. Para el caso positivo se tendrá en cuenta la fecha aleatoria de la fecha de referencia asociada a las contrataciones.
  - **Para los casos negativos**, es decir, clientes que no contratarán una tarjeta, se tomará como fecha de referencia cualquiera de los posibles meses del periodo de entrenamiento, eligiéndose esta fecha de manera aleatoria entre las posibles.

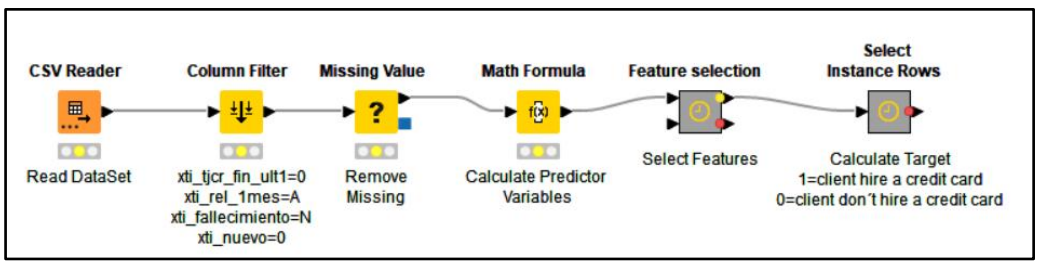


Figura 17. Workflow para Preparación de los datos (Elaboración Propia)

- ***Particionamiento de tablones: entrenamiento, test y validación.***

  Para la construcción de los tablones de entrenamiento y test se seleccionarán las fechas de referencia de junio 2018 a marzo 2019. Se excluirá del tablón de entrenamiento el 20% de los clientes y estos formarán el conjunto de test. Estos clientes no entrarán en el algoritmo de aprendizaje y servirán para medir el modelo sobre población distinta a la usada en entrenamiento. Para el tablón validación se seleccionará la fecha de referencia de marzo 2019.

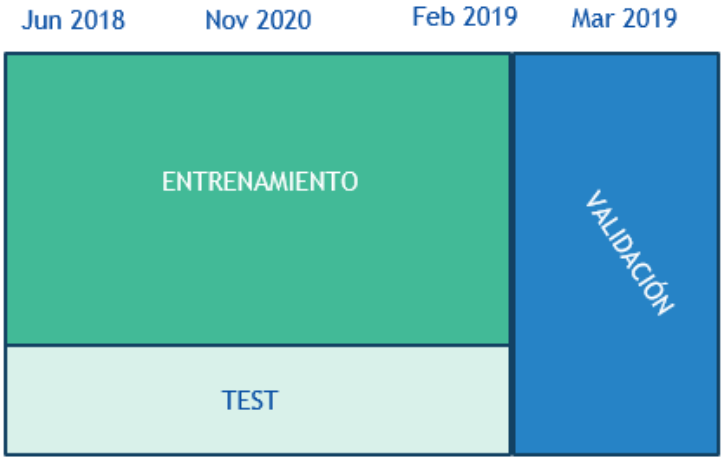


Figura 18. Particionamiento de Tablones de Entrenamiento, Test y Validación (Elaboración Propia)

- ***Selección del algoritmo.***

  Para el trabajo descrito se construyó del modelo de predicción aplicando tres tipos de algoritmos de clasificación: Árbol de Decisión, *RandomFforest*, *GradientBoosting*.

  - *Modelo predictivo de Árbol de Decisión.*

    Para la implementación del modelo predictivo aplicando el algoritmo de Árbol de Decisión (*Apartado 3.2.4.2.*) se utilizará la librería sklearn.tree.DecisionTreeClassifier [SCIKIT-LEARN s.f.-a], con los parámetros:

    max_depth=4

    min_weight_fraction_leaf=0.04

    Ambos parámetros se configuraron para limitar el tamaño de las hojas y evitar el overfitting.

    El siguiente gráfico muestra el flujo aplicado para la implementación del modelo predictivo empleando el algoritmo del árbol de decisión en el que luego del particionamiento de los datos pre-procesados se empleará el 80% de

ellos para el entrenamiento y el 20% para el test del modelo. El resultado de la ejecución del modelo será el fichero con el score por cada cliente del *dataset* de entrada.

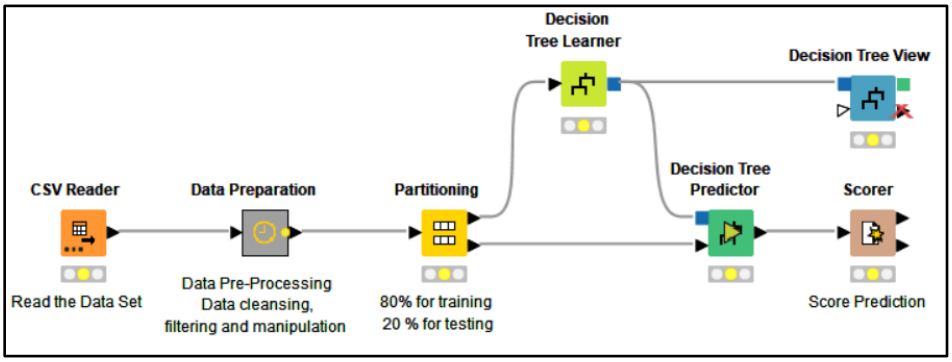


Figura 19. Workflow del Árbol de Decisión (Elaboración Propia)

Posterior al entrenamiento del modelo se obtuvo la representación gráfica del árbol de decisión, Figura 20. Esta representación gráfica nos permitióidentificar los patrones tomados en cuenta para la predicción, así como también analizar las variables con mayor poder predictivo. Analizando los resultados podemos analizar las variables seleccionadas por el algoritmo:

- *inc_productos_total_0_3_m*: Incremento del total de productos desde hace tres meses al mes actual. Esta variablediscrimina aquellos clientes que tiene una tendencia de contratación de productos.
- *num_prods_total:* Sumatorio del total de productos contratados por el cliente. Estavariable discrimina a los clientes según los productos contratados.
- *flag_cancelacion_total_0_3m:* Indicador que se marca a 1 en el caso que en hace tres meses el cliente haya realizado una cancelación de algún producto y 0 en caso contrario.Esta variable discrimina a los clientes según su tendencia de cancelación de productos.
- *xti_recibo_ult1:* Indicador se marca a 1 en el caso de que el cliente tenga contratado el Débito Directo y 0 en caso contrario. Esta variable también discrimina a los clientes según los productos contratados.
- *xti_ctop_fin_ult1:* Indicador se marca a 1 en el caso de que el cliente tenga contratado la *C*uenta Particulary 0 en caso contrario.Esta variable también discrimina a los clientes según los productos contratados.

- *xti_cno_fin_ult1:* Indicador se marca a 1 en el caso de que el cliente tenga contratado la *Cuenta Nómina* y 0 en caso contrario.Esta variable también discrimina a los clientes según los productos contratados.

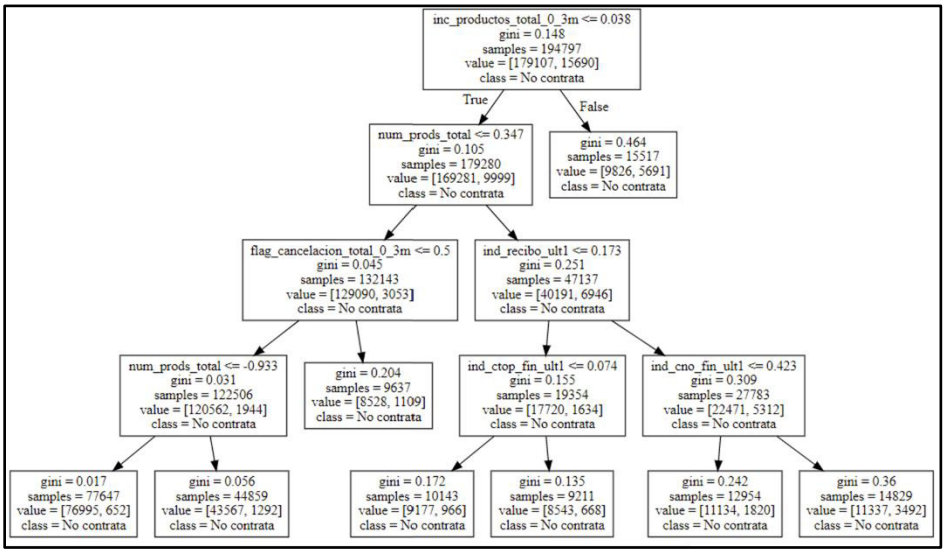


Figura 20. Representación Gráfica de Árbol de Decisión (Elaboración Propia)

- <u>*Modelo predictivo de Random Forest*</u>.

  Para la implementación del modelo predictivo aplicando el algoritmo de Random Forest (*Apartado 3.2.4.2.*) se utilizará la librería sklearn.ensemble.RandomForestClassifier [SCIKIT-LEARN s.f.-b], con los parámetros:

  max_depth=4

  n_estimators=30

  El parámetro max_depth limita la profundidad del árbol para evitar *overfitting* y el n_estimators representa la cantidad de árboles decisión que se generarán, en este caso se limita el número de árboles para no ralentizar la ejecución del modelo.

  El siguiente gráfico muestra el flujo aplicado para la implementación del modelo predictivo empleando el algoritmo de *RandomForest* en el que luego del particionamiento de los datos pre-procesados se empleará el 80% de ellos para el entrenamiento y el 20% para el test del modelo. El resultado de la

ejecución del modelo será el fichero con el score por cada cliente del *dataset* de entrada.

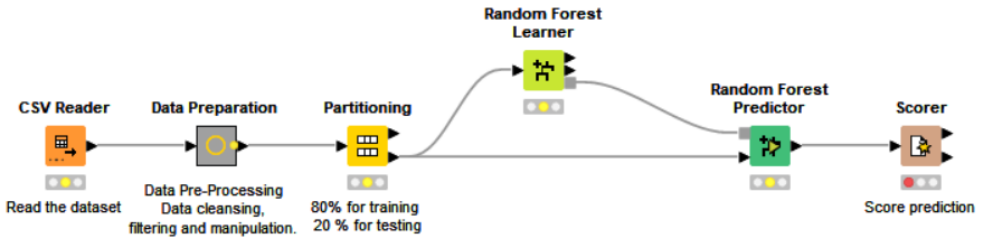


Figura 21. Workflow del Random Forest (Elaboración Propia)

- *Modelo predictivo de GradientBoosting*

  Para la implementación del modelo predictivo aplicando el algoritmo de *GradientBoosting* (*Apartado 3.2.4.2.*) se utilizará la librería sklearn.ensemble.GradientBoostingClassifier [SCIKIT-LEARN s.f.-c], con los parámetros:

  max_depth=4

  n_estimators=30

  learning_rate=0.1 (valor por defecto)

  Parámetros similares utilizados en los algoritmos anteriores. Adicionalmente el parámetro learning_rate se deja con el valor por defecto, este parámetro permite controlar que tan rápido aprende el modelo y también el riesgo de llegar al *overfitting*.

  El siguiente gráfico muestra el flujo aplicado para la implementación del modelo predictivo empleando el algoritmo de *RandomForest* en el que luego del particionamiento de los datos pre-procesados se empleará el 80% de ellos para el entrenamiento y el 20% para el test del modelo. El resultado de la ejecución del modelo será el fichero con el score por cada cliente del *dataset* de entrada.

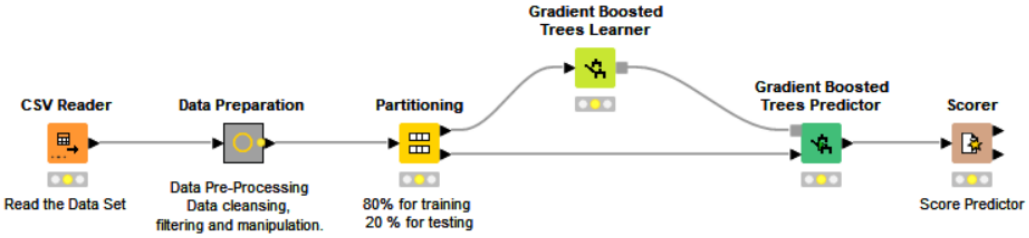


Figura 22. Workflow de GradientBoosting (Elaboración Propia)

3.2.5.5 Fase 5. Evaluación

Posterior a la ejecución de los tres algoritmos se evaluaron los resultados de los algoritmos. La métrica utilizada para la evaluación de los modelos es el AUC, cuyas siglas en español significan Área Bajo la Curva ROC (*Apartado 3.2.4.3.*). Como muestra la Figura 23, los algoritmos *Random Forest* y *GradientBoosting* tienen los mejores resultados.

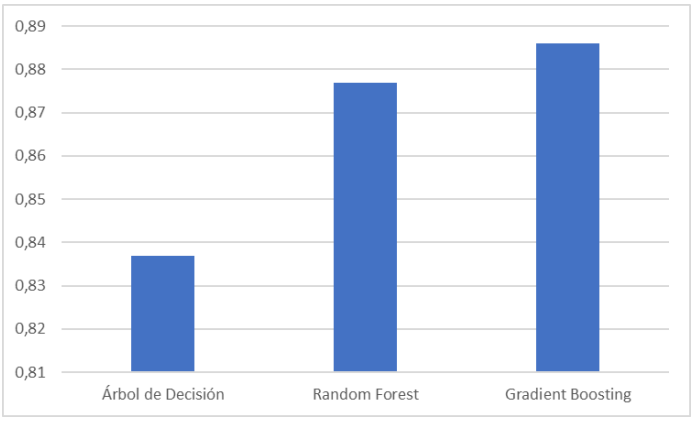


Figura 23. Comparativa Métrica AUC (Elaboración Propia)

Para seleccionar los algoritmos de *Random Forest* y *GradientBoosting* se tuvo en cuenta los resultados de AUC entre los modelos con datos de *"train"* y los datos de "*test*". A pesar que el algoritmo de *Random Forest* tiene menos AUC que el algoritmo de *GradientBoosting*, la diferencia entre los resultados de *"train"* y "test" es menor a diferencia del otro algoritmo. Esto quiere decir que el modelo será más estable en el tiempo y no necesitará reentrenar en un corto periodo.

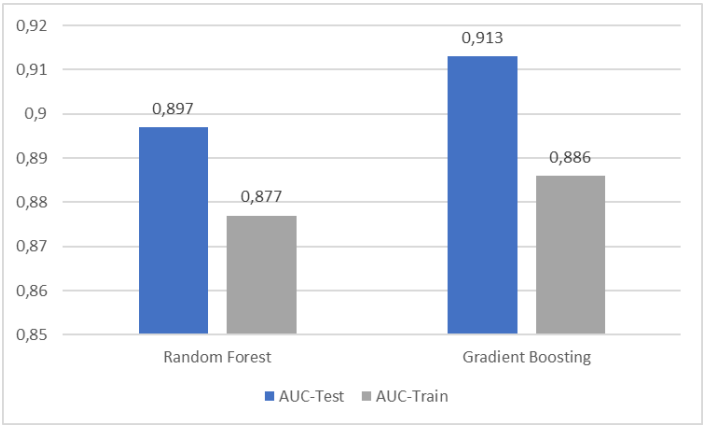


Figura 24. Comparativa de Métrica AUC para "Train" y "Test" (Elaboración Propia)

Otro punto que se ha tenido en cuenta en la selección del algoritmo es el tiempo de ejecución, el tiempo de ejecución del algoritmo *Random Forest* es mucho menor al algoritmo de *GradientBoosting*y la ganancia al seleccionar este algoritmo no es significativa por lo que finalmente se selecciona el algoritmo de *Random Forest* por ser más estable en el tiempo y por su mejor rendimiento de ejecución.

El modelo genera un score de clientes según su propensión a contratar una tarjeta de crédito, tal como se muestra en la imagen a continuación:

	cod_cliente	score
0	75632	0.070884
1	41285	0.071062
2	29662	0.038746
3	87141	0.667298
4	87137	0.459390

Figura 25. Muestra de Score de Clientes (Elaboración Propia)

3.2.5.5. Fase 6. Despliegue

Dentro de la entidad bancaria se siguieron los procedimientos definidos por la organización para la implantación de la solución. La Figura 26 muestra la arquitectura definida para la solución dentro del ecosistema de la entidad.

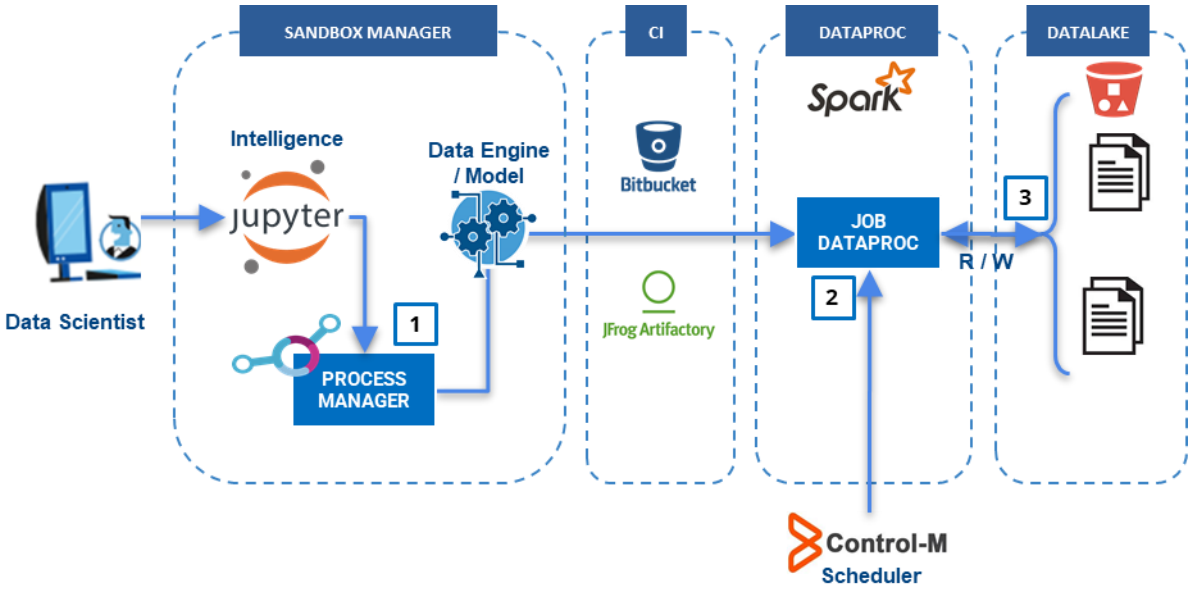


Figura 26Arquitecturadel Modelo de Solución (Elaboración Propia)

Dentro del modelo de solución (Figura 26), las fases para la implantación de la solución son las siguientes:

1. Posterior a la implementación del modelo nos encargamos de crear un Motor de Datos o *Data Engine* mediante *Process Manager*. El Motor de datos es un Job de *DataProc*solo ejecutable en el entorno Sandbox y el código se publicó en un repositorio de *Bitbucket*. Ese motor de datos es un archivo ejecutable .jar en *Artifactory*, dentro del repositorio asignado para el proyecto.
2. Una vez que se realizaron las pruebas y estas fueron satisfactorias se creó el Job que utiliza el *ProcessManager* para que invoque a la API de *DataProc*.Una vez publicado el Job se programó para que se pueda ejecutar en ambiente productivo pormedio de Control-M. En el caso del modelo se deberá ejecutar posterior a la generación de la vista en el *DataLake*. Esta productivización la realizó el área de *AdvanceAnalytics* de la entidad.
3. La lectura del input del modelo se realizadesde el *DataLake*, donde mensualmente se generará la vista de Clientes-Productos necesaria para la ejecución del modelo. La ejecución del modelo persistirá el score resultante en el *DataLake* para posteriormente ser consumido por el área de negocio.

A continuación, se detallan los componentes propios de la organización que intervienen en la arquitectura de la solución:

- **Sandbox:** En el contexto de Big Data, es una plataforma escalable y de desarrollo que se utiliza para explorar los grandes conjuntos de información de una organización a través de la interacción y la colaboración. Trabajar en un entorno Sandbox permite a los *Data Scientist* trabajar en una plataforma Big Data basada en *Spark*, dimensionada para satisfacer todas las necesidades. El entorno de producción está aislado y para entornos previos proporciona un servidor de pruebas, un servidor de desarrollo o un directorio de trabajo.
- **Sandbox Manager:** Sandbox Manager es el principal espacio de trabajo de los analistas de la entidad, encapsula herramientas de *Big Data, Data Science*, Inteligencia artificial&*Business Intelligence*para trabajar con datos reales con los requisitos de seguridad implementados.
- **Data Lake:** Es un repositorio centralizado que almacena todos los datos estructurados y no estructurados de la organización. Adicionalmente también permite compartir y gobernar datos en la organización.
- **Dataproc:** Es un servicio administrado de Spark y Hadoop con el que puede aprovechar herramientas de datos de código abierto para procesamientos por lotes, búsquedas, transmisiones y aprendizaje automático.
- **Process Manager:**Herramienta de la organización, permite desplegar proyectos de forma más rápida y con una curva de aprendizaje mínimaen comparación con el flujo de trabajo común anterior. El objetivo principal de este documento es ayudar al usuario a productivizar un código a partir de un notebook.
- **Control-M:** La herramienta utilizada para programar trabajos de la API de Dataproc. La programación determina cuándo debe comenzar o terminar una actividad, según su duración, la actividad (o actividades) predecesora, las relaciones predecesoras, la disponibilidad de recursos y la fecha prevista de finalización del proyecto.

El modelo de predicción fue desarrollado en *Python* 3.6 y el *Pipeline* de ejecucióndel modelo consta de tres fases el preprocesamiento de los datos, la construcción de variables predictoras y la ejecución del modelo entrenado, tal como se muestra en la siguiente figura:

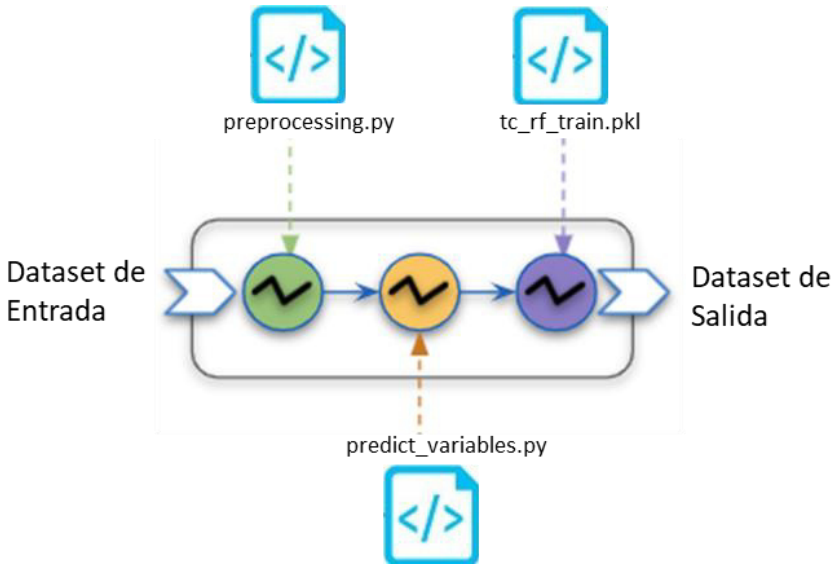


Figura 27. Pipeline Modelo Predictivo de Tarjetas de Crédito (Elaboración Propia)

3.3. Evaluación

3.3.1. Evaluación de los modelos de predicción

Tal como se ha descrito en el apartado anterior, para la construcción del modelo de predicción de contratación de tarjetas se seleccionaron tres algoritmos comúnmente usados en problemas de clasificación. Los algoritmos seleccionados fueron: Árbol de Decisión, *Random Forest* y *GradientBoosting*.

Posterior a las fases de entrenamientoy *test* de cada modelo, se aplicaron las métricas de *Accuracy* y AUC para evaluar los modelos. Con la métrica de Accuracy medimos la precisión de la clasificación del modelo y con el AUC el poder predictivo del modelo.

Tal como se muestra en la Figura 28, los resultados obtenidos en los tres modelos son muy buenos y para la selección se tuvieron en cuenta los resultados de las métricas y las características propias de la entidad donde se iba a aplicar la solución. Por ello, finalmente se seleccionó el algoritmo de *Random Forest*.

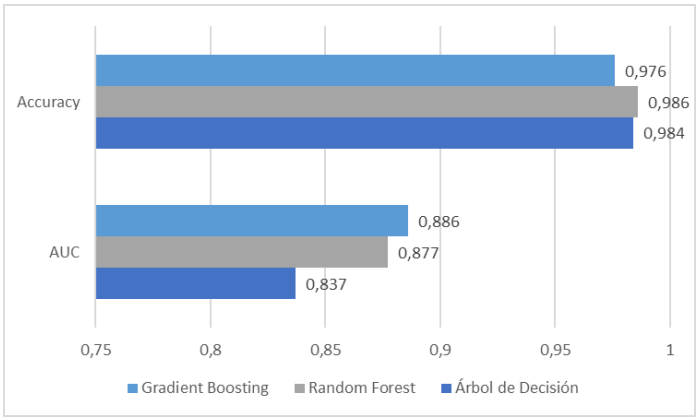


Figura 28. Comparativa de Métricas Accuracy y AUC (Elaboración Propia)

Los resultados del modelo implementado aplicando el algoritmo de Random Forest tiene un Accuracy de 0,986. Esto quiere decir que el modelo es capaz de predecir correctamente un 98.6% de las observaciones. Aplicando la métrica de AUC se obtuvo un 8,77. Interpretando el resultado podemos afirmar que hay un 87.7% de probabilidad de que el modelo pueda distinguir entre los clientes que son propensos a contratar una tarjeta y los que no. Además,

según la clasificación del poder predictivo de un modelo de predicción según la métrica de AUC, el modelo implementado está clasificado dentro del rango de "Modelo bueno".

El modelo implementado fue muy exitoso porque permitió identificar el perfil correcto para la venta de tarjetas de crédito y así orientar las campañas comerciales a estos clientes, optimizando la efectividad de las campañas.

3.3.2. Evaluación económica

3.3.2.1. Coste del proyecto

El proyecto tuvo una duración de tres meses y en él participaron dos recursos con el perfil de *Data Scientist*. A continuación, se muestra el total de horas invertidas para el desarrollo del proyecto:

Tabla 2. Tabla del Total de Horas Invertidas en el Proyecto (Elaboración Propia)

Perfiles	Unidades	% Total de Asignación	# Total Horas
Data Scientist 1	1	20	96
Data Scientist 2	1	100	480
Total	**1**	**120**	**576**

El coste total del proyecto para la entidad bancaria se calcula aplicando la tarifa de la empresa Bluetab, que tiene acordada con la entidad una tarifa de 60.00 € horas/hombre. En la Tabla 3 se indica el coste total del proyecto, que culminará con la entrega del desarrollo del modelo.

Tabla 3. Coste Total del Proyecto (Elaboración Propia)

	Coste
Modelo de Predicción	34.560,00 €
Otros Costes	0,00 €
IVA	7.257,60 €
Total	**41.817,60 €**

3.3.2.2. Beneficios para la entidad

Posterior a la implantación del modelo, el área de Inteligencia Comercial de la entidad aplicó el score generado para priorizar sus acciones comerciales en base a la propensión generada. Los resultados obtenidos a finales del 2019 producto de la aplicación del modelo fueron muy prometedores. La aplicación del modelo permitió mejorar la efectividad en un 15%. El modelo predictivo aumentó las ventas de tarjetas de crédito,superando el KPI marcado para ese año, así como el porcentaje de efectividad marcado en2,7%.

En la Figura 28 se muestra una comparativa de las ventas de las campañas de tarjetas de crédito de los últimos 4 meses delaño 2018 y 2019.En esta comparativa se observa que la efectividadpromedio de las campañas en 2018 es de 2,42% y la venta promedio es de 326.355 unidades al mes. En cambio, en el2019 con el uso del modelo, la efectividad de las campañas se incrementó a 2,78% con una media de 379.007 unidades de venta mensual. Esto supone un incremento medio en la venta de tarjetas de crédito de 52.000 unidades mensuales.

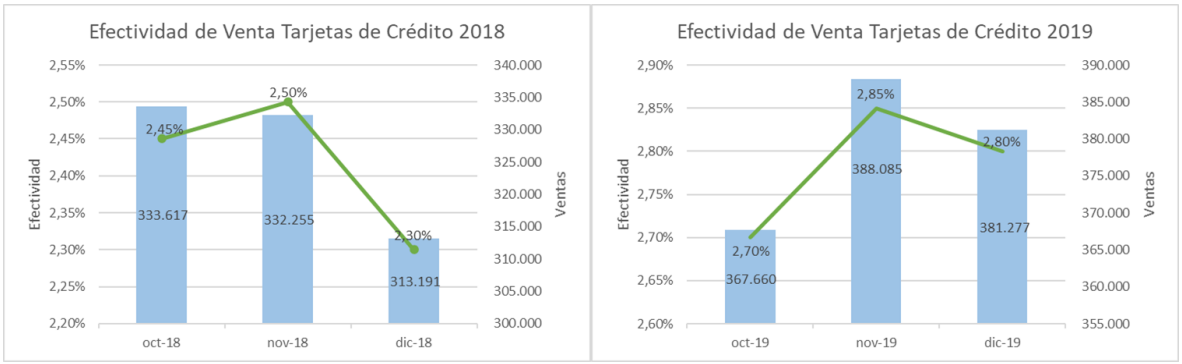


Figura 29. Comparativa de Efectividad en la Venta de Tarjetas de Crédito 2018-2019 (Elaboración Propia)

El impacto económico del proyecto para la entidad lo calculamos teniendo en cuenta el PRV (Valor Relativo del Producto) estimado de una tarjeta para la entidad que tiene el valor de 1.329€. Esto quiere decir que el PRV estimado anual es de69.108.000 €.

CAPÍTULO IV. REFLEXIÓN CRÍTICA DE LA EXPERIENCIA

El uso de una metodología adecuada a un proyecto puede marcar la diferencia entre éxito y fracaso del proyecto. En el desarrollo del proyecto fue acertada la selección de la metodología CRISP-DM porque aportó una visión global y alineada a los objetivos del negocio.

En el desarrollo del proyecto, fue muy importante mantener una buena comunicación con el usuario de la entidad, sobre todo en las fases iniciales donde se requiere de su conocimiento de negocio y se definen los resultados esperados. La presentación Diseño Preliminar, para su posterior aprobación, fue vital para aclarardudas o interpretaciones erróneas en la definición.

Dentro de la fase de definición de nuevas variables predictores fue acertada la comunicación con el usuario de negocio, aplicamos la experiencia del usuario para la creación de nuevas variables y resultaron ser de gran poder predictivo en el modelo.

En la presentación de los resultados del modelo fue necesario explicar conceptos teóricos a los usuarios de la entidad para que comprendan la justificación del algoritmo seleccionado, así como también la evaluación realizada para garantizar que el modelo cumple con las expectativas.

La participación en este proyecto me permitió aplicar mis conocimientos teóricos en técnicas de *Machine Learning*. En el proyecto también participó un compañero senior en esta disciplina, lo cual me permitió aprender de su experiencia y apoyarme en él ante cualquier duda en la ejecución del proyecto.

CAPÍTULO V. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

- El modelo de aprendizaje supervisado para la predicción de contratación de tarjetas de crédito obtuvo una eficacia de 0,897en AUC (Área Bajo la Curva ROC).Este resultado quiere decir que hay un 89,7% de probabilidad de que el modelo pueda distinguir entre los clientes propensos a contratar una tarjeta de crédito y los que no.

- Se tomó la decisión de seleccionar el algoritmo de *Random Forest* para implementar el modelo de clasificación. Como criterios de selección, no solo se tuvo en cuenta los resultados de las métricas, sino también la diferencia entre los resultados de las métricas con datos de "*train"* y "*test".* Mientras menor sea la diferencia entre estos resultados, mayor será la durabilidad del modelo.

- La aplicación del modelo predictivo para incrementar la contratación de tarjetas de crédito tuvo resultados muy positivos, cumpliendo los objetivos marcados. En el año 2019 se logró incrementar la efectividad de las campañas a 2,78% y la venta de tarjetas a 379.007 unidades en promedio al mes. Esto supuso un incremento en la venta de tarjetas en 52.000 unidades mensuales con respecto a las ventas previas a la implantación del modelo.

- Adicionalmente, la solución permitió identificar las variables con mayor poder predictivo para el modelo de predicción de contratación de tarjetas de crédito, pero también permitió identificar variables que podrían aplicarse en la resolución de otros problemas como la ratio de cancelación del producto. Este punto incrementa el valor del proyecto desarrollado para la entidad.

5.2. Recomendaciones

- Para futuras iteraciones del modelo, se sugiere una activa participación del área de negocio para compartir los conocimientos obtenidos en la aplicación del modelo predictivo para el desarrollo de nuevas accionescomerciales.

- El conocimiento obtenido en la implementación del modelo predictivo, es aplicable a nuevos casos de uso para la entidad, tales como aplicar el modelo para pronosticar contratación de otros productos o identificar clientes propensos a cancelar un producto.

- Se sugiere incluir nuevas fuentes datos de orígenes externos a la entidad como datos de navegación, redes sociales o herramientas de CRM (CustomerRelationship Management) para obtener mayor conocimiento del cliente y ser capaces de desarrollar algoritmos de predicción más potentes aplicables a por ejemplo recomendar productos al cliente o identificar cuellos de botella que evitan que el cliente cierre una contratación.

FUENTES DE INFORMACIÓN

- *aprendeIA*. (2021). Obtenido de https://aprendeia.com/curvas-roc-y-area-bajo-la-curva-auc-machine-learning/

- Arana, C. (2021). *Modelos de Aprendizaje Automático Mediante Árboles de Decisión.*

- Boehmke , B., & Greenwell, B. (2020). *https://bradleyboehmke.github.io/HOML/gbm.html*. Obtenido de Hands-On Machine Learning with R. Chapter 12 Gradient Boosting.

- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*.

- Breiman, L. (2001). *Machine Learning. Random Forests.*

- Cimarra Muñoz, D. (2018). *Experimientos de Predicción con Gradient Boositng y Random Forest.*

- Ejea Carbonell, D. G. (2017). *Árboles de Regresión. Algunos algoritmos.*

- Espinosa-Zúñiga, J. (2020). *Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito.*

- Fernández, A., López, V., Galar, M., José del Jesus, M., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *ScienceDirect*.

- *Kaggle*. (s.f.). Obtenido de https://www.kaggle.com/.

- Medina Merino, R. F., & Ñique Chacón, C. I. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. *Dialnet*.

- Mitchell, T. M. (1990). *Machine Learning.*

- Pallarés, Á. A. (2019). *Aplicación y comparación de modelos de machine learning destinados a la puntuación del riesgo de crédito.*

- Trujillo Fernandez, D. (2017). *Aplicación de Metodologías Machine Learning en la Gestión de Riesgo de Crédito.*

- Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining.*

GLOSARIO

- **Big Data:**Se trata de un conjunto de datos muy grande en el que se incluyen datos estructurados y no estructurados, siendo difícil de gestionar con medios habituales. Además, debe cumplir conlas tres “V”: volumen, variedad y velocidad.
- **Data Lake:** Es un repositorio centralizado que almacena todos los datos estructurados y no estructurados de la organización. Es utilizado por *Data Scientist*para manipular datos.
- **Data Scientist:**Es aquel experto que crea código de programación y lo combina con conocimientos estadísticos para crear conocimientos sobre datos del negocio.
- **Business Intelligence:** Es el método que transforma y analiza datos almacenados de la organización con el objetivo de generar información estratégica para una organización.
- **Machine Learning:** El aprendizaje automático, una rama de la inteligencia artificial, se refiere a la construcción y estudio de sistemas que pueden aprender de los datos.
- **Dataset**:Un conjunto de datos es una colección de datos. Generalmente, un conjunto de datos corresponde al contenido de una tabla de base de datos.
- **Algoritmo:** En Ciencias de la Computación, un algoritmo es una secuencia lógica, finita y con instrucciones que forman una fórmula matemática o estadística para realizar el análisis de datos.
- **Overfitting:** Crear un modelo que coincida tan estrechamente con los datos de entrenamiento que el modelo no pueda hacer predicciones correctas sobre nuevos datos, en otras palabras, no se generalice.
- **KPIS:** Indicador de medición.
- **Efectividad:**Es un indicador que define cual es el impacto que tiene una o varias campañas comerciales. Calculada a partir del cociente del número de ventas realizadas entre el número de intentos realizados para vender el producto.
- **Campaña Comercial:**Conjunto de estrategias comerciales enfocadas a conseguir un objetivo específico, tales como: captar, fidelizar o retener un cliente.

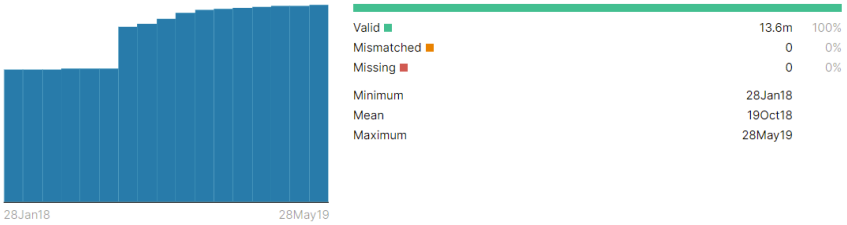
ANEXOS

Anexo 1: Análisis Univariante

Para la representación del análisis univariante de las variables de entrada nos hemos apoyado en la herramienta (Kaggle, s.f.). Los resultados obtenidos en este análisis son:

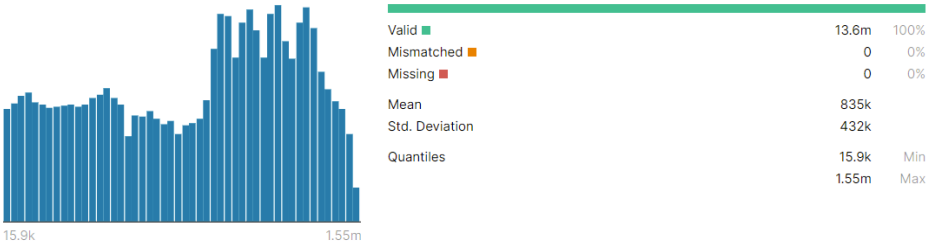
- *Variables asociadas a la extracción de los datos:*

fec_dato: Fecha de generación de los datos.

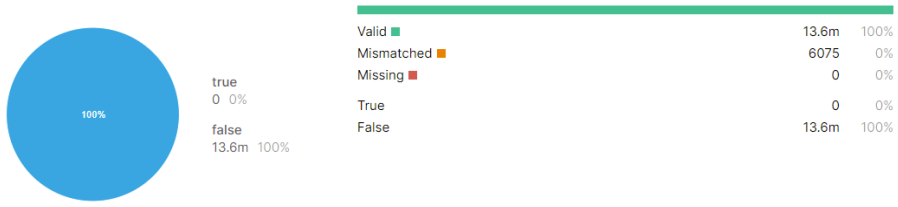


- *Variables asociadas a las características del cliente:*

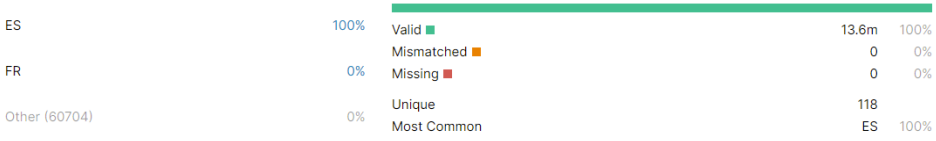
cod_cliente: Identificador único del cliente en la entidad.



xti_empleado: Indicador de empleado.



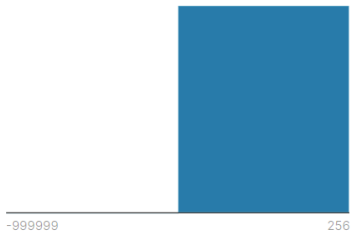
nom_pais_resid: Nombre de país de residencia del cliente.



xti_nuevo: Indicador de nuevo cliente (antigüedad menor a 6 meses).

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.06	
Std. Deviation	0.24	
Quantiles	0	Min
	1	Max

num_antiguedad: Antigüedad en meses.

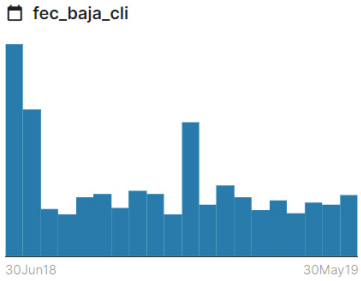


Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	76.6	
Std. Deviation	1.67k	
Quantiles	-999999	Min
	256	Max

xti_estado: Indicador de estado de clientetitular.

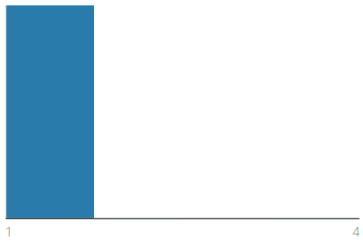
Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.18	
Std. Deviation	4.18	
Quantiles	1	Min
	99	Max

fec_baja_cli: Fecha de baja del cliente.



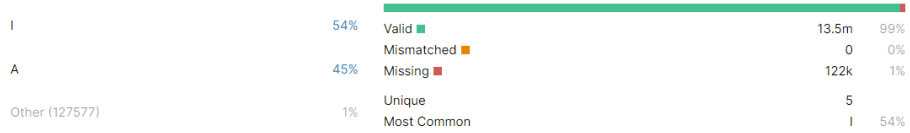
Valid	24.8k	0%
Mismatched	0	0%
Missing	13.6m	100%
Minimum	30Jun18	
Mean	18Nov18	
Maximum	30May19	

xti_tipo_cliente: Tipo de cliente.

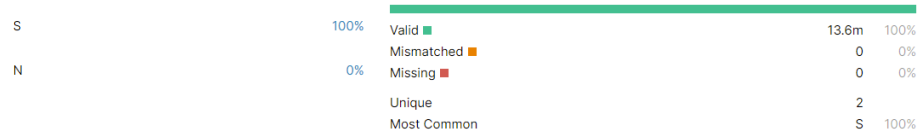


Valid	13.5m	99%
Mismatched	874	0%
Missing	122k	1%
Mean	1	
Std. Deviation	0.04	
Quantiles	1	Min
	4	Max

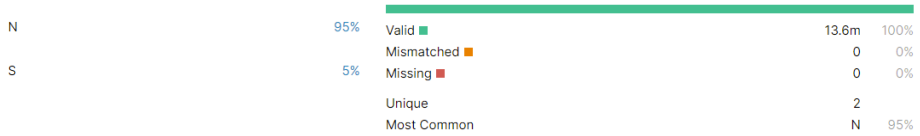
xti_relacion: Relación del cliente con la entidad.



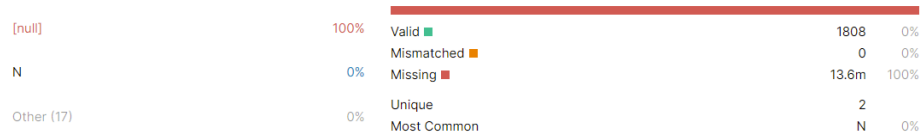
xti_residencia: Indicador de residencia en el país de la entidad.



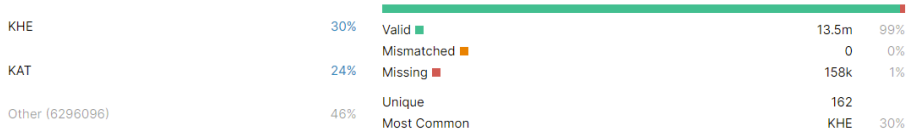
xti_extranjero: Indicador residencia en elextranjero.



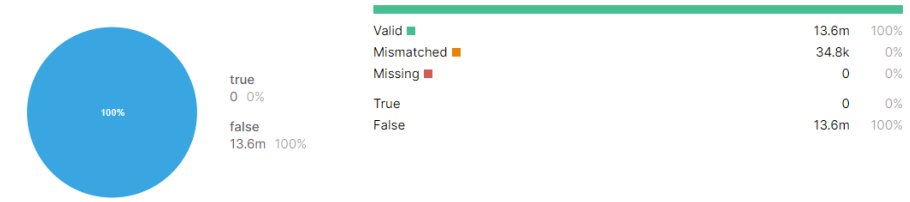
xti_conyugue: Indicador de conyugue.



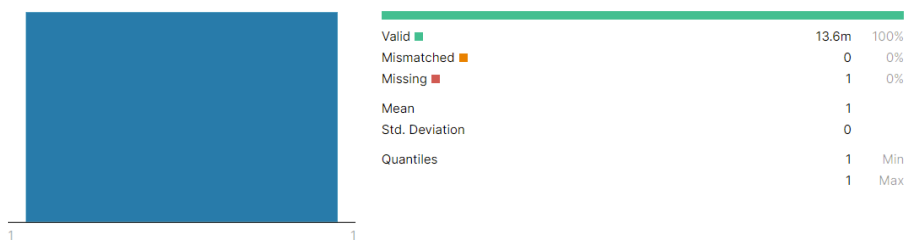
nom_canal: Siglas que representan el canal de contacto del cliente.



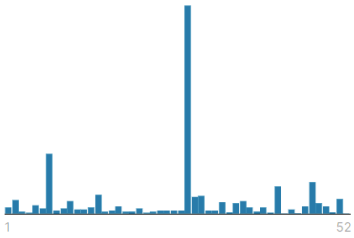
xti_fallecimiento: Indicador de fallecimiento.



xti_domicilio: Indicador de primer domicilio informado.



cod_provincia: Código de provincia donde se encuentra el domicilio.



Valid	13.6m	100%
Mismatched	0	0%
Missing	65.9k	0%
Mean	26.6	
Std. Deviation	12.8	
Quantiles	1	Min
	52	Max

nom_provincia: Nombre de provincia.

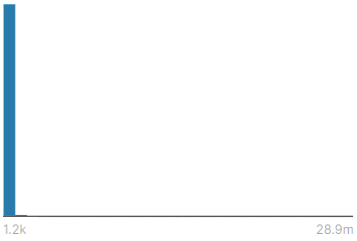
MADRID	32%
BARCELONA	9%
Other (7934756)	58%

Valid	13.6m	100%
Mismatched	0	0%
Missing	65.9k	0%
Unique	52	
Most Common	MADRID	32%

xti_activo: Indicador de cliente activo.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.46	
Std. Deviation	0.5	
Quantiles	0	Min
	1	Max

imp_renta: Importe de renta anual del cliente.



Valid	10.9m	80%
Mismatched	0	0%
Missing	2.77m	20%
Mean	134k	
Std. Deviation	231k	
Quantiles	1.2k	Min
	28.9m	Max

cod_segmento: Código de segmento al que pertenece el cliente.

02 - PARTICULARES	58%
03 - UNIVERSITARIO	36%
Other (723776)	5%

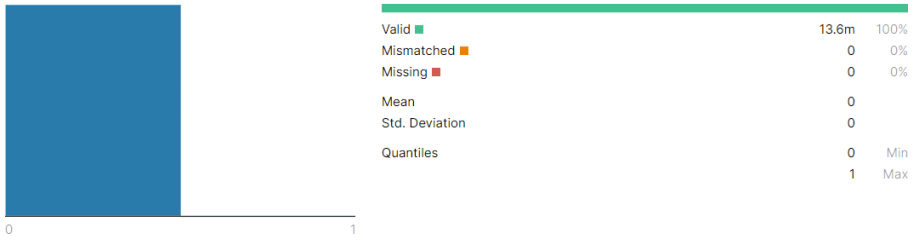
Valid	13.5m	99%
Mismatched	0	0%
Missing	162k	1%
Unique	3	
Most Common	02 - PARTIC...	58%

- *Variables asociadas a la cartera de productos del cliente:*

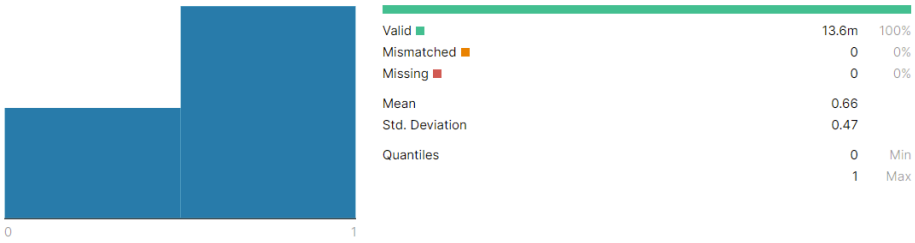
xti_ahor_fin_ult1: Indicador de cuenta de ahorro.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0	
Std. Deviation	0.01	
Quantiles	0	Min
	1	Max

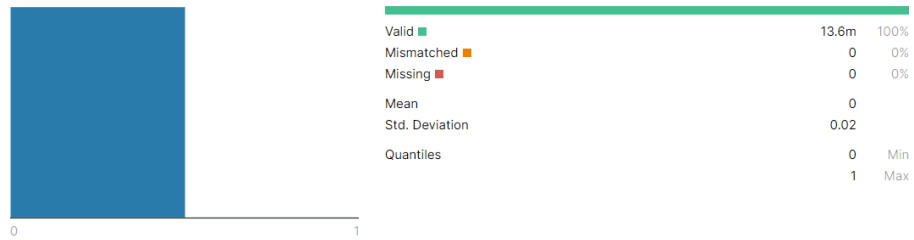
xti_aval_fin_ult1: Indicador de cuenta de garantías.



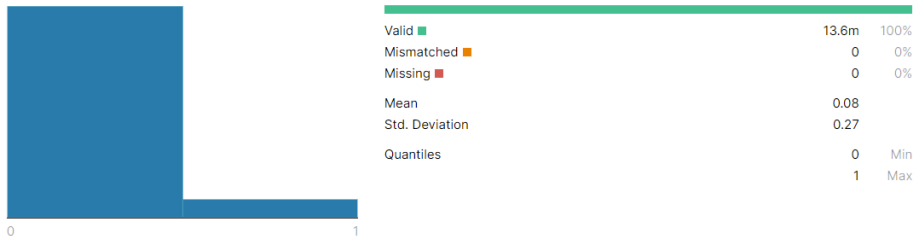
xti_cco_fin_ult1: Indicador de cuenta corriente.



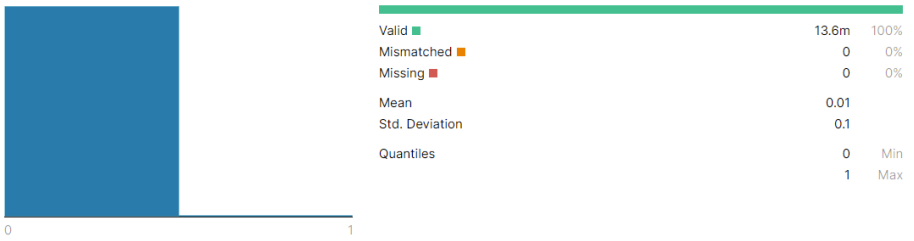
xti_cder_fin_ult1: Indicador de cuenta derivada.



xti_cno_fin_ult1: Indicador de cuenta nómina.



xti_ctju_fin_ult1: Indicador de cuenta junior.



xti_ctma_fin_ult1: Indicador de cuenta más particular.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.01	
Std. Deviation	0.1	
Quantiles	0	Min
	1	Max

xti_ctop_fin_ult1: Indicador de cuenta particular.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.13	
Std. Deviation	0.34	
Quantiles	0	Min
	1	Max

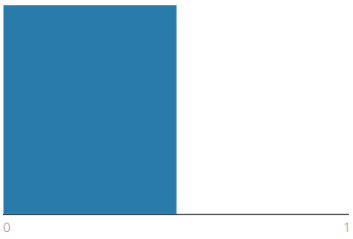
xti_ctpp_fin_ult1: Indicador de cuenta de particular plus.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.04	
Std. Deviation	0.2	
Quantiles	0	Min
	1	Max

xti_deco_fin_ult1: Indicador de depósitos a corto plazo.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0	
Std. Deviation	0.04	
Quantiles	0	Min
	1	Max

xti_deme_fin_ult1: Indicador de depósitos a medio plazo.



Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0	
Std. Deviation	0.04	
Quantiles	0	Min
	1	Max

xti_dela_fin_ult1: Indicador de depósitos a largo plazo.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.04	
Std. Deviation	0.2	
Quantiles	0	Min
	1	Max

xti_ecue_fin_ult1: Indicador de cuenta online.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.08	
Std. Deviation	0.28	
Quantiles	0	Min
	1	Max

xti_fond_fin_ult1: Indicador de fondos.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.02	
Std. Deviation	0.13	
Quantiles	0	Min
	1	Max

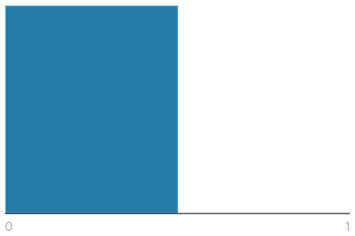
xti_hip_fin_ult1: Indicador de hipoteca.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.01	
Std. Deviation	0.08	
Quantiles	0	Min
	1	Max

xti_plan_fin_ult1: Indicador de pensiones.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.01	
Std. Deviation	0.1	
Quantiles	0	Min
	1	Max

xti_pres_fin_ult1: Indicador de préstamos.



Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0	
Std. Deviation	0.05	
Quantiles	0	Min
	1	Max

xti_reca_fin_ult1: Indicador de impuestos.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.05	
Std. Deviation	0.22	
Quantiles	0	Min
	1	Max

xti_tjcr_fin_ult1: Indicador de tarjeta de crédito.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.04	
Std. Deviation	0.21	
Quantiles	0	Min
	1	Max

xti_valo_fin_ult1: Indicador de valores.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.03	
Std. Deviation	0.16	
Quantiles	0	Min
	1	Max

xti_viv_fin_ult1: Indicador de cuenta de vivienda.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0	
Std. Deviation	0.06	
Quantiles	0	Min
	1	Max

xti_nomina_ult1: Indicador de nómina.

Valid	13.6m	100%
Mismatched	0	0%
Missing	217	0%
Mean	0.05	
Std. Deviation	0.23	
Quantiles	0	Min
	1	Max

xti_nom_pens_ult1: Indicador de pensiones.

Valid	13.6m	100%
Mismatched	0	0%
Missing	217	0%
Mean	0.06	
Std. Deviation	0.24	
Quantiles	0	Min
	1	Max

xti_recibo_ult1: Indicador de débito directo.

Valid	13.6m	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.13	
Std. Deviation	0.33	
Quantiles	0	Min
	1	Max

Anexo 2: Diseño Preliminar

1. Objetivo del documento

El objetivo del presente documento es recoger la definición a alto nivel del diseño del modelo predictivo de Contratación de Tarjetas de Crédito para una entidad financiera, incluyendo el objetivo del modelo, las fuentes de datos a utilizar, el target, horizonte de predicción y otras premisas relevantes.

El diseño y desarrollo del modelo se realizará según las premisas recogidas en el presente documento.

2. Objetivos principales del estudio

El objetivo del estudio es desarrollar un algoritmo predictivo capaz de localizar patrones comunes de clientes propensos a la contratación de una tarjeta de crédito dos meses después a la ejecución del algoritmo. La salida de este algoritmo será un scoring (o una probabilidad) que permitirá ordenar a los clientes por su propensión a la contratación de una tarjeta de crédito.

El algoritmo predictivo se desarrollará en Python 3.6.

3. Fuentes de datos

Se han identificado las siguientes fuentes de datos:

Fuente	Necesidad de histórico
Vista de clientes-productos	17 fotos

4. Público objetivo

Se entiende por público objetivo la población sobre la que se va a ejecutar este modelo. Es importante que esta población este definida previamente porque es la misma con la que se debe entrenar el algoritmo.

En este caso, el público objetivo lo forman los clientes que no hayan contratado anteriormente una tarjeta de crédito.

A nivel técnico, el público objetivo lo formarán los clientes activos que no tengan contratada una tarjeta de créditos según la definición del siguiente apartado.

Sobre este público objetivo se aplican los filtros definidos en la siguiente sección.

5. Filtros aplicados

Los filtros utilizados en la selección del público son:

- Clientes que no hayan contratado una tarjeta de crédito

- Clientes activos
- Cliente que no hayan fallecido
- Se excluirán los clientes nuevos

6. Definición del target

El target o variable objetivo en los modelos de este tipo se construye como un indicador binario en el que se marca como 1 o positivo aquellos clientes que cumplen el evento a predecir y como 0 o negativo los que no lo cumplen.

Se consideran casos positivos en la fecha de referencia, aquellos clientes que, sin haber contratado una tarjeta de crédito en el mes de ejecución, son propensos a contratar una tarjeta dos meses después de la ejecución, sin haberla contratado en el mes anterior.

Se consideran casos negativos en la fecha de referencia, aquellos clientes que no son propensos a contratar una tarjeta en cualquiera de los meses del periodo de entrenamiento.

Aunque el objetivo del modelo es predecir qué clientes contratarán una tarjeta de crédito dos meses después a la ejecución, desde un punto de vista de entrenamiento del modelo es necesario incluir un mes intermedio adicional ya que los datos se generar a cierre de mes.

7. Definición del horizonte de predicción

Se entiende por horizonte de predicción (meses ciegos) la diferencia en tiempo entre la última actualización de datos y la fecha para la que el modelo está entrenado para predecir. Es muy importante ser cuidadoso en la definición de este horizonte porque es el que nos permite no llegar tarde con las acciones que se pueda derivar de los resultados de este análisis.

Se plantea utilizar como horizonte de predicción 1 mes. Esto significa que con los datos de cierre Julio 2019 se predecirán los clientes que realizar contrataciones de tarjetas de crédito en el mes de Septiembre.

En el escenario ideal de implantación totalmente automática, el scoring estará calculado un día después de la última de las cargas en los sistemas de Santander de los datos necesarios para la ejecución del modelo.

8. Tablones de datos

Se plantea crear varios tablones de datos con la información existente; cada tablón deberá contener:

- Identificación del cliente.
- Fecha de referencia.
- Variables predictoras.
- Target o variable objetivo.

Estos tablones de datos contendrán una única fila por cliente. Las fechas de referencia asociadas a cada cliente se calcularán de la siguiente forma:

- Para los casos positivos, es decir los clientes que contratarán una tarjeta de crédito, se tomará como fecha de referencia dos meses antes al mes en el que ocurre la contratación.

  En el caso de clientes que en algún periodo dieron de alta una tarjeta en varios periodos. Para el caso positivo se tendrá en cuenta la fecha aleatoria de la fecha de referencia asociada a las contrataciones.
- Para los casos negativos, es decir, clientes que no contratarán una tarjeta, se tomará como fecha de referencia cualquiera de los posibles meses del periodo de entrenamiento, eligiéndose esta fecha de manera aleatoria entre las posibles.

9. Tablón de entrenamiento

El tablón de entrenamiento se construirá utilizando datos desde Agosto de 2018 hasta Abril 2019.

Las fechas de referencia podrán variar desde Junio 2018 hasta Febrero 2019 y el resto de mes se utilizan para construir variables promedios y acumuladas de los meses anteriores.

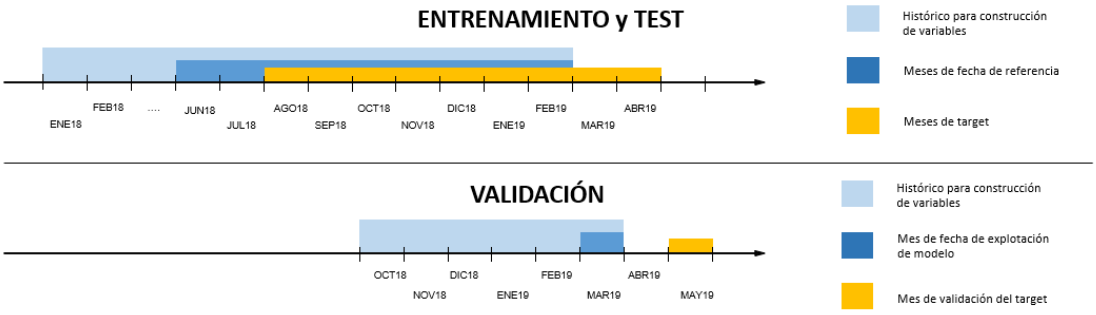
Como las fechas de referencia se sitúan entre Junio 2018 hasta Febrero 2019 y el horizonte de predicción es 1 mes, los target se construirán usando las posibles contrataciones de tarjeta desde Agosto 2018 hasta Abril 2019.

10.Conjunto de test

Se excluirá del tablón de entrenamiento el 20% de los clientes y estos formarán el conjunto de test. Estos clientes no entrarán en el algoritmo de aprendizaje y servirán para medir el modelo sobre población distinta a la usada en entrenamiento

11.Conjunto de validación

Se destina un mes completo, el de Marzo 2019 cuyas contratación de tarjetas a predecir son las de Mayo 2019 para hacer una prueba de validación del modelo real. Con un mes completo de datos y con datos que nunca han entrado en el algoritmo de aprendizaje.



Anexo 3: Documento de Depuración

1. Objetivo del documento

La mayoría de los algoritmos predictivos obtienen las estimaciones de probabilidad en base a ajustes de las medias de las variables numéricas. Esto hace que sean muy sensibles a los outliers (valores extremos o atípicos), ya que un único valor muy lejano al rango de valores habituales puede distorsionar las medias y generar estimaciones que no tendrán tanto acierto para la población en global.

Por ello, es necesario realizar labores de depuración en los datos. Estas labores consisten en eliminar o sustituir ciertos registros del conjunto de entrenamiento, incluso sabiendo que son datos reales y correctos, pero incluirlos como son originalmente tendría efectos negativos en el acierto del modelo.

El objetivo del presente documento es detallar las depuraciones que se van a tener en cuenta para la construcción del conjunto de entrenamiento.

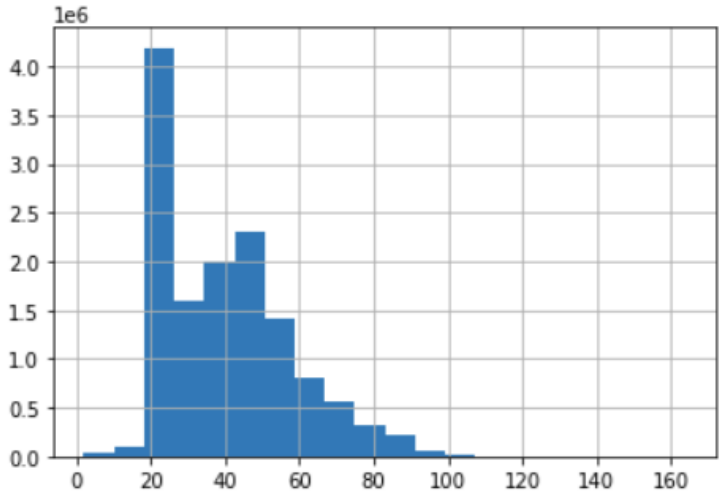
2. Técnicas de depuración

En la situación actual, en la que el número de observaciones (clientes) que tenemos en nuestro conjunto tiene un tamaño muy elevado, podemos sin perder generalidad y representatividad, eliminar o modificar los registros considerados outliers.
A modo de ejemplo, se incluyen mediante histogramas el proceso de depuración para la variable Edad.

2.1. Valores originales leídos directamente del fichero recibido

En el análisis de los datos se detectan valores muy extremos, edad mínima de 2 años y máxima de 164.

Histograma edad



Min.	1st Quartile	Mediana	Media	3rd Quartile	Max.	Valores NA's (nulos)
2	24	39	40	51	164	0

Para esta variable, la depuración es simple porque todos sabemos que estos valores son incorrectos y deben ser eliminados o sustituidos. Pero en otros casos, como importes de saldos o cuotas, en ocasiones se requiere conocimiento de negocio para poder proponer la solución.

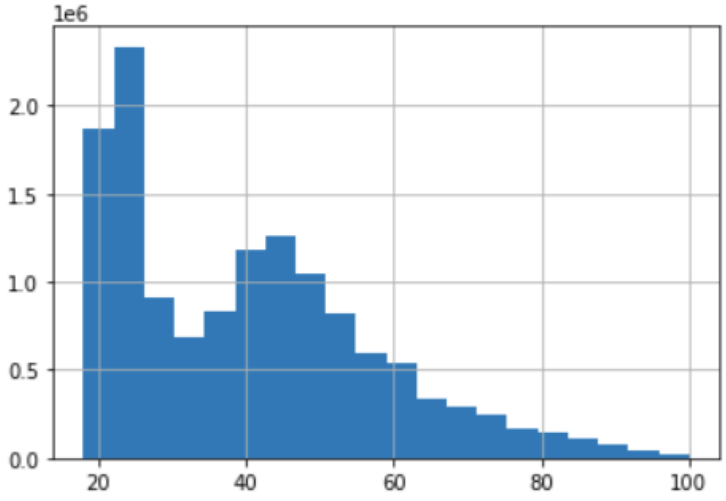
2.2. Valores erróneos excluidos

Una vez detectada la presencia de algunos valores que necesitan depuración hay que establecer un criterio para definir cuáles exactamente son los que se van a depurar. Hay dos opciones:

a. Si se tiene el conocimiento de los límites que debe tener la distribución de la variable correcta se aplican estos límites y los valores que estén fuera se eliminan o sustituyen. Como el objetivo del proyecto es realizar un modelo predictivo
b. Si no se conocen los límites de la distribución, se puede utilizar cualquiera de las técnicas estadísticas para la limpieza de outliers. Como el objetivo del proyecto es realizar un modelo predictivo.

En el ejemplo que estamos tratando de la variable edad, según criterios de negocio se tendrá en cuenta el rango de edades válidas entre 18 y 100 años. Aplicando esta depuración obtenemos los siguientes resultados.

Histograma edad



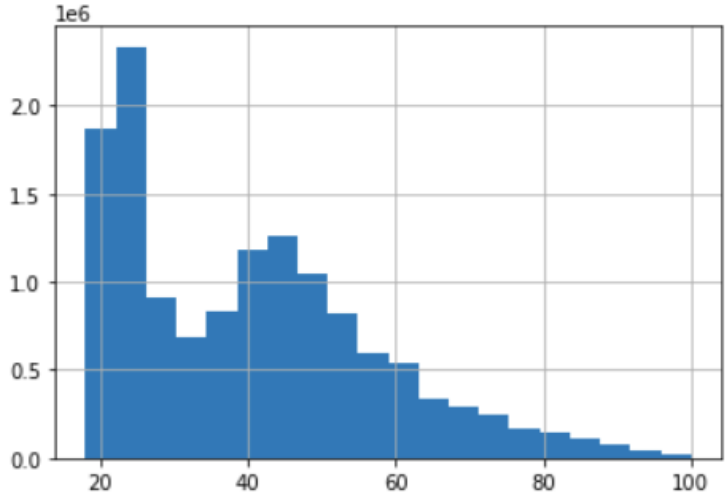
Min.	1st Quartile	Mediana	Media	3rd Quartile	Max.	Valores NA's (nulos)
18	24	39	40	51	100	3050

2.3. Valores erróneos sustituidos.

Para sustituir un valor necesitamos comprender la variable y buscar el valor más adecuado. Las opciones habituales son:

- Utilizar la media
- Utilizar un valor mínimo teórico
- Utilizar un valor máximo teórico
- Opciones mixtas según criterio para cada grupo de registros.

En el caso del ejemplo vamos a utilizar el máximo y mínimo teórico. Es decir, los 3050 valores erróneos que en el paso anterior se excluyeron sustituyéndose por nulos, le asignaremos el valor máximo 100 y mínimo 18.



Min.	1st Quartile	Mediana	Media	3rd Quartile	Max.	Valores NA's (nulos)
18	24	39	40	50	100	0

La propuesta de depuración final para el ejemplo de la variable edad es sustituir los valores extremos (edades menos a 18 y mayores a 100 años) por el valor de máximo y mínimo teórico.

3. Propuesta de depuración

3.1. FUENTES

Vista cliente-productos

Se ha recibido información de clientes-productos desde los meses de Enero de 2018 hasta Mayo de 2019 a modo de cierres mensuales. El volumen recibido es 13.619.575 registros, que corresponden a 956.645 clientes distintos

Los volúmenes recibidos por fecha de dato son:

fec_dato	# filas
2018-01-27	618.504
2018-02-27	621.454
2018-03-27	624.118
2018-04-27	626.075
2018-05-27	628.360
2018-06-27	630.249
2018-07-27	829.817
2018-08-27	843.201
2018-09-27	865.440
2018-10-27	892.251
2018-11-27	906.109
2018-12-27	912.021
2019-01-27	916.269
2019-02-27	920.904
2019-03-28	925.076
2019-04-28	928.274
2019-05-28	931.453

La vista estará incluirá datos del clientes y cartera de productos contratados por el cliente. A continuación, detallamos las variables recibidas:

Datos Clientes

La información recibida en el fichero es:

fec_dato	Fecha de extracción del dato
cod_cliente	Código cliente
xti_empleado	Indicador empleado
nom_pais_resid	Nombre país de residencia
xti_genero	Genero del cliente
num_edad	Edad del cliente

fec_alta_cli	Fecha de alta del cliente
xti_nuevo	Indicador nuevo cliente (antigüedad menor a 6 meses)
num_antiguedad	Antigüedad del cliente en meses
xti_relacion	No utilizado
fec_baja_cli	No utilizado
xti_tipo_cliente	Indicador de tipo de cliente
xti_estado	Estado de cliente
xti_residencia	No utilizado
xti_extranjero	Indicador país de nacimiento del cliente en el extranjero
xti_conyugue	No utilizado
nom_canal	Canal usado por el cliente
xti_fallecimiento	Indicador de fallecimiento
xti_domicilio	No utilizado
cod_provincia	Código de provincia de domicilio
nom_provincia	No utilizado
xti_activo	Indicador activo cliente
imp_renta	Importe de renta del cliente
cod_segmento	Código de segmento

Se han realizado depuraciones de las siguientes variables:

Nombre de canal: Según el análisis realizado, el 1% de los registros tiene valores nulos para esta variable. Al ser una variable categórica, se asignará una nueva clase “NA” a los valores nulos.

Código de provincia: Según el análisis realizado, el 0.5% de los registros tiene valores nulos para esta variable. Al ser una variable categórica, se asignará una nueva clase “0” a los valores nulos.

Código de segmento: Según el análisis realizado, el 1.2% de los registros tiene valores nulos para esta variable. Al ser una variable categórica, se asignará una nueva clase “NA” a los valores nulos.

Importe de renta: Según el análisis realizado, el 20% de los registros tiene valores nulos para esta variable. En este caso los valores nulos se han sustituido siguiendo el siguiente criterio.
Mediana de renta por provincia. Este criterio se aplicará para los clientes con provincia informada, representa el 42.19% de los registros. La renta se calculará teniendo en cuenta la mediana renta de la provincia en la que se encuentra el domicilio del cliente.
Mediana de renta global. Este criterio se aplicará para los clientes que no tienen informada la provincia, representa el 2.37% de los registros. La renta se calculará teniendo en cuenta la mediana renta del total de clientes.

Para el resto de variables no se proponen depuraciones ya que se ha comprobado que los valores existentes son los valores que aparecen en los catálogos de posibles valores para cada variable.

Cartera de productos contratados

La información recibida en cada fichero es:

xti_ahor_fin_ult1	Indicador de cuenta de ahorro
xti_aval_fin_ult1	Indicador de cuenta de garantías
xti_cco_fin_ult1	Indicador de cuenta corriente
xti_cder_fin_ult1	Indicador de cuenta derivada
xti_cno_fin_ult1	Indicador de cuenta nómina
xti_ctju_fin_ult1	Indicador de cuenta junior
xti_ctma_fin_ult1	Indicador de cuenta más particular
xti_ctop_fin_ult1	Indicador de cuenta particular
xti_ctpp_fin_ult1	Indicador de cuenta de particular plus
xti_deco_fin_ult1	Indicador de depósitos a corto plazo
xti_deme_fin_ult1	Indicador de depósitos a medio plazo
xti_dela_fin_ult1	Indicador de depósitos a largo plazo
xti_ecue_fin_ult1	Indicador de cuenta online
xti_fond_fin_ult1	Indicador de fondos
xti_hip_fin_ult1	Indicador de hipoteca
xti_plan_fin_ult1	Indicador de pensiones
xti_pres_fin_ult1	Indicador de préstamos
xti_reca_fin_ult1	Indicador de impuestos
xti_tjcr_fin_ult1	Indicador de tarjeta de crédito

xti_valo_fin_ult1	Indicador de valores
xti_viv_fin_ult1	Indicador de cuenta de vivienda
xti_nomina_ult1	Indicador de nómina
xti_nom_pens_ult1	Indicador de pensiones
xti_recibo_ult1	Indicador de débito directo

Variables predictoras

Una variable predictora (o sintética) es una variable que se construye a partir de variables originales, de esta forma se resume la información original.

Una vez auditada y depurada la información de los datos maestros y la información disponible, se construyen las variables predictoras.

Con estas variables se construirá el tablón de estudio que será utilizado en todas las fases de la creación del modelo: entrenamiento, validación y ejecución. Por este motivo, es necesario disponer para cada una de las fases de una fotografía temporal que refleje la situación de los individuos previa a la contratación de una tarjeta de crédito.

Se adjunta el listado de variables predictoras que se han desarrollado para el modelo:

Fuente datos	Variable	Descripción	Límite Inferior	Límite Superior
Vista Cliente-Productos	num_prods_total	Número de productos contratados en el mes actual.	0	24
Vista Cliente-Productos	inc_total_productos_0_3m	Incremento de productos contratados en los últimos 3 meses.	0	24
Vista Cliente-Productos	inc_total_productos_0_6m	Incremento de productos contratados en los últimos 6 meses	0	24
Vista Cliente-Productos	flag_cancelacion_total_0_3m	Flag que indica si se ha cancelado un producto en los últimos 3 meses.	0	1
Vista Cliente-Productos	flag_cancelacion_total_0_6m	Flag que indica si se ha cancelado un producto en los últimos 6 meses.	0	1
Vista Cliente-Productos	flag_contratacion_total_0_3m	Flag que indica si se ha contratado un producto en los últimos 3 meses.	0	1
Vista Cliente-	flag_contratacion_total_0_6m	Flag que indica si se ha	0	1

Productos		contratado un producto en los últimos 6 meses.		