UNIVERSIDAD RICARDO PALMA
ESCUELA DE POSGRADO

MAESTRÍA EN CIENCIA DE LOS DATOS

Tesis para optar el Grado Académico de Maestro en Ciencias de los Datos

Determinación de la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente usando algoritmos de aprendizaje automático.

Autor: Bach. Rivera Bardales John Frank

Asesor: Mg. Roque Paredes Ofelia

LIMA - PERÚ

2020

PÁGINA DEL JURADO

El Jurado Examinador para la evaluación de la sustentación de la presente tesis, se encuentra integrado por los siguientes miembros:

1. Presidente :	PhD. Oscar Efraín Ramos Ponce
2. Miembro :	Mg. José Antonio Cárdenas Garro
3. Miembro :	Mg. Walter Edwin Marticorena Ramos
4. Asesor Interno :	Mg. Ofelia Roque Paredes
5. Asesor Externo:	Dr. Edwyn Javier Aldana Bobadilla

DEDICATORIA

A mis padres y hermanos,

a mi esposa,

a mis hijas,

y principalmente a Dios,

gracias por apoyarme siempre.

AGRADECIMIENTO

Un agradecimiento especial a mis padres por los valores formados en mí y a mi familia por su incondicional apoyo en cada etapa de mi vida. Asimismo agradezco a la universidad por haber confiado en este programa, a todas las personas que trabajaron y sacaron adelante el programa en especial al fundador de la maestría Dr. Erwin Kraenau quien fue un visionario, líder y guía, a la coordinadora de la maestría Mg. Ofelia Roque quien siempre nos apoyó y estuvo pendiente de nuestro avance además de saber llevar este programa de la mejor manera posible. Gracias a mi asesor externo Dr. Edwyn Aldana que a pesar de la distancia siempre tuvo un tiempo para guiarme en este proceso.

ÍNDICE

PÁGINA DEL JURADO i
DEDICATORIA ii
AGRADECIMIENTO iii
LISTADO DE TABLAS vi
LISTADO DE FIGURAS vii
RESUMEN viii
ABSTRACT ix
INTRODUCCIÓN 1
CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA 3
1.1 Descripción del Problema 3
1.2 Formulación del problema y justificación del estudio 4
1.2.1 Problema general. 4
1.2.2 Problemas específicos. 5
1.3 Importancia y Justificación del estudio 6
1.4 Delimitación del estudio 6
1.5 Objetivos de la Investigación 7
1.5.1 Objetivo general. 7
1.5.2 Objetivos específicos. 7
CAPÍTULO II: MARCO TEÓRICO 9
2.1 Marco histórico 9
2.2 Investigaciones relacionadas con el tema 11
2.2.1 Antecedentes Internacionales 11
2.3 Estructura teórica y científica que sustenta el estudio 13
2.3.1 Aprendizaje de un programa de computadora 13
2.3.1.1 Aprendizaje automático 14
2.3.1.2 Aprendizaje Supervisado 15
2.3.1.2.1 Árbol de clasificación. 15
2.3.1.2.2 Redes Neuronales. 18
2.3.1.2.3 Modelos ensamble basado en árboles. 21

2.3.1.2.3.1 Bagging. 22
2.3.1.2.3.1.1 Random Forest. 24
2.3.1.2.3.2 Boosting. 27
2.3.1.2.3.2.1 Gradient Boosting Machine (GBM). 29
2.3.1.2.3.2.2 Extreme Gradient Boosting (XGBOOST). 31
2.3.1.3 Aprendizaje No Supervisado. 32
2.3.2 Gestión de la base de datos para la adquisición de nuevos clientes vía telefónica en una campaña vigente de productos financieros. 32
2.4 Definición de términos básicos 33
2.5 Hipótesis 36
2.6 Variables 37
CAPÍTULO III: MARCO METODOLÓGICO 39
3.1 Diseño de investigación 39
3.2 Población y muestra 39
3.3 Técnicas e instrumentos 39
3.4 Recolección de datos 39
CAPÍTULO IV: RESULTADOS Y ANÁLISIS DE RESULTADOS 40
4.1 Resultados 40
4.2 Análisis de resultados 47
4.2.1 Indicadores de desempeño del modelo. 47
4.2.2 Gestión de la base de datos. 48
CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES 50
5.1 Conclusiones 50
5.2 Recomendaciones 50
REFERENCIAS BIBLIOGRÁFICAS 52
6.1 Referencias bibliográficas 52
6.2 Referencias electrónicas consultadas 53
ANEXOS 54

LISTADO DE TABLAS

Tabla 1 Calidad del modelo en el test de acuerdo al intervalo de AUC 34
Tabla 2 Matriz de confusión para clasificación binaria 35
Tabla 3 Métricas para evaluar problemas de clasificación binaria 35
Tabla 4 Diccionario de variables 38
Tabla 5 Descripción resumida de las variables 40
Tabla 6 Partición muestral estratificada en base a la proporción de la target 43
Tabla 7 Importancia de variables en el modelo base 44
Tabla 8 Indicadores del modelo 1 46
Tabla 9 Indicadores del modelo 2 46
Tabla 10 Indicadores del modelo 3 47
Tabla 11 Indicadores de los tres modelos desarrollados 47

LISTADO DE FIGURAS

Figura 1. Valor agregado por venta vs costo de transacción 4
Figura 2. Visión general de como el aprendizaje automático se utiliza para abordar una tarea determinada..... 14
Figura 3. Posibles árboles caso estudiantes que juegan cricket en su tiempo libre. 16
Figura 4. Componentes de una neurona biológica..... 19
Figura 5: (a) Modelo McCullochPitts (M-P) (b) Estructura de red de avances múltiples..... 20
Figura 6. Pasos seguidos en Bagging..... 23
Figura 7. Algoritmo Gradient Bosst..... 30
Figura 8. Algoritmo GBM 31
Figura 9. Gráfico de boxplot para X10, X12 y X12 42
Figura 10. Gráfico de boxplot para X20, X21, X22 y X23 42
Figura 11. Gráfico de boxplot para X20, X21, X22 y X23 cuando X24=1..... 43
Figura 12. Importancia de los predictores en el modelo base..... 45
Figura 13. Importancia de variables modelo 1..... 46

RESUMEN

En este trabajo de tesis se planteó abordar la determinación de la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente, para ello se usaron algunos algoritmos de aprendizaje automático. El estudio tuvo un enfoque de aprendizaje supervisado donde la variable objetivo es la aprobación del producto financiero y de variables específicas de la gestión. Asimismo con la determinación de la aceptación se buscó generar futuras eficiencias para el área de inteligencia de televentas y tomar mejores decisiones en la gestión de llamadas. Además se analizó si favorece el uso del protocolo de venta basado en el perfil de la gestión.

Los datos utilizados pertenecen a un *call center* propio de una entidad financiera, recolectados en los cuatro primeros meses del 2018. Se contó con variables propias de la base de datos de clientes potenciales a la campaña y otras variables propias de la gestión de llamadas. Para la determinación de la aceptación de un producto financiero se desarrollaron tres modelos mediante algoritmos de aprendizaje automático y se seleccionó el mejor modelo basado en el indicador AUC sin descuidar el indicador de sensibilidad pues interesa al negocio.

Palabras Claves: aprendizaje supervisado de clasificación, *telemarketing*, producto financiero, xgboost, gbm, redes neuronales.

ABSTRACT

This thesis proposed the determination of the acceptance of a financial product based on the management of calls to potencial clients in a current campaign so some machine learning algorithms has been used. This research has a supervised learning approach where the target variable is the approval of the financial product and of the specific management drivers. Besides, with the determination of acceptance, the aim was to generate future efficiencies for the telesales intelligence area and define better decisions in call management. In addition, it was analyzed whether it favors the use of the sales protocol based on the management profile.

The data used belongs to a call center of a financial institution, these were collected in the first four months of 2018 and there are own variables of the database of potential customers to the campaign and other own variables of call management. Consequently, three models were developed using machine learning algorithms and the best model based on the AUC indicator was selected without neglecting the sensitivity indicator as it interests the business.

Keywords: supervised classification learning, *telemarketing*, financial product, xgboost, gbm, neural networks.

INTRODUCCIÓN

Por lo general el negocio de adquirir clientes vía telefónica para una entidad financiera que cuenta con un *call center* propio se realiza de la siguiente manera: mensualmente el equipo de CRM de las entidades financieras entrega una base de datos de potenciales clientes que calificarían para adquirir uno o más productos y/o servicios financieros al área de inteligencia de televentas, para que sean contactados por los ejecutivos de ventas y de esa manera crear la oportunidad de venta. El área de inteligencia de televentas debe gestionar esta base de datos de manera eficiente; priorizando ciertas variables, habitualmente las variables que vienen siendo más efectivas para el negocio.

El estudio realizado planteó abordar la determinación de la aceptación de un producto financiero, basado en la gestión de llamadas a clientes potenciales en una campaña vigente, para lo cual se usaron algunos algoritmos de aprendizaje automático. Cabe señalar que existen pocas fuentes teóricas que hablen del uso de modelos de aprendizaje automático para la adquisición de clientes vía telefónica, al menos en el contexto latinoamericano, por lo que fue necesario recurrir a otras fuentes para conseguir las bases teóricas que sustenten esta investigación.

Asimismo con la determinación de la aceptación se buscó generar futuras eficiencias para el área de inteligencia de televentas y tomar mejores decisiones en la gestión de llamadas. Además se analizó si favorece el uso del protocolo de venta basado en el perfil de la gestión.

La presente investigación consta de 5 capítulos:

1. Capítulo I, se explica el problema y los objetivos a conseguir.

2. Capítulo II, contiene el marco histórico, investigaciones relacionadas con el tema, marco teórico, definición de términos básicos y las hipótesis.
3. Capítulo III, muestra el marco metodológico de la investigación.
4. Capítulo IV, detalla los resultados y análisis de los mismos.
5. Capítulo V, se expresan las conclusiones y recomendaciones en base a lo analizado.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción del Problema

En el mercado de la adquisición de nuevos clientes para la colocación de productos financieros, el uso de *analytics* es cada vez más frecuente; esto hace que el competidor que mejor explote estos conocimientos y/o herramientas será quien mayores beneficios obtenga.

Los datos y el *analytics* están cambiando las bases de la competencia. Las empresas líderes utilizan sus capacidades no sólo para mejorar sus operaciones base, sino para crear modelos de negocios totalmente nuevos. Aquellas empresas que aprovechen eficazmente estas capacidades crearán un valor significativo para su negocio y se diferenciarán del resto, mientras las demás tendrán una desventaja para competir. (McKinsey Global Institute, 2016)

Existen diversas maneras de adquirir clientes, por ejemplo, se tienen canales presenciales, telefónicos y medios sociales.

Las empresas exitosas de hoy en día suelen emplear marketing multicanal, utilizando dos o más canales de mercadeo para llegar a los segmentos de clientes en un área de mercado. Cada canal puede dirigirse a un segmento diferente de compradores, o diferentes estados de necesidad para un comprador, permitiendo entregar los productos adecuados en los lugares correctos de la manera correcta al menor costo. (Kotler & Keller, 2016).

La Figura 1, muestra cómo seis diferentes canales de ventas se acumulan en términos de valor agregado por venta y el costo por transacción.

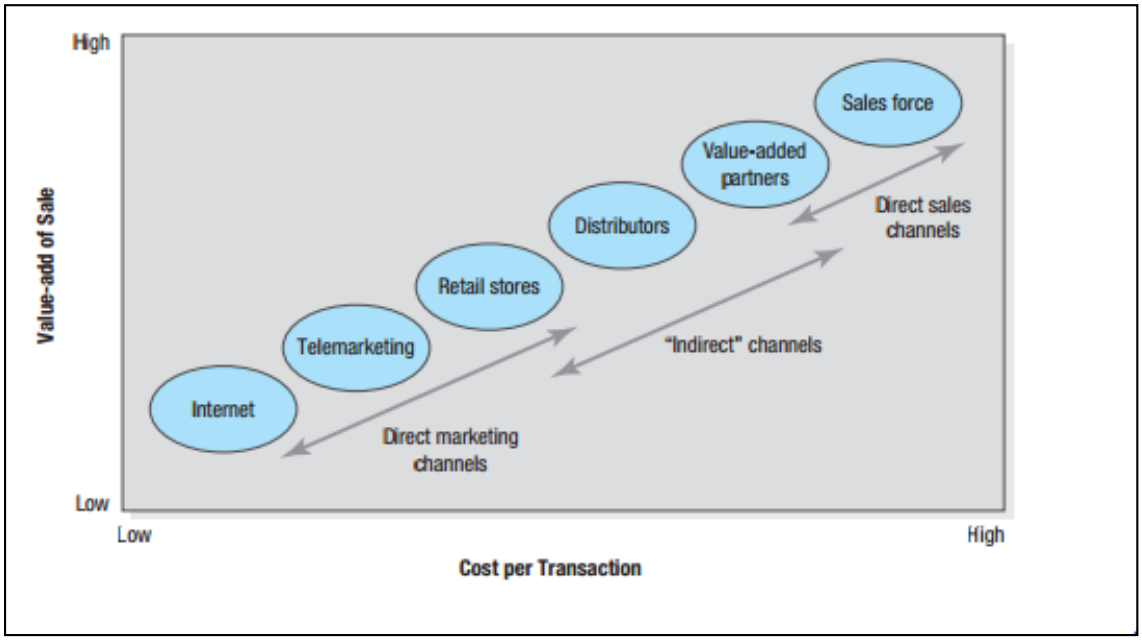


Figura 1. Valor agregado por venta vs costo de transacción
Fuente: (Kotler & Keller, 2016)

En este estudio se trató la adquisición de clientes vía telefónica por parte de un banco que cuenta con un *call center* propio y una base de datos de clientes potenciales donde la problemática fue: El no ser eficiente en la gestión de la base de datos al no tener un modelo predictivo de referencia que determine la aceptación de nuevos clientes para adquirir un producto financiero vía telefónica en una campaña vigente. Asimismo, su impacto se observó en el incumplimiento de metas comerciales iniciales ya que se tuvo que hacer ajustes sobre ellas o también el sobrecumplimiento de metas por parte de los ejecutivos de ventas, por ende la productividad de los ejecutivos de venta no era óptima.

1.2 Formulación del problema y justificación del estudio

1.2.1 Problema general.

En base a lo expuesto anteriormente, se formula la pregunta de esta investigación: ¿Se puede determinar la aceptación de un producto financiero basado en la gestión de llamadas a

clientes potenciales en una campaña vigente de productos financieros usando algoritmos de aprendizaje automático?

1.2.2 Problemas específicos.

Asimismo, tenemos los siguientes problemas específicos:

- ¿Cómo afecta en el negocio de adquisición de clientes vía telefónica el uso del algoritmo *Gradient Boosting Machine* (GBM) en la determinación de la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales ?
- ¿ Cómo afecta en el negocio de adquisición de clientes vía telefónica el uso del algoritmo *Extreme Gradient Boosting* (XGBOOST) en la determinación de la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales?
- ¿Cómo afecta en el negocio de adquisición de clientes vía telefónica el uso del algoritmo de Redes Neuronales Artificiales (RNA) en la determinación de la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales?
- ¿Qué diferencias existen al comparar los algoritmos de aprendizaje automático para la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros en base a los indicadores: AUC y sensibilidad?

1.3 Importancia y Justificación del estudio

El estudio realizado ayudó al área de inteligencia de televentas en priorizar los clientes potenciales en la gestión de la base de datos para cumplir con los objetivos trazados inicialmente de una manera eficaz y eficiente. Para ello se propuso desarrollar un modelo predictivo para la adquisición de nuevos clientes vía telefónica en una campaña vigente de productos financieros; usando una técnica adecuada de aprendizaje automático.

El modelo elegido deberá ser capaz de aprender en base al avance de los días de la campaña vigente y en una fecha determinada (ideal a mitad del mes gestionable) pueda determinar cuánto es el potencial de ventas pendiente pues de esta manera se tendría más claro si se cumplirá con el objetivo mensual, si se necesitará mayores recursos (base de clientes) o se necesita mejorar la calidad de base (mejora de teléfonos para incrementar el ratio de contacto efectivo) ambas son proporcionadas por el área de CRM del banco, si se tiene exceso de personal (determinar la dotación de ejecutivos de ventas y aumentar la productividad ya que se podría colocar otros productos con el exceso de personas y esto aumentaría el *cross-selling*) , si se va a sobre cumplir el objetivo mensual (por ende se tiene que pagar más sueldo variable y esto afectaría al presupuesto), etc.

1.4 Delimitación del estudio

El estudio fue realizado para una entidad bancaria peruana que cuenta con un call center propio y la base de datos recolectada hace referencia a la gestión de llamadas a los clientes potenciales en los cuatro primeros meses del 2018.

1.5 Objetivos de la Investigación

1.5.1 Objetivo general.

- Determinar un modelo predictivo de referencia que determine la aceptación de un producto financiero basado en la gestión de la base de datos para la adquisición de nuevos clientes vía telefónica en una campaña vigente de productos financieros.

1.5.2 Objetivos específicos.

- Aplicar el algoritmo de aprendizaje automático de *Gradient Boosting Machine* (GBM) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.
- Aplicar el algoritmo de aprendizaje automático de *Extreme Gradient Boosting* (XGBOOST) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.
- Aplicar el algoritmo de aprendizaje automático de Redes Neuronales Artificiales (RNA) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.
- Comparar los algoritmos de aprendizaje automático para la aceptación de un producto financiero basado en la gestión de llamadas a clientes

potenciales en una campaña vigente de productos financieros en base a los indicadores: AUC y sensibilidad.

CAPÍTULO II: MARCO TEÓRICO

2.1 Marco histórico

La importancia de los *call center* en la economía de Estados Unidos creció dramáticamente desde 1878, cuando Bell Telephone Company comenzó a usar operadores para conectar llamadas. (Pinedo, Michael; Seshadri, Sridhar; Shanthikumar, J. George;, 2000)

Un *call center* puede servir para diferentes propósitos en una empresa, esto dependerá de la industria en la que se encuentre y de la estrategia general de la empresa. Se puede usar para proporcionar información, manejar pedidos, reservas o realizar transacciones más complejas como proporcionar asesoramiento médico o realizar ciertas transacciones financieras. Se puede apreciar que los propósitos mencionados anteriormente están ligados a la gestión *inbound* (el cliente o el cliente potencial llama al *call center* de la empresa) pero también existe la gestión *outbound* (cuando el ejecutivo de la empresa llama al cliente y/o cliente potencial) y está más relacionada a la venta de un producto y/o servicio que la empresa brinde.

En julio de 1997, Fleet Financial Group comenzó a utilizar el software de *call center* que permite a los operadores ver simultáneamente información sobre clientes y solicitudes de venta con scripts. "De esta forma, un cliente que solicita información sobre una cuenta corriente también puede recibir información sobre la última tasa de un certificado de depósito, una tarjeta de crédito a bajo interés u otro producto que se adapte al estilo de vida y las necesidades de inversión del cliente" (Hamblen, 1997). (Schwartz, 1998) observó que el Fleet Financial Group incrementó en un 30% la cantidad de clientes convertidos en compradores de aquellos que solicitan información. Los ingresos también aumentaron porque el nuevo sistema permitió una venta cruzada de

productos más efectiva (es decir, productos que están relacionados con aquellos sobre los que los clientes informan).

Pero no sólo bastaba con tener de aliado al software predictivo en los *call center*, se debía trabajar en mejorar la orientación de las campañas de *telemarketing*, es decir en encontrar a nuestro público objetivo al cuál enfocar nuestras campañas y ser más eficientes en la gestión. Trabajos relacionados a este enfoque han sido investigados principalmente por Sergio Moro.

Otra tendencia diferente de la investigación es el análisis de la receptividad de los clientes hacia las campañas de *telemarketing*, utilizando dicho conocimiento para mejorar futuras campañas. En julio del 2009, (Mehrotra, Ankit ; Agarwal, Reeti ;, 2009) menciona que no se debería enfocarse mucho en la tecnología sin entender las actitudes y preferencias de los clientes. El objetivo del artículo fue utilizar técnicas inteligentes como selección de variables y técnicas de clasificación y regresión (C&RT en sus siglas en inglés) para clasificar a los clientes según sus actitudes positivas o negativas hacia el *telemarketing*.

En octubre de 2011, (Moro, Sérgio; Laureano, Raul; Cortez, Paulo;, 2011) implementaron un proyecto de minería de datos; donde recopilaron datos de una campaña de marketing portuguesa relacionada con la suscripción a un depósito bancario. El objetivo comercial fue encontrar un modelo que pueda explicar el éxito de un contacto, es decir, si el cliente suscribe el depósito. Dicho modelo pudo aumentar la eficacia de la campaña identificando las principales características que afectan el éxito, ayudando a una mejor administración de los recursos disponibles (esfuerzo humano, llamadas telefónicas, tiempo) y la selección de un grupo de compradores potenciales de alta calidad y asequible.

En enero del 2015, (Moro, Sérgio; Cortez, Paulo; Rita, Paulo;, 2015) presentaron el concepto de valor de vida útil del cliente (LTV) aplicado al *telemarketing* específicamente a campañas bancarias para vender depósitos a largo plazo. El objetivo era beneficiarse de la historia de contactos pasados para extraer conocimiento adicional. Para su estudio utilizaron redes neuronales. Como resultados se extrajo conocimiento explicativo del modelo propuesto, revelando dos características de LTV altamente relevantes que son: el último resultado de la campaña anterior para vender el mismo producto y la frecuencia de los éxitos pasados de los clientes.

En 2017, el mercado peruano de *call center*s movía US$ 520 millones al año y contaba con una tasa de crecimiento promedio de 7% interanual. Se espera que, para los próximos cinco años, se coloque en un 8% interanual. (Diario Gestión, 2017)

2.2 Investigaciones relacionadas con el tema

2.2.1 Antecedentes Internacionales.

(Moro, Sérgio; Cortez, Paulo; Rita, Paulo;, 2014) Propusieron un enfoque de minería de datos (DM) para predecir el éxito de las llamadas de *telemarketing* para vender depósitos bancarios a largo plazo. Se abordó un banco minorista portugués, con datos recopilados de 2008 a 2013, los datos incluían los efectos de la crisis financiera del 2008. Analizaron un conjunto de 150 variables relacionadas con los atributos del cliente del banco, producto y atributos socio-económicos. Se exploró una selección de variables semiautomáticas en la fase del modelado, realizado con los datos anteriores a julio de 2012, esto permitió seleccionar un conjunto reducido de 22 variables. También compararon cuatro modelos de DM: regresión logística, árboles de decisión (DT), red neuronal (NN) y máquina de vectores de soporte (SVM). Para evaluar el desempeño de los modelos usaron dos métricas: AUC *(Area under curve)* y ALIFT (*Area of the*

LIFT cumulative curve), los cuatro modelos fueron probados en un conjunto de evaluación, usando los datos más recientes (después de julio de 2012) y un esquema de ventana móvil. La NN presentó los mejores resultados (AUC = 0.8 y ALIFT = 0.7), lo que permitió llegar al 79% de los suscriptores al seleccionar los clientes clasificados medio mejor. Además, se aplicaron dos métodos de extracción de conocimiento, un análisis de sensibilidad y un DT al modelo NN y revelaron varios atributos clave (por ejemplo, la tasa de Euribor, la dirección de la llamada y la experiencia del agente bancario). Tal extracción de conocimiento confirmó que el modelo obtenido era creíble y valioso para los administradores de campañas de *telemarketing*.

(Moro, Sérgio; Cortez, Paulo; Rita, Paulo;, 2015) aplicaron el concepto de LTV (*custormer lifetime value*) mediante la incorporación de información histórica para mejorar las capacidades de predicción de un sistema de soporte de decisión de referencia ya sólido que utiliza redes neuronales para vender depósitos bancarios en un contexto de campaña de *telemarketing*. El valor de vida útil del cliente (LTV) permite utilizar las características del cliente, como la actualidad, la frecuencia y el valor monetario (RFM por sus siglas en inglés), para describir el valor de un cliente a lo largo del tiempo en términos de rentabilidad. Se realizó una técnica de selección directa, en la que se probaron doce características de entrada de candidatos de LTV. El procedimiento de evaluación, utilizando un esquema de ventana móvil robusto y realista, y dos métricas, favoreció un modelo basado en datos que incluía cinco características de LTV. Cuando se compara con el modelo de referencia (sin características de LTV), el modelo de LTV mejorado produjo una mejora de 6 pp (puntos porcentuales) en el área de la curva ROC (*Receiver Operating Characteristic*), con un AUC total = 0,86 y 4 pp en la curva de elevación acumulada, con un ALIFT total = 0.70, para clientes con historial de telemercadeo anterior. Además se extrajo conocimiento explicativo del modelo propuesto, revelando dos características de LTV altamente relevantes que son: el último resultado de la campaña anterior para vender el mismo producto y la frecuencia de los éxitos pasados de los clientes. Los resultados obtenidos son particularmente valiosos para las empresas de *call center*, ya que pueden mejorar el rendimiento predictivo sin siquiera tener que solicitar más información a las empresas a las que prestan servicios.

2.3 Estructura teórica y científica que sustenta el estudio

2.3.1 Aprendizaje de un programa de computadora

El objetivo principal del aprendizaje de un programa de computadora es mejorar en alguna tarea con experiencia. Para realizar esto se requiere de la definición de tres componentes (Rokach & Maimon, 2014):

1. La tarea T que nos gustaría mejorar con el aprendizaje.
2. La experiencia E que se usará para el aprendizaje.
3. Una medida de desempeño P que se utilizará para medir la mejora.

Aunque cabe señalar que fue Tom Mitchell quién definió estos tres componentes (Mitchell, 1997). Brindo un ejemplo para entender mejor los componentes mencionados. Se tiene el problema de spam en los correos electrónicos, se considera spam a todo correo que el usuario no quiere recibir y no pidió recibir. En este caso se pueden usar técnicas de aprendizaje automático para filtrar automáticamente estos tipos de correos. Para aplicar el aprendizaje automático en este caso se requiere los tres componentes ya mencionados, como sigue:

1. La tarea T es para identificar los emails no deseados (spam).
2. La experiencia E es un conjunto de correos que fueron clasificados por los usuarios como correos no deseados (spam) y deseados (no spam).
3. La medida de desempeño P es el porcentaje de correos spam que fueron clasificados como spam y el porcentaje de correos deseados (no spam) que fueron clasificados como spam.

2.3.1.1 Aprendizaje automático

El aprendizaje automático se define como un proceso automatizado que extrae patrones de los datos (Kelleher, John D.; Namee, Brian Mac; Aoife, D'Arcy;, 2015).

Uno de los objetivos del aprendizaje automático es construir un sistema inteligente. Los dos componentes principales que pueden ayudar a que los enfoques del aprendizaje automático alcancen este objetivo son los modelos de aprendizaje y los algoritmos de aprendizaje (Suthaharan, 2016).

El aprendizaje automático es todo sobre el uso de las variables correctas para construir los modelos correctos que logran realizar las tareas correctas (Flach, 2012).

La figura 2, muestra una visión general de como el aprendizaje automático se utiliza para abordar una tarea determinada. Una tarea (cuadro rojo) requiere un mapeo apropiado que está representado por un modelo que es descrito por las características de las salidas. Obtener tal mapeo de los datos de entrenamiento es lo que constituye un problema de aprendizaje.

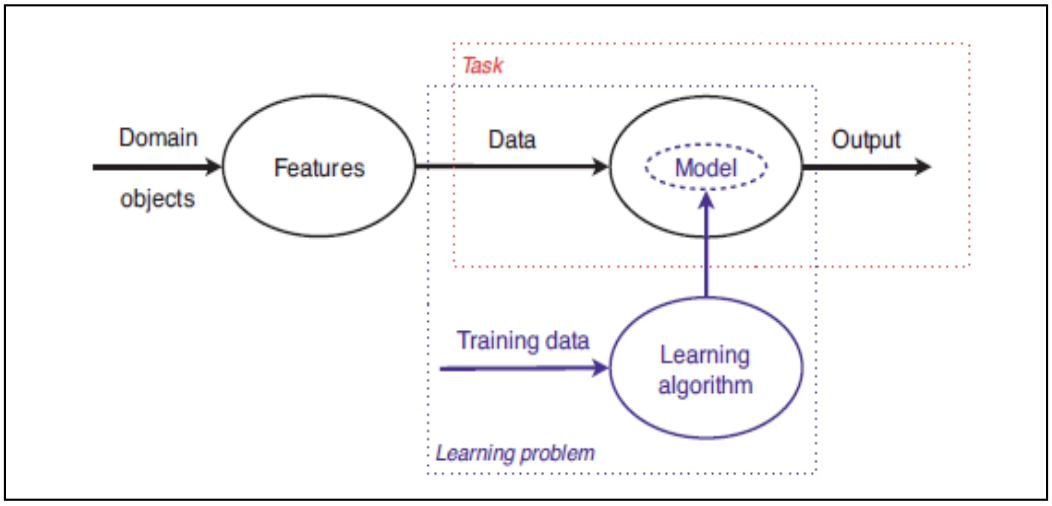


Figura 2. Visión general de como el aprendizaje automático se utiliza para abordar una tarea determinada. Fuente: (Flach, 2012)

2.3.1.2 Aprendizaje Supervisado.

Dentro del aprendizaje automático los métodos de predicción son comúnmente referidos al aprendizaje supervisado. Los métodos supervisados intentan descubrir la relación que existe entre las variables de entrada (conocido como variables independientes) y una variable de salida (variable dependiente o target). Esta relación encontrada será representada en una estructura y se le denominará modelo. Por lo general, los modelos describen y explican fenómenos que están ocultos en los datos. Estos pueden usarse para la predicción de la variable de salida siempre y cuando se conozca las variables de entrada. Los métodos supervisados se pueden implementar en diversas áreas como marketing, finanzas, producción, entre otras (Rokach & Maimon, 2014).

2.3.1.2.1 Árbol de clasificación.

Es un tipo de algoritmo de aprendizaje automático, teniendo una variable objetivo pre-definida. Su principal uso se da en los problemas de aprendizaje supervisado. Los árboles de decisión trabajan con variables categóricas y continuas. En esta técnica se separa la población o una muestra en dos o más conjuntos homogéneos o sub-poblaciones basados en el mejor separador o diferenciador en las variables de entrada.

Ejemplo, Se tiene una muestra de 30 estudiantes con tres variables que juegan cricket en su tiempo libre (Analytics Vidhya, 2016):

Género (femenino y masculino), clase (IX y X) y altura (5 a 6 ft), 15 estudiantes fuera de estos 30 juegan cricket en su tiempo libre. Se desea crear un modelo para predecir ¿quién jugaría cricket durante su tiempo libre? Se necesita segregar los estudiantes que juegan cricket en su tiempo libre basado en la variable de entrada más significativa entre todos los árboles.

En la Figura 3 se observa que el género es capaz de identificar los mejores conjuntos homogéneos en comparación de las otras dos variables.

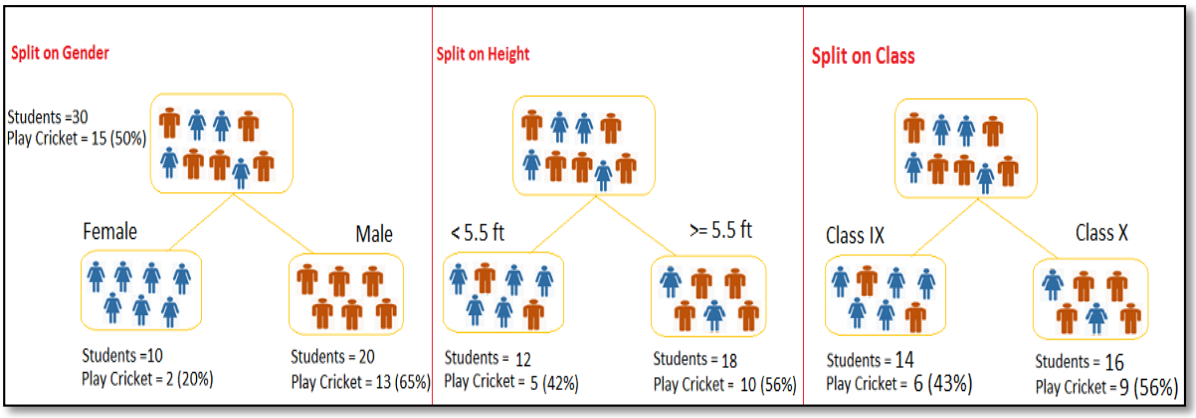


Figura 3. Posibles árboles caso estudiantes que juegan cricket en su tiempo libre.

Fuente: (Analytics Vidhya, 2016)

Para determinar la variable y la división que nos da los mejores conjuntos homogéneos en un árbol se usa diferentes algoritmos como son índice gini, chi-cuadrado, beneficio de la información, reducción de la varianza; donde los tres primeros algoritmos son usados para variables categóricas y el último para variable continuas.

Tipo de Árboles de decisión: Se basan según el tipo de variable objetivo (target).

I. Árboles de decisión de variable categórica (Variable respuesta binaria).

II. Árboles de decisión de variable continúa.

Terminología relacionada a árboles de decisión:

1. Nodo raíz: Representa la población entera o la muestra y además se divide en 2 o más conjuntos homogéneos.
2. *Splitting* (división): Es el proceso de dividir un nodo en 2 o más sub-nodos.
3. Nodo decisor: Cuando un sub-nodo se divide en más sub-nodos, luego esto es llamado nodo decisor.
4. Hoja o nodo terminal: Es un nodo que no divide.
5. Poda: Cuando retiramos sub-nodos de un nodo decisor, a este proceso lo llamamos poda
6. Rama o sub-árbol: Una sección de un árbol entero.
7. Nodo padre y nodo hijo: Un nodo que es dividido en sub-nodos es llamado nodo padre y los nodos productos de la división vendrían a ser los nodos hijos.

Ventajas de usar árboles:

1. Fácil de entender: La salida de un árbol de decisión es muy fácil de entender para personas que no tienen un *background* analítico. No requiere de conocimientos estadísticos para leer e interpretar los resultados. Esta representación gráfica es intuitiva y los usuarios pueden relacionar fácilmente sus hipótesis.
2. Útil en la explotación de datos: Árboles de decisión es uno de los caminos más rápidos para identificar las variables más significativas y las relaciones entre dos o más variables. Además con la ayuda de los árboles de decisión se pueden crear nuevas variables/ características mejor poder para predecir la variable objetivo.

3. Menos datos para limpiar: Se requiere menos datos para limpiar a comparación de otras técnicas, los árboles no se ven afectos con datos atípicos y valores perdidos hasta cierto grado.
4. El tipo de dato no es una restricción: Se puede manipular variables numéricas y categóricas
5. Método no paramétrico: Al ser un método no paramétrico no necesita supuestos acerca el espacio de distribución y la estructura del clasificador.

Desventajas:

1. Sobreajuste: Es una de las dificultades más prácticas. Este problema se resuelve estableciendo restricciones sobre los parámetros del modelo y la poda.

No apto para variables continuas: Mientras se trabaja con variables numéricas continuas se pierde información al categorizar estas.

2.3.1.2.2 Redes Neuronales.

Las redes neuronales, también llamadas redes neuronales artificiales, se originaron al simular redes neuronales biológicas. Fueron usadas ampliamente en los años 80's y 90's. Por algunas razones su popularidad cayó en los 90's pero ahora han vuelto a resurgir. Por lo general son algoritmos altamente costosos computacionalmente y por lo tanto su resurgimiento se debe al avance tecnológico que hoy se tiene y son una de las técnicas más vanguardistas que se tiene por la diversidad de aplicaciones que tiene desde clasificación, predicción, reconocimiento de imagen, etc.

La base de la red neuronal es una neurona, un elemento que imita el trabajo de las neuronas del cerebro. Las dendritas de una neurona reciben la información por conexiones especiales llamadas sinapsis y produce una salida que está conectada a las entradas (dendritas) de otras neuronas. Además, hay una salida (axón) de una señal que llega a las sinapsis de otras neuronas (Kriesel, 2007). En la Figura 4 se puede ver la ilustración de los componentes de una neurona biológica.

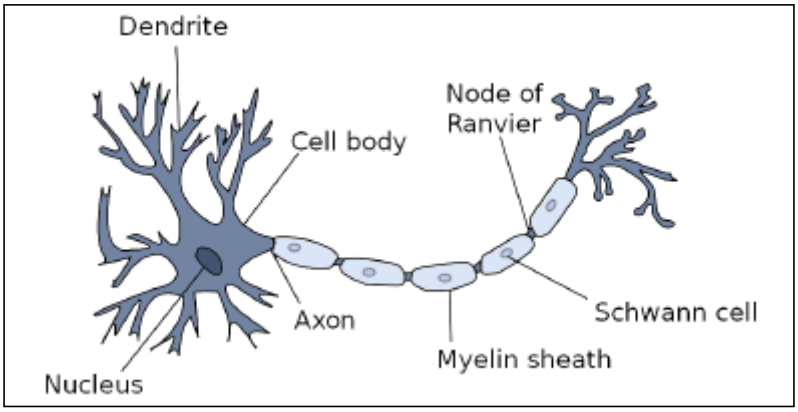


Figura 4. Componentes de una neurona biológica
Fuente: (Kriesel, 2007)

La función de la neurona está determinada por el modelo de la neurona, la estructura de la red y el algoritmo de aprendizaje.

La neurona también se denomina unidad, que es el componente computacional básico en las redes neuronales. El modelo de neurona más popular, es el modelo McCullochPitts (modelo M-P), se ilustra en la Figura 5(a). En este modelo, las señales de entrada (x1, x2, x3) se multiplican con los pesos de conexión (w1, w2, w3) correspondientes al principio, y luego las señales se agregan y se comparan con un umbral (θ), también denominado sesgo de la neurona. Si la señal agregada es más grande que el sesgo, la neurona se activará y la señal de salida se generará

mediante una función de activación, también llamada función de transferencia o función de aplastamiento.

Las neuronas están vinculadas por conexiones ponderadas para formar una red. Hay muchas estructuras de red posibles, entre las cuales la más popular es la red de avance de múltiples capas, como se ilustra en la Figura 5(b).

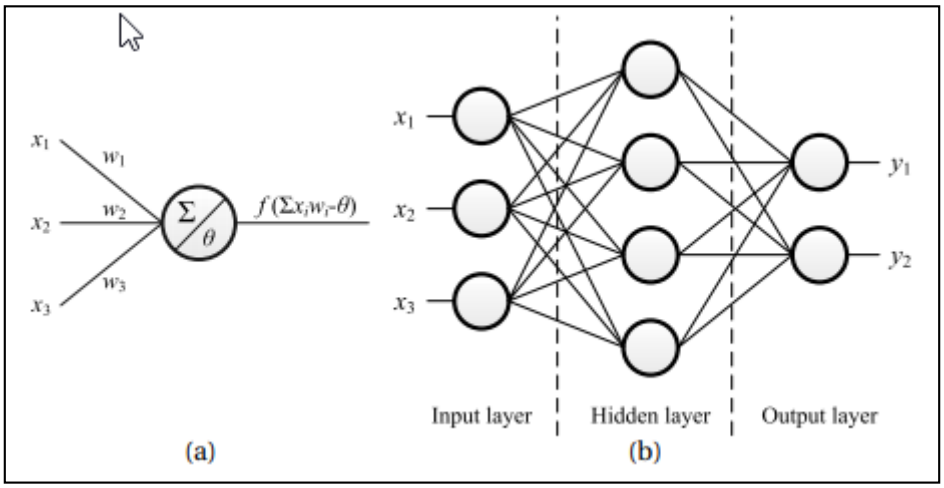


Figura 5: (a) Modelo McCullochPitts (M-P)
(b) Estructura de red de avances múltiples
Fuente: (Zhou, 2012)

Como se puede ver en la estructura de la red de múltiples capas, hay una capa de entrada que recibe vectores de entidades de entrada, donde cada neurona generalmente corresponde a un elemento del vector de características. La función de activación o transferencia para las neuronas de entrada generalmente se establece como f (x) = x. Existe una capa de salida que genera etiquetas, donde cada neurona normalmente corresponde a una etiqueta posible, o un elemento de un vector de etiquetas. Las capas entre las capas de entrada y salida se denominan capas ocultas. Las

neuronas ocultas y las neuronas de salida son unidades funcionales, y una función de activación popular para ellas es la función sigmoide que es definida por:

$$f(x) = \frac{1}{1 + e^{-x}}$$

El objetivo de entrenar una red neuronal es determinar los valores de los pesos de conexión y los sesgos de las neuronas. Una vez que se deciden estos valores, se decide la función calculada por la red neuronal. Hay muchos algoritmos de aprendizaje de redes neuronales. La idea más comúnmente aplicada para entrenar una red neuronal de avance de múltiples capas es que, siempre que la función de activación sea diferenciable, toda la red neuronal puede considerarse como una función diferenciable que puede optimizarse mediante el método de descenso de gradiente. El algoritmo más exitoso es el *Backpropagation*, el cual funciona de la siguiente manera: Al principio, las entradas se envían desde la capa de entrada a través de la capa oculta a la capa de salida, en la que el error se calcula comparando la salida de la red con la verdad fundamental. Luego, el error se propagará nuevamente a la capa oculta y la capa de entrada, durante la cual los pesos y sesgos de conexión se ajustan para reducir el error. El proceso se realiza sintonizando la dirección con el degradado. Dicho proceso se repetirá en muchas veces, hasta que se minimice el error de entrenamiento o se finalice el proceso de entrenamiento para evitar el sobreajuste (Zhou, 2012).

2.3.1.2.3 Modelos ensamble basado en árboles.

La idea del aprendizaje en conjunto (ensamble) es construir un modelo de predicción combinando las fortalezas de una colección de modelos básicos más simples. El aprendizaje en conjunto se puede dividir en dos tareas: desarrollar una población de estudiantes base a partir de

los datos de capacitación y luego combinarlos para formar el predictor compuesto (Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome;, 2009).

No hay una definición clara de aprendizaje en conjunto. En general, se cree que una de las características más importantes de este tipo de aprendizaje es aprender por el mismo problema. El aprendizaje conjunto produce varios clasificadores con diferentes tipos o parámetros, y entrena diferentes muestras por muchas veces. El principio del modelo de aprendizaje conjunto se expresa de la siguiente manera. En primer lugar, se producen varios clasificadores y se obtienen resultados de clasificación entrenados en diferentes muestras. Luego, elije los clasificadores correctos como miembros del conjunto según ciertos criterios. Finalmente, agrega el enfoque conjunto de visas de estos miembros del conjunto y obtiene el resultado del conjunto (Li, Xiao-Lin; Yu, Zhong;, 2012).

En resumen, los métodos de ensamble envuelven un grupo de modelos predictivos que buscan conseguir una mayor precisión y estabilidad al modelo. Algunos de los métodos de ensamble más usados son *Bagging, Boosting y Stacking*.

2.3.1.2.3.1 Bagging.

Bagging (*bootstrap aggregation*), es un procedimiento de propósito general para reducir la varianza de un método de aprendizaje estadístico; es particularmente útil y se usa con frecuencia en el contexto de los árboles de decisión (James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert;, 2014).

(Analytics Vidhya, 2016), menciona que es una técnica utilizada para reducir la varianza de nuestras predicciones combinando el resultado de múltiples clasificadores modelados en diferentes sub-muestras del mismo conjunto de datos.

La Figura 6, brinda una idea de la forma como trabaja *bagging*.

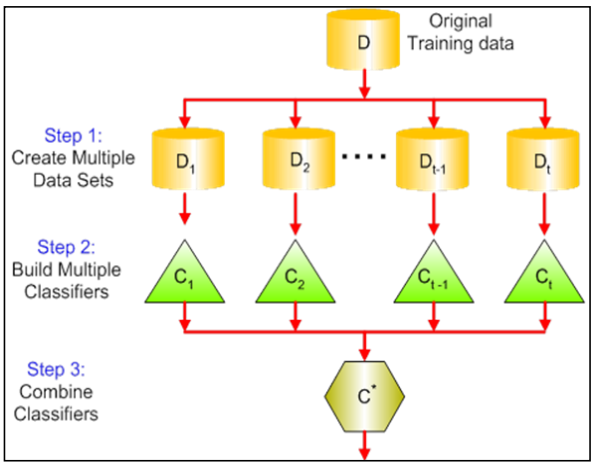


Figura 6. Pasos seguidos en Bagging.

Fuente: (Analytics Vidhya, 2016)

Pasos:

1. Crear múltiples conjuntos de datos de nuestro conjunto original de entrenamiento, para la creación de estos múltiples conjuntos el muestreo se realiza con reemplazo. Los nuevos conjuntos de datos pueden tener una parte de los campos (columnas) así como filas de los datos originales, generalmente son los hiper-parámetros en un modelo *Bagging*. Esto ayuda a la construcción de modelos robustos, menos propensos al sobreajuste.
2. Los clasificadores se construyen en cada conjunto de datos generados de los datos de entrenamiento original, generalmente el mismo clasificador es modelado en cada conjunto de datos y se hacen predicciones.

Los clasificadores se combinan usando la media, mediana, moda dependiendo del problema a resolver. Los valores combinados de los clasificadores por lo general son más robustos. Una de las técnicas más usadas en bagging es *Random Forest*.

2.3.1.2.3.1.1 Random Forest.

Random Forest proporciona una mejora sobre los árboles con *Bagging*, ya que puede tomar árboles que no estén correlacionados. Al igual que en *Bagging*, se construye un número de árboles de decisión en las muestras de entrenamiento (mediante *bootstrap*). Pero al construir estos árboles de decisión, cada vez que se considera una división en un árbol, se elige una muestra aleatoria de m predictores como candidatos divididos del conjunto completo de predictores p. La división sólo permite usar uno de esos m predictores. Luego se toma una nueva muestra de m predictores en cada división, y típicamente elegimos $m \approx \sqrt{p}$, es decir, el número de predictores considerados en cada división es aproximadamente igual a la raíz cuadrada del número total de predictores. En otras palabras, al construir un bosque aleatorio, en cada división del árbol, el algoritmo ni siquiera tiene en cuenta la mayoría de los predictores disponibles. Esto puede parecer una locura, pero tiene un razonamiento ingenioso. Supongamos que hay un predictor muy fuerte en el conjunto de datos, junto con una serie de otros predictores moderadamente fuertes. Luego, en la colección de árboles con *Bagging*, la mayoría o todos los árboles utilizarán este fuerte predictor en la división superior. En consecuencia, todos los árboles con *Bagging* se verán bastante similares entre sí. Por lo tanto, las predicciones de los árboles con *Bagging* estarán altamente correlacionados. Como sabemos, promediar muchas cantidades altamente correlacionadas no conduce a una reducción de la varianza tan grande como promediar muchas cantidades no correlacionadas. Esto significa que *Bagging* no conducirá a una sustancial reducción en la varianza sobre un sólo árbol con esta configuración. *Random Forest* supera este problema forzando a cada división a considerar sólo un

subconjunto de los predictores. Por lo tanto, en promedio (p - m) / p de las divisiones ni siquiera considerarán el fuerte predictor, por lo que otros predictores tendrá más oportunidad. Podemos pensar en este proceso como descorrelacionar los árboles, lo que hace que el promedio de los árboles resultantes sea menos variable y por lo tanto más confiable (James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert;, 2014).

(Analytics Vidhya, 2016), manifiesta que *Random Forest* es considerado una panacea (solución para cualquier problema) de todos los problemas de ciencia de datos. Dado que es un método versátil del aprendizaje automático capaz de realizar tareas de regresión y clasificación. También lleva a cabo métodos de reducción dimensional, valores faltantes, atípicos y otros pasos que se necesita para la exploración de datos. Es un tipo de aprendizaje en conjunto, donde un grupo de modelos débiles se combinan y forman un modelo poderoso.

¿Cómo trabaja *Random Forest*? En *Random Forest*, crecemos varios árboles de manera aleatoria en oposición a un solo árbol como en el modelo CART. Ahora según sea el problema abordado los autores recomiendan lo siguiente:

- Para clasificación, el valor predeterminado para m es $[\sqrt{p}]$ y el tamaño mínimo del nodo es uno.
- Para regresión, el valor predeterminado para m es [p/3] y el tamaño mínimo del nodo es cinco.

Luego para la clasificación, cada árbol da una clasificación y podemos decir el árbol "vota" para esa clase. *Random Forest* escoge la clasificación que tiene más "votos" (sobre todos los

árboles que contiene el bosque) y en el caso de regresión elige el promedio de los resultados de los diferentes árboles (Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome;, 2009).

Ventajas:

1. *Random Forest* puede resolver problemas de clasificación y regresión, además realiza una estimación decente en ambos.
2. Uno de los beneficios de *Random Forest* que más entusiasma es el poder manejar grandes conjuntos de datos con mayor dimensionalidad, pues puede manejar miles de variables de entrada e identificar las variables más significativas por lo que se considera uno de los métodos de reducción de la dimensionalidad. Además, el modelo produce la importancia de la variable de salida, que puede ser una característica muy útil (en algún conjunto de datos aleatorios).
3. Dispone de un método eficaz para estimar los datos que faltan y mantiene la precisión cuando falta una gran parte de los datos.
4. Tiene métodos para balancear errores en conjuntos de datos donde las clases están desbalanceadas.
5. Las capacidades de lo anterior pueden ampliarse a datos no etiquetados, lo que lleva a clústeres sin supervisión, vistas de datos y detección de valores atípicos.
6. *Random Forest* implica el muestreo de los datos de entrada con el reemplazo llamado como muestreo Bootstrap. Aquí un tercio de los datos no se utiliza para el entrenamiento y se puede utilizar para la prueba. Estos se llaman las muestras fuera de bolsa. El error estimado en estas muestras fuera de bolsa se conoce como error fuera de bolsa. El estudio de estimaciones de errores por Fuera de bolsa, da

evidencia para mostrar que la estimación fuera de bolsa es tan precisa como usar un conjunto de prueba del mismo tamaño que el conjunto de entrenamiento. Por lo tanto, el uso de la estimación de error fuera de bolsa elimina la necesidad de un conjunto de prueba de retirada.

Desventajas:

1. Hace un buen trabajo para el caso de clasificación, pero para el caso de regresión no da predicciones exactas dado que las variables continuas fueron categorizadas a rangos.

Es como una caja negra ya que se tiene poco control sobre lo que hace el modelo. En el mejor de los casos puede probar diferentes parámetros y semillas al azar.

2.3.1.2.3.2 Boosting.

Boosting es una de las ideas de aprendizaje más potentes introducidas en los últimos veinte años. Fue diseñado para problemas de clasificación, pero también puede extenderse a la regresión. La motivación fue el desarrollo de un procedimiento que combine los resultados de muchos clasificadores "débiles" para producir uno “poderoso” (Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome;, 2009).

En resumen, Boosting se refiere a convertir aprendices débiles en aprendices fuertes.

Ejemplo, clasificación de spam (Analytics Vidhya, 2016):

¿Cómo se puede clasificar si un email es spam o no? Para esto se puede usar el siguiente criterio, si:

1. El correo electrónico sólo tiene un archivo de imagen (imagen promocional), es un spam.
2. El correo electrónico solo tiene enlaces, es un spam.
3. El cuerpo del email consiste en la oración como "Ud. ganó un dinero del premio de $ xxxxxx", es un spam.
4. Correo electrónico de una fuente conocida, no es spam.

Ahora ¿cree usted que esos criterios son lo necesariamente fuerte para clasificar correctamente un correo electrónico como spam o no? Pues no lo es. A estas reglas antes mencionadas se las conoce como aprendices débiles.

Para convertir un aprendiz débil en fuerte se combinará la predicción de cada aprendiz débil usando:

1. Promedio o promedio ponderado.
2. Tomar en cuenta la predicción con mayor “voto”.

Ahora la pregunta es: ¿Cómo *Boosting* identifica las reglas débiles? Ya que estas reglas débiles serán convertidas en aprendices fuertes. Para la identificación de reglas débiles aplicamos algoritmos de aprendizaje base (ML) con una distribución diferente. Cada vez que se aplica el algoritmo de aprendizaje base genera una nueva regla débil. Esto es un proceso iterativo, después de muchas iteraciones el algoritmo de *Boosting* combina estas reglas en una predicción fuerte.

Ahora vuelve la pregunta ¿Cómo elegir una distribución diferente para cada ronda? Para esto debemos hacer los siguientes pasos:

Paso 1: El aprendiz base toma todas las distribuciones y asigna igual peso o atención a cada observación.

Paso 2: Si hay algún error de predicción causado por el algoritmo de aprendizaje de la primera base, entonces se presta mayor atención a las observaciones que tienen error de predicción. Luego se aplica el siguiente algoritmo de aprendizaje básico.

Paso 3: Itere el Paso 2 hasta que se alcance el límite del algoritmo de aprendizaje base o se logre una mayor precisión.

Paso 4: Por último, combine los resultados del aprendiz débil y cree un aprendiz fuerte que finalmente mejore el poder de predicción del modelo. *Boosting* se centra más en los ejemplos que están mal clasificados o que tienen errores mayores al preceder a las reglas débiles.

Existen diferentes algoritmos basados en *boosting* como por ejemplo ADABOOST, GBM, XGBOOST, LIGHTGBM, etc. En este estudio se hizo uso de GBM y XGBOOST.

2.3.1.2.3.2.1 Gradient Boosting Machine (GBM).

(Friedman, 2001) Realiza una estimación de funciones desde la perspectiva de la optimización numérica en el espacio de las funciones en lugar del espacio de parámetros. Muestra la conexión entre las expansiones aditivas (árboles) y la minimización de una función de costo adecuada según sea el caso (regresión o clasificación) mediante el gradiente de descenso. Se presentan algoritmos específicos para mínimos cuadrados, desviación mínima absoluta y funciones de pérdida de Huber-M para regresión y probabilidad logística multiclase para clasificación. Se obtienen mejoras especiales para el caso particular en el que los componentes aditivos individuales son árboles de regresión, y se presentan herramientas para interpretar tales

modelos "*TreeBoost*". El aumento gradual de los árboles de regresión produce procedimientos competitivos, altamente robustos e interpretables tanto para la regresión como para la clasificación.

1. $F_0(\mathbf{x}) = \arg\min_\rho \sum_{i=1}^{N} L(y_i, \rho)$
2. For $m = 1$ to M do:
3. $\tilde{y}_i = -\left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, \; i = 1, N$
4. $\mathbf{a}_m = \arg\min_{\mathbf{a}, \beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(\mathbf{x}_i; \mathbf{a})]^2$
5. $\rho_m = \arg\min_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i; \mathbf{a}_m))$
6. $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}; \mathbf{a}_m)$
7. endFor
   end Algorithm

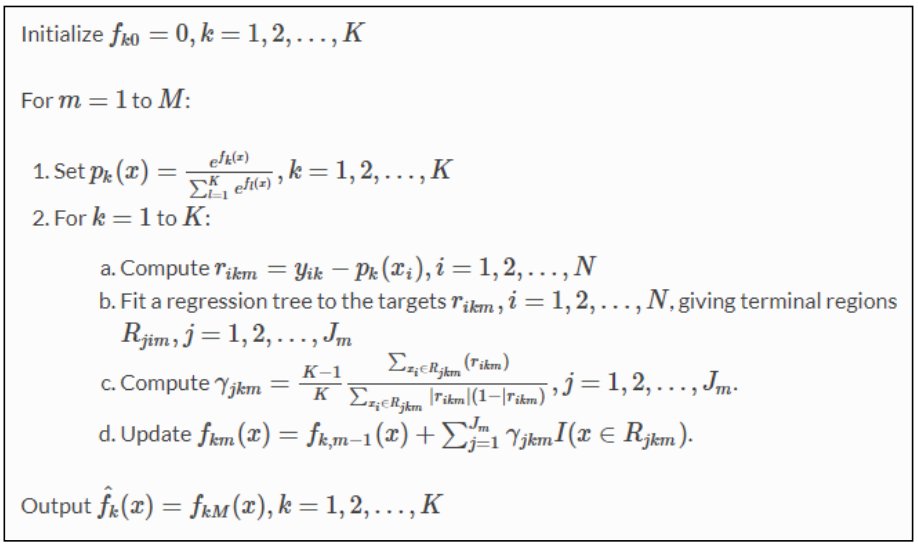
Figura 7. Algoritmo Gradient Boost
Fuente: (Friedman, 2001)

La Figura 7, muestra el primer algoritmo génerico de boosting basado en árboles desarrollado por (Friedman, 2001), donde:

1. Inicializa F(x), en base al argmin de la función de costo L en un punto fijo.
2. Inicia el bloque for.
3. Computa la gradiente negativa.
4. Ajusta el modelo con el fin de encontrar los parámetros a que pertenecen al árbol.
5. Elije un tamaño de paso.
6. Actualiza la estimación de F(x).
7. Fin del bloque for

Tener en cuenta que la elección de la función de costo L depende del tipo de problema a resolver (regresión o clasificación). (Friedman, 2001) Muestra en total seis algoritmos según sea el problema a resolver pero la esencia parte del algoritmo genérico mostrado en la Figura 7.

Como el estudio que se realizó trata de un problema de clasificación, la Figura 8 muestra el algoritmo GBM que usa H2O.ai, el cual se usó para la solución del estudio.



Initialize $f_{k0} = 0, k = 1, 2, \ldots, K$

For $m = 1$ to M:

1. Set $p_k(x) = \frac{e^{f_k(x)}}{\sum_{l=1}^{K} e^{f_l(x)}}, k = 1, 2, \ldots, K$
2. For $k = 1$ to K:
   a. Compute $r_{ikm} = y_{ik} - p_k(x_i), i = 1, 2, \ldots, N$
   b. Fit a regression tree to the targets $r_{ikm}, i = 1, 2, \ldots, N$, giving terminal regions $R_{jim}, j = 1, 2, \ldots, J_m$
   c. Compute $\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{x_i \in R_{jkm}} (r_{ikm})}{\sum_{x_i \in R_{jkm}} |r_{ikm}|(1-|r_{ikm}|)}, j = 1, 2, \ldots, J_m.$
   d. Update $f_{km}(x) = f_{k,m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jkm} I(x \in R_{jkm}).$

Output $\hat{f}_k(x) = f_{kM}(x), k = 1, 2, \ldots, K$

Figura 8. Algoritmo GBM
Fuente: (H2O.ai, 2019)

2.3.1.2.3.2.2 Extreme Gradient Boosting (XGBOOST).

(Chen, Tianqi; Guestrin, Carlos;, 2016) desarrollaron un sistema escalable de árboles boosting llamado XGBOOST, el cual es ampliamente utilizado por los científicos de datos para lograr resultados de vanguardia en muchos desafíos de aprendizaje automático como por ejemplo los que se desarrollan en Kaggle. En dicho estudio proponen un nuevo algoritmo el cuál trabaja muy bien con datos sparse (presentan muchos 0) y pesos ponderados para el aprendizaje aproximados de los árboles. Además incluyen información sobre los patrones de acceso a la memoria cache, la compresión y la fragmentación de datos. Todo esto en busca de construir un sistema escalable de árboles. Por ende XGBOOST escala muy bien en comparación a otros

algoritmos basados en árboles por que hace un uso eficiente de sus recursos y puede apoyarse en la computación paralela y distribuida para agilizar el aprendizaje del modelo.

2.3.1.3 Aprendizaje No Supervisado.

El aprendizaje no supervisado abarca todo tipo de aprendizaje automático donde no hay salida conocida (variable target), se dice también que no existe un maestro para instruir al algoritmo de aprendizaje (Guido, Sarah; Mueller, Andreas;, 2016).

Según Kohavi y Provost et al. (1998), el término "aprendizaje no supervisado" se refiere a "técnicas de aprendizaje que agrupan instancias sin un atributo dependiente pre especificado".

Por lo general los métodos no supervisados se usan en un entorno exploratorio, por ejemplo, se desea segmentar clientes en grupos con similares características. También son usados para describir asociaciones y secuencias (Tsiptsis, Konstantinos; Chorianopoulos, Antonios;, 2010).

2.3.2 Gestión de la base de datos para la adquisición de nuevos clientes vía telefónica en una campaña vigente de productos financieros.

En el entorno de los *call center* que ofrecen productos y/o servicios financieros, lo que se busca es poder tener contacto vía telefónica con el cliente potencial (si esto ocurre se le conoce como contacto efectivo) y colocarle un producto financiero al cuál este califique.

Por lo general estás campañas son mensuales es decir el cliente potencial tiene la oferta en el mes que es contactado vía telefónica.

Se explica brevemente como se realiza el proceso de adquirir clientes vía *call center* de un banco:

Mensualmente el equipo de CRM de las entidades financieras, entrega una base de datos de potenciales clientes que calificarían para adquirir uno o más productos y/o servicios financieros al área de inteligencia de televentas, para que sean contactados por los ejecutivos de ventas y de

esa manera crear la oportunidad de venta. El área de inteligencia de televentas debe gestionar esta base de datos de manera eficiente; priorizando ciertas variables, por lo general las variables que sean más efectiva para el negocio. Para esta gestión los *call center* utilizan marcadores predictivos, los cuales realizan las llamadas y van direccionándolas según la disponibilidad del ejecutivo de ventas. Toda llamada tiene una tipificación, es decir un resultado de llamada; que es asignada por el discador predictivo (cuando la llamada no llega a ser derivada al ejecutivo de venta) o por el ejecutivo de ventas (cuando la llamada es derivada por el discador al ejecutivo de venta). Las tipificaciones que asigna el ejecutivo de ventas son definidas por lo general por el equipo de inteligencia de televentas y son usadas para la gestión de las bases de datos cargadas en el discador predictivo. Una vez que el ejecutivo de ventas contacte con el potencial cliente se crea la oportunidad de venta y dependerá de él el cierre de la venta. Cabe señalar que el ejecutivo de venta debe seguir un protocolo en su interacción con el cliente.

Para medir que tan buena es la gestión de la base de datos para la adquisición de nuevos clientes vía telefónica en una campaña vigente de productos financieros, se usa el ratio de efectividad.

2.4 Definición de términos básicos

- *Alift:* Es una medida de desempeño para evaluar un modelo. Está basado en acumular el lift (ratio de predicción / ratio promedio) obtenido en cada grupo (por lo general se trabajan con deciles y se ordenan de menor a mayor valor de lift). Un alto *ALIF* confirma que el modelo predictivo concentra la mayor cantidad de target en los deciles superiores.
- *Analytics:* Es el descubrimiento, la interpretación y la comunicación de patrones significativos en los datos.
- *Área CRM (Customer relationship management):* Área encargada de la administración de los clientes y/o potenciales clientes del banco.
- *AUC (Area under curve):* Es una medida de desempeño para evaluar el modelo de clasificación elegido. Matemáticamente se expresa $AUC = \int_0^1 ROCdD$, donde ROC es la curva (*Receiver Operating Characteristic*) y D es el umbral.

(Myazina, 2017) Muestra intervalos de AUC asociados a la calidad del modelo desarrollado y para comparar dos o más modelos entre sí, compara el área bajo las curvas ROC o AUC de los *test*.

Tabla 1
Calidad del modelo en el test de acuerdo al intervalo de AUC

Intervalo AUC	Calidad del modelo
[0.9 – 1]	Excelente
[0.8 - 0.9>	Muy bueno
[0.7 - 0.8>	Bueno
[0.6 - 0.7>	Promedio
[0.5 - 0.6>	Insatisfactorio

- *Bootstrap:* Es una método de remuestreo, extremadamente aplicable y potente que se puede usar para cuantificar la incertidumbre asociada con un estimador determinado o un método de aprendizaje estadístico.
- *Call Center:* Es un área donde agentes, asesores, supervisores o ejecutivos, especialmente entrenados, realizan llamadas (llamadas salientes o *outbound*) y/o reciben llamadas (llamadas entrantes o *inbound*) desde o hacia: clientes (externos o internos), socios comerciales, compañías asociadas u otros.
- *Cliente potencial (lead):* Persona que calificaría para adquirir un producto y/o servicio. También se le conoce como lead.
- *Contacto efectivo:* Es cuando existe una interacción vía telefónica entre el cliente potencial y el ejecutivo de venta.
- *Cross-Selling:* Conocido como venta cruzada, en nuestro caso es la venta adicional que realiza un ejecutivo de venta respecto a su producto principal.
- *Discador predictivo:* Software usado para realizar las llamadas de manera automática. Una vez que detecta la voz humana derivada dicha llamada al ejecutivo de venta disponible.
- *Ejecutivo de ventas:* Persona a la cual el discador predictivo transfiere la llamada. Es la persona que interactúa con el cliente potencial buscando colocar un producto y/o servicio.

- *Matriz de confusión:* Es muy usada para medir el desempeño de los modelos en problemas de clasificación, esta tabla o matriz muestra la distribución de los valores observados y de los valores estimados. Los valores observados son los valores reales y los valores estimados se obtienen a partir del modelo de clasificación. (Hossin, Mohammad; Sulaiman, Nasir ;, 2015) nos muestra el uso de una matriz de confusión para evaluar problemas de clasificación binaria y algunas métricas de interés que derivan de ellas, como se muestra en la Tabla 1 y Tabla 2 respectivamente.

Tabla 2
Matriz de confusión para clasificación binaria

	Clase Positiva Predicha	**Clase Negativa Predicha**
Clase Positiva Real	Verdadero Positivo (VP)	Falso Negativo (FN)
Clase Negativa Real	Falso Positivo (FP)	Verdadero Negativo (VN)

Donde VP y VN denota el número de aciertos tanto de la clase positiva y negativa respectivamente. Mientras que FP denota el número de desaciertos en la clase positiva predicha y FN denota el número de desaciertos en la clase negativa predicha.

Tabla 3
Métricas para evaluar problemas de clasificación binaria

Métrica	**Fórmula**	**Enfoque de evaluación**
Exactitud	(VN+VP)/(VN+VP+FN+FP)	Porcentaje de acierto global
Sensibilidad (s)	(VP)/(VP+FN)	Porcentaje de acierto respecto al total real de la clase positiva
Precisión (p)	(VP)/(VP+FP)	Porcentaje de acierto respecto al total predicho de la clase positiva
F1	(2*s*p)/(s+p)	Media armónica entre la sensibilidad y la precisión

- *Modelo:* Es una representación formal de una teoría.
- *Ratio de contacto efectivo:* Cantidad total de llamadas con contacto efectivo / cantidad total de llamadas atendidas por los ejecutivos de ventas.

- *Ratio de aceptación:* Cantidad total que aprobaron en una campaña específica / cantidad total de llamadas con contacto efectivo.
- *Ratio de efectividad:* Cantidad de leads que aprobaron en una campaña específica / cantidad total de leads de la campaña específica.
- *ROC (Receiver Operating Characteristic):* Se traza para estimar la calidad de un modelo de clasificación, y muestra la dependencia del número de resultados positivos correctamente clasificados del número de resultados negativos clasificados incorrectamente.
- *Telemarketing:* Es un proceso interactivo entre una empresa y sus clientes que utiliza un sistema integral de medios y métodos para obtener una respuesta. Uno de los medios más usados es el teléfono.

2.5 Hipótesis

2.5.1 General.

Afecta de manera positiva el uso de un modelo predictivo de referencia que determine la aceptación de un producto financiero en la gestión de la base de datos para la adquisición de nuevos clientes vía telefónica en una campaña vigente de productos financieros.

2.5.2 Específicas.

- H_1 Se podrá aplicar el algoritmo de aprendizaje automático de *Gradient Boosting Machine* (GBM) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.
- H0 No se podrá aplicar el algoritmo de aprendizaje automático de *Gradient Boosting Machine* (GBM) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.

- H_1 Se podrá aplicar el algoritmo de aprendizaje automático de *Extreme Gradient Boosting* (XGBOOST) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.
- H0 No se podrá aplicar el algoritmo de aprendizaje automático de *Extreme Gradient Boosting* (XGBOOST) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.
- H_1 Se podrá aplicar el algoritmo de aprendizaje automático de Redes Neuronales Artificiales (RNA) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.
- H0 No se podrá aplicar el algoritmo de aprendizaje automático de Redes Neuronales Artificiales (RNA) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.
- H_1 Existen diferencias al comparar los algoritmos de aprendizaje automático para la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros en base a los indicadores: AUC y sensibilidad.
- H0 No existen diferencias al comparar los algoritmos de aprendizaje automático para la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros en base a los indicadores: AUC y sensibilidad.

2.6 Variables

A continuación se detalla las variables y su significado.

Tabla 4
Diccionario de variables

Variable	Descripción
X1	Indica el mes de la gestión para el lead tipificado como contacto efectivo.
X2	Identificador del lead tipificado como contacto efectivo.
X3	Indica la prioridad para la gestión del lead, el menor valor es más prioritario.
X4	Indica que tan propenso es el lead para ser contactado, el menor valor es más contactable.
X5	Indica que tan propenso es el lead para aceptar el producto, el menor valor es más propenso.
X6	Indica el flujo de riesgo que tiene el lead si acepta el producto, el menor valor es mejor.
X7	Indica que tan recurrente es el lead en el mes de gestión, el menor valor es más reciente en la gestión.
X8	Indica la zona de riesgo del lead, a menor valor menor zona de riesgo.
X9	Indica el nivel de riesgo del lead, a menor valor menor riesgo.
X10	Indica la edad del lead.
X11	Indica a que categoría de edad pertenece el lead.
X12	Indica el monto para cual califica el lead.
X13	Indica a que categoría de monto califica el lead.
X14	Indica cuántas veces se llamó al cliente (incluye llamadas que no llegaron al ejecutivo).
X15	Indica a que segmento pertenece el cliente.
X16	Indica el tipo de cliente.
X17	Indica a que departamento del Perú pertenece el lead.
X18	Indica a que categoría de departamento pertenece el lead.
X19	Indica si se tiene transcripción de la llamada.
X20	Índice de protocolo de servicio, asociado a la atención que tuvo el lead con los ejecutivos de ventas (sólo tienen a los que se les pudo transcribir la llamada).
X21	Índice de información de beneficios, asociado a la atención que tuvo el lead con los ejecutivos de ventas (sólo tienen a los que se les pudo transcribir la llamada).
X22	Índice de información financiera del producto, asociado a la atención que tuvo el lead con los ejecutivos de ventas (sólo tienen a los que se les pudo transcribir la llamada)
X23	Índice de gestión comercial, asociado a la atención que tuvo el lead con los ejecutivos de ventas (sólo tienen a los que se les pudo transcribir la llamada)
X24	Indica si el lead obtiene el producto, donde 1 es la clase positiva.

Dado la naturaleza del estudio, la variable *target* o dependiente es X24, las demás variables se las considera como variables *drivers* o independientes.

CAPÍTULO III: MARCO METODOLÓGICO

3.1 Diseño de investigación

La investigación fue de tipo explicativa, ya que utiliza los conocimientos previos para aplicarlos ante una nueva situación y el diseño transversal, puesto que el estudio de las variables es en un solo periodo tiempo.

3.2 Población y muestra

El universo y el estudio fueron los 253,759 leads de la entidad financiera que fueron tipificados como contactos efectivos en la gestión para la adquisición de nuevos clientes vía telefónica en una campaña vigente de productos financieros en los 4 primeros meses del 2018. No se trabajó con muestra.

3.3 Técnicas e instrumentos

Los datos se encuentran almacenados en diversas tablas dentro de la base de datos del área de inteligencia de televentas de la entidad financiera. Los datos cuentan con la información de gestión del *call center*, información del cliente potencial y del negocio al cuál el cliente potencial tiene campaña.

3.4 Recolección de datos

Se realizó procedimientos de ETL (extracción, transformación y carga) para la recolección e integración de los datos. Asimismo se compromete a tener un compromiso de confidencialidad y protección de datos de los clientes potenciales.

CAPÍTULO IV: RESULTADOS Y ANÁLISIS DE RESULTADOS

4.1 Resultados

En este apartado, se muestra los resultados obtenidos desde el análisis descriptivo de los datos y la aplicación de los modelos de aprendizaje automático. Todo el análisis y los resultados se obtuvieron usando el software R.

Tabla 5
Descripción resumida de las variables

Nombre	Tipo	Nulos	Media	Mediana	Mad	Mín	Max	Niveles
X1	Factor	-	-	-	-	56506	68494	4
X3	Factor	-	-	-	-	148	53919	16
X4	Factor	-	-	-	-	2000	157660	5
X5	Factor	-	-	-	-	2448	164788	5
X6	Factor	-	-	-	-	1	199712	6
X7	Factor	-	-	-	-	10718	188582	4
X8	Factor	-	-	-	-	19988	85299	4
X9	Factor	-	-	-	-	17787	92767	5
X10	Integer	-	39.71	38.00	11.86	21	65	-
X11	Factor	-	-	-	-	13012	82808	6
X12	Integer	-	8606.74	3400.00	3261.72	700	100000	-
X13	Factor	-	-	-	-	12866	52268	9
X14	Integer	-	7.35	5.00	4.45	1	56	-
X15	Factor	-	-	-	-	1223	127155	6
X16	Factor	-	-	-	-	79129	174630	2
X17	Factor	-	-	-	-	182	133968	25
X18	Factor	-	-	-	-	14337	153016	5
X19	Factor	-	-	-	-	17893	235866	2
X20	Numeric	17893	0.45	0.50	0.19	0	1	-
X21	Numeric	17893	0.13	0.10	0.15	0	0.75	-
X22	Numeric	17893	0.19	0.00	0.00	0	1	-
X23	Numeric	17893	0.17	0.00	0.00	0	1	-
X24	Factor	0				25143	228616	2

Para una descripción breve de los datos se usa la Tabla 2, se obvia la variable X2 pues sólo nos identifica al lead.

Mayor detalle del análisis descriptivo se encuentra en el anexo al usar la función describe de la librería Hmisc, pues esta función nos muestra la cantidad de registros, la cantidad de valores nulos, la cantidad de valores distintos de todas las variables. Si las variables son categóricas se muestra la frecuencia y proporción por cada valor que toma y si son cuantitativas adiciona la media, algunos percentiles, los valores más altos y mínimos, etc. Al observar la variable target se muestra que los datos se encuentran desbalanceados pues sólo el 10% pertenece a la clase de interés 1 (si el lead obtuvo el producto). Sólo se tiene como variables cuantitativas X10, X12, X14, X20, X21, X22 y X23. Cómo la edad está representada por la variable X10 se observa que la edad mínima y máxima es 21 y 65 años respectivamente, tiene una media de 40 años, mediana de 38 años, además el percentil 75 muestra que la edad es 47 años lo cual hace sospechar que tiene una distribución asimétrica positiva. El monto de línea de crédito que calificaría el lead representada por la variable X12 tiene un valor mínimo y máximo de S/700 y S/ 100,000 soles respectivamente esto hace sentido por los diversos segmentos que pertenecen los leads, una media de S/8,607 soles, una mediana de S/3,400 soles lo cual hace pensar en una distribución asimétrica positiva. El número de intentos de llamadas que se le realiza a todos los teléfonos del lead está representada por la variable X14 vemos que tiene un mínimo y máximo de 1 y 56 intentos, una media de 7 intentos y una mediana de 5 intentos lo cual también hace pensar que la distribución es asimétrica positiva. Los valores de las variables X20, X21, X22 y X23 si presentan valores nulos pues sólo se tienen registros de aquellas llamadas que pudieron ser transcritas (92% como lo indica la variable X19), en estas variables si se espera tener una distribución asimétrica (salvo X20 que es un índice de protocolo de servicio) dado que los datos se encuentran desbalanceados y estas variables son propias del protocolo de venta que debe seguir el ejecutivo.

Lo descrito anteriormente para las variables cuantitativas se observa mejor en un gráfico de cajas o *boxplot*, los cuáles son muy útiles para visualizar la distribución de los datos y ver los valores atípicos (*outliers*).

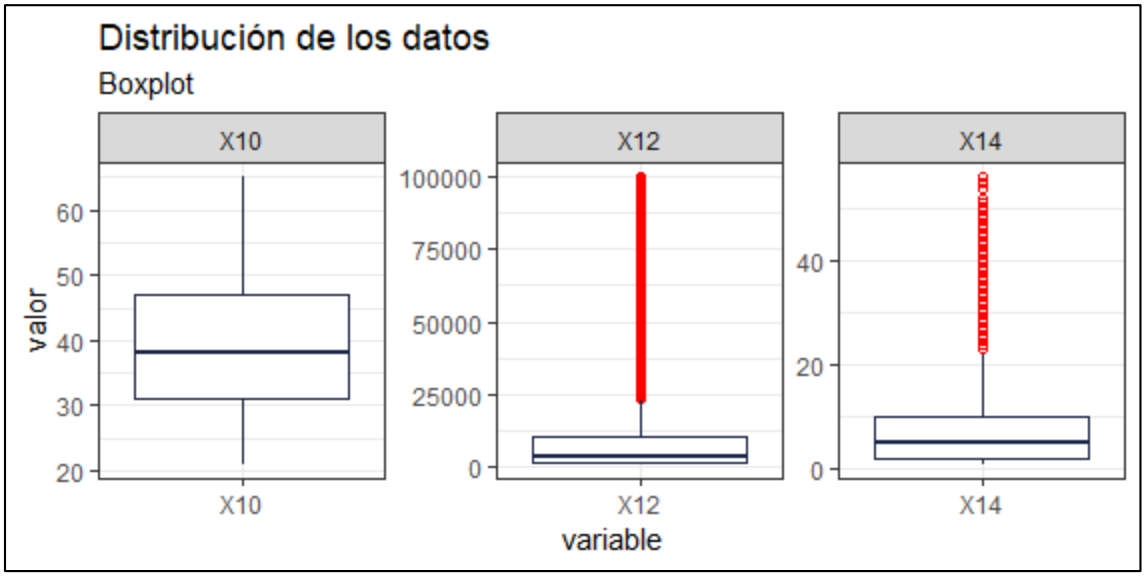


Figura 9. Gráfico de boxplot para X10, X12 y X12
Fuente: Elaboración propia

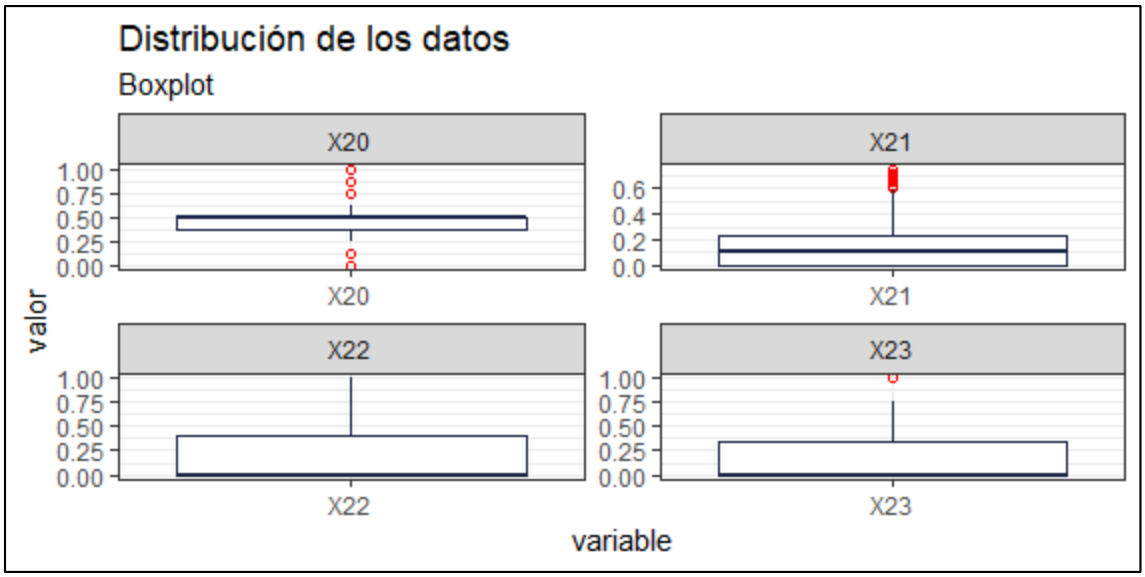


Figura 10. Gráfico de boxplot para X20, X21, X22 y X23
Fuente: Elaboración propia

Se realizó dos gráficos para nuestras variables numéricas pues hay un grupo de ellas (X20, X21, X22 y X23) que están relacionadas con el uso del protocolo de venta por parte del ejecutivo y se obtienen de la transcripción de las llamadas. A este grupo se le realiza un análisis adicional porque nuestro interés es en el uso del protocolo de venta cuando se aprueba el producto financiero.

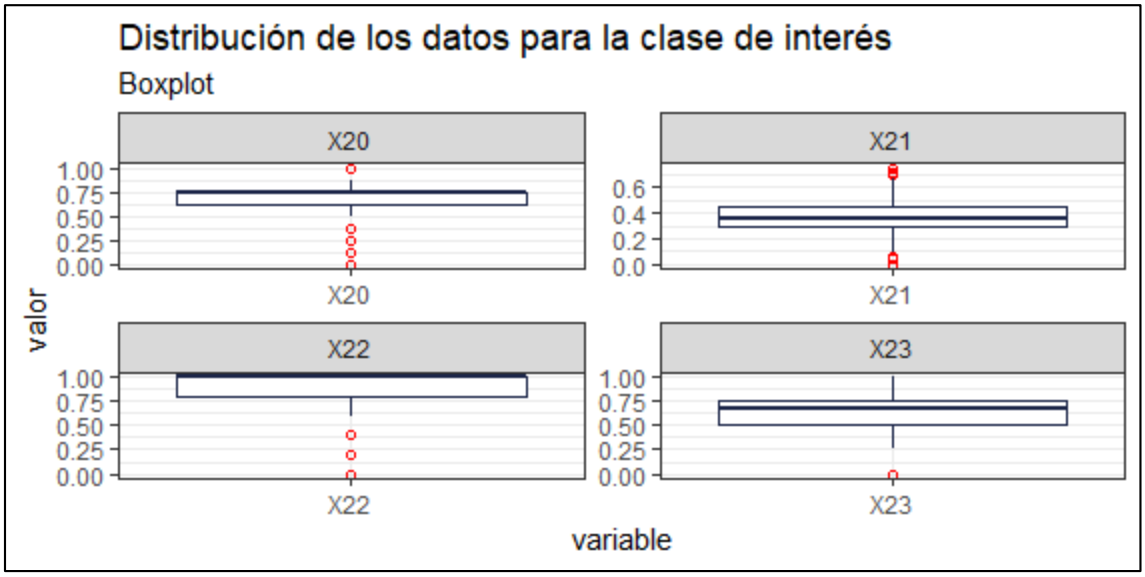


Figura 11. Gráfico de boxplot para X20, X21, X22 y X23 cuando X24=1
Fuente: Elaboración propia

Para trabajar en el modelado se usó la librería h2o, la cual está desarrollada en Java y se puede usar desde R, es muy eficiente al trabajar con grandes volúmenes de datos e incluye algoritmos de aprendizaje automático, además puede solucionar problemas de imputación de datos, balanceo de datos, etc. También se procedió a realizar una partición muestral estratificada en base a la variable *target* para garantizar que se mantenga la proporción de ésta en el entrenamiento y validación de los mismos, los datos se dividieron en 70% para el entrenamiento, 10% para validar el entrenamiento y 20% para la evaluación (*test*) del modelo.

Tabla 6
Partición muestral estratificada en base a la proporción de la target

Partición	Cantidad	%	Clase Negativa	Clase Positiva	% Clase Negativa	% Clase Positiva
Entrenamiento	177798	70	160257	17541	90	10
Validación	25352	10	22818	2534	90	10
Evaluación	50609	20	45541	5068	90	10

Como se observa en la Tabla 6, se mantiene la proporción de la variable *target* (10%) en las 3 particiones.

Se realizó un modelo base usando el algoritmo de *gradient boosting machine (gbm)*, con todas las variables que permiten determinar la aceptación para nuevos leads en futuras campañas, por lo tanto se excluyeron las variables X1, X2, X14, X19, X20, X21, X22, X23. El objetivo de realizar un modelo base es poder mejorarlo y si fuera posible tener una idea de los predictores que mayor influyen en el modelo.

Tabla 7
Importancia de variables en el modelo base

variable	importancia relativa	importancia escalada	porcentaje
X7	85457.7969	1	0.379552
X8	38226.6563	0.447316	0.16978
X3	26471.6875	0.309763	0.117571
X4	18811.9023	0.220131	0.083551
X12	15746.6084	0.184262	0.069937
X13	11778.4893	0.137828	0.052313
X9	9961.67383	0.116568	0.044244
X6	9086.57031	0.106328	0.040357
X17	2668.06714	0.031221	0.01185
X18	1703.34009	0.019932	0.007565
X11	1492.53833	0.017465	0.006629
X5	1165.86792	0.013643	0.005178
X10	1062.26709	0.01243	0.004718
X15	866.78949	0.010143	0.00385
X16	654.251587	0.007656	0.002906
X10	1062.26709	0.01243	0.004718
X15	866.78949	0.010143	0.00385
X16	654.251587	0.007656	0.002906

En la Tabla 7, se puede ver la importancia de todas las variables que ingresaron al modelo base ya sea de manera relativa, bajo la misma escala o porcentaje. Para una mejor observación de la importancia de las variables o predictores se muestra la figura 10.

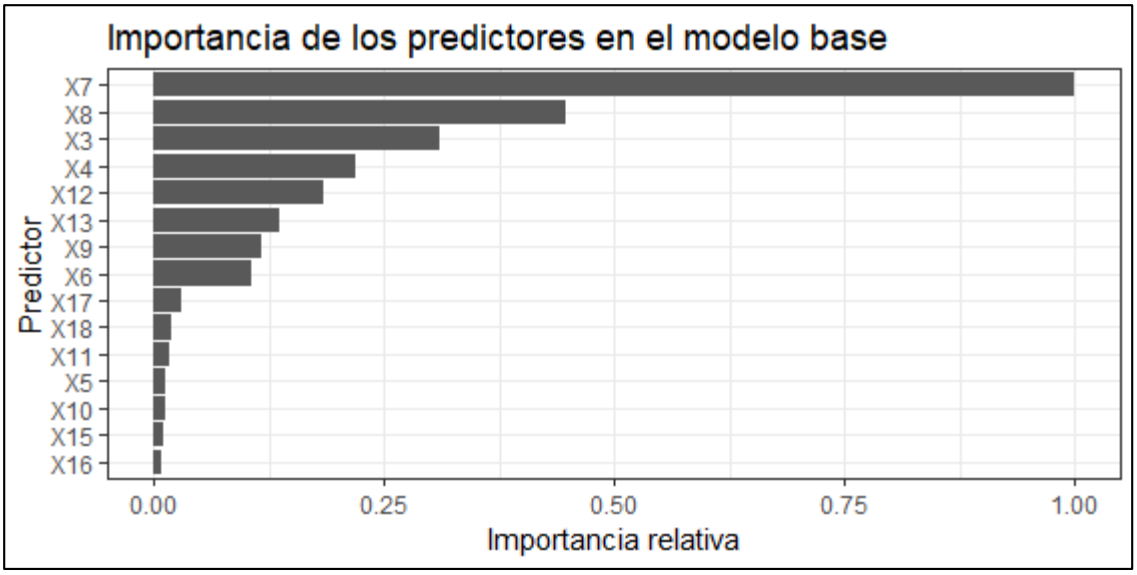


Figura 12. Importancia de los predictores en el modelo base
Fuente: Elaboración propia

Se observa que las variables (predictores) que mayor influyen en el modelo son la que se refiere a la recurrencia del lead en la campaña (X7), la zona de riesgo (X8), la prioridad de la gestión (X3), la propensión de ser contactado (X4), las referentes al monto que califica (X12 y X13), el nivel de riesgo (X9) y el flujo de riesgo (X6), el resto de las variables no influyen mucho.

Teniendo esta información se procedió a ejecutar tres modelos de aprendizaje automático para poder compararlos, de los cuáles dos de ellos están basados en árboles y uno en redes neuronales. Todos los modelos se trabajaron balanceando los datos por sobremuestreo *(oversampling)* como se puede ver en el anexo.

1. Modelo 1: Modelo basado en el algoritmo de *gradient boosting machine (gbm).*
2. Modelo 2: Modelo basado en el algoritmo de extreme *gradient boosting (xgboost).*
3. Modelo 3: Modelo basado en redes neuronales.

Al realizar el primer modelo se seleccionaron las siguientes variables dado su importancia y guardando relación con el negocio:

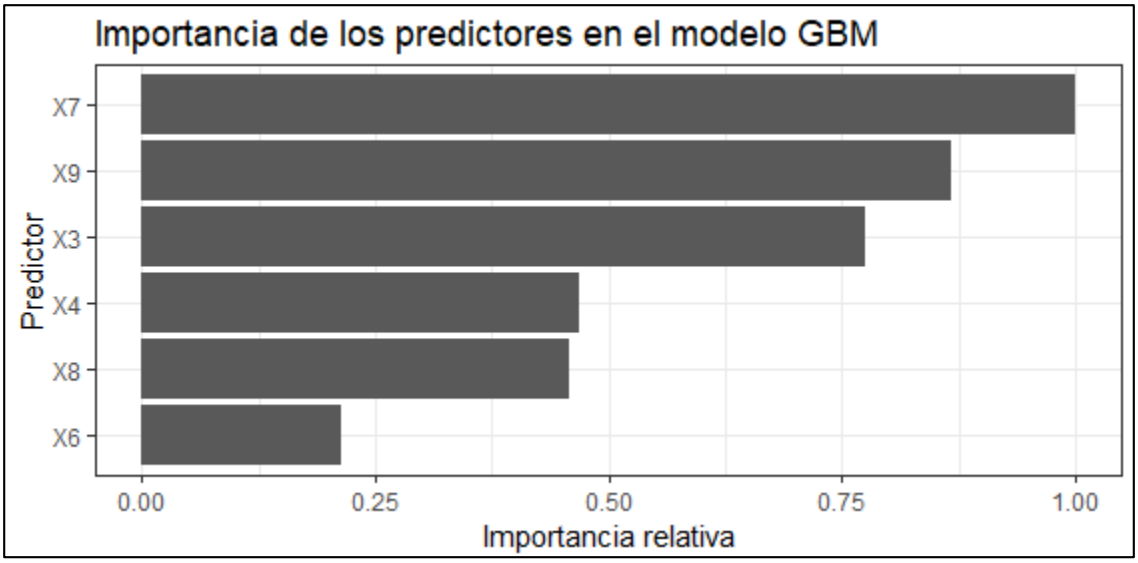


Figura 13. Importancia de variables modelo 1
Fuente: Elaboración propia

Las variables mostradas en la figura 13 también fueron usadas para los otros modelos, quedaron todas las variables que se maneja en la gestión propia de la base de datos.

Del modelo 1 se obtuvo los siguientes indicadores:

Tabla 8
Indicadores del modelo 1

Indicador	Valor (%)
AUC	73.11%
Sensibilidad	53.39%

Para el modelo 2, se ejecutó mediante el algoritmo *xgboost* y se obtuvo los siguientes resultados:

Tabla 9
Indicadores del modelo 2

Indicador	Valor
AUC	74.26%
Sensibilidad	82.89%

Para el modelo 3, se aplicó un algoritmo de redes neuronales con la siguiente arquitectura: (10, 10, 10, 10, 10, 10) es decir 6 capas ocultas con 10 neuronas por cada capa oculta. La función de activación usada fue ReLu (rectified linear unit), la cual está definida como $f(x) = Max(0, x)$. Obteniéndose los siguientes resultados:

Tabla 10
Indicadores del modelo 3

Indicador	Valor
AUC	58.70%
Sensibilidad	55.52%

Se procedió a construir una tabla, la cual contiene los tres modelos desarrollados con sus respectivos indicadores mostrados en las tablas anteriores.

Tabla 11
Indicadores de los tres modelos desarrollados

Modelo	AUC	Sensibilidad
Gbm	73.11%	53.39%
Xgboost	74.26%	82.89%
Redes Neuronales	58.70%	55.52%

Estos resultados se analizaron en la sección 4.2

4.2 Análisis de resultados

4.2.1 Indicadores de desempeño del modelo.

De la Tabla 11, se procedió a realizar la comparación de los 3 modelos desarrollados y se mostró que el modelo xgboost obtuvo el mejor desempeño pues obtuvo el mayor valor en el indicador AUC (74.26%), además se observó un alto valor en el indicador de sensibilidad 82.89% el cual es otra métrica dentro de los modelos de

clasificación y de gran importancia para el negocio pues esta métrica indica el porcentaje de acierto respecto al total real de la clase de interés, en este caso que el cliente haya obtenido el producto financiero. (Hossin, Mohammad; Sulaiman, Nasir ;, 2015) Detallan algunas métricas adicionales a las mostradas en las tablas anteriores desde su definición e interpretación y refuerzan su importancia en la elección de las mismas además de brindar bibliografía adicional sobre métricas de clasificación.

4.2.2 Gestión de la base de datos.

Una vez que se determinó el modelo, el área de televentas procedió a validarlo con nuevos datos de prueba conformado por 2 meses siguientes (mayo y junio 2018), para la medición de los resultados se usó el reporte de gestión de la base de datos el cual contiene los ratios intermedios que conforman la efectividad, especialmente contacto efectivo y aceptación. En ellos se pudo notar que se realizaba un esfuerzo similar en la gestión de la base de datos en *leads* a los que el modelo calificaba como 0 (*leads* que no calificarían al producto financiero) con los *leads* que el modelo calificaba como 1 (*leads* que calificarían al producto financiero) y representaba aproximadamente 45% de los *leads* que fueron contactados de manera efectiva y un 18% en ventas. Por lo tanto el grupo que el modelo calificaba como 1 tenía mayor efectividad que los que eran calificados como 0. Estos valores iban de acorde a los resultados que se esperaba mediante el modelo elegido.

En base a los resultados descritos anteriormente se procedió a realizar ciertas estrategias para la toma de decisiones en la gestión de la base de datos en busca de maximizar su eficiencia, como por ejemplo:

- Predecir la clasificación de los *leads* una vez recibida cualquier base de datos a ser gestionada.
- Dar prioridad de discado a los *leads* que el modelo predice en que obtendrán el producto sin descuidar el contacto efectivo.
- En caso no se tenga futuras recargas realizar un enriquecimiento de teléfonos a los *leads* que el modelo determina que obtendrán el producto.

- Perfilar los *leads* en busca del cumplimiento de los objetivos comerciales.

Se tiene en cuenta el punto de vista de (Mehrotra, Ankit ; Agarwal, Reeti ;, 2009) de entender las actitudes y preferencias de los clientes se procedió a analizar el cumplimiento del protocolo de venta en la gestión de la base de datos por parte de los ejecutivos, por lo tanto se procedió a analizar las variables X20, X21, X22 y X23 pero sólo de los *leads* que obtuvieron el producto financiero (cuando X24=1) en base a lo mostrado en la Figura 9. Se observó que la mediana del índice de servicio (X20) para aquellos *leads* está alrededor del 0.75 es decir hay un alto uso de este indicador, también llamó la atención que existan ventas con un índice menor a 0.5. La mediana del índice de información de beneficios (X21) está por debajo del 0.4 y con la presencia de valores atípicos. La mediana del índice de la información financiera (X22) está cerca de 1 y es de esperarse porque se da prácticamente cuando el ejecutivo de venta lee la parte legal del contrato al adquirir el producto aunque existen algunas ventas con un bajo índice. Por último tenemos el índice de gestión comercial (X23) y se observó que tiene un alto valor en la mediana esto es de esperarse porque muestra que el ejecutivo de venta utiliza argumentos para conseguir el cierre de la venta. Todo este análisis sirvió para determinar el perfil de la gestión de venta.

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

1. Se muestra la determinación de un modelo predictivo de referencia para la gestión de base datos para el área de televentas usando algoritmos de aprendizaje automático, como se pudo observar se realizaron tres modelos (GBM, XGBOOST y Redes Neuronales Artificales) y se seleccionó el modelo de XGBOOST pues obtuvo el mejor desempeño basado en AUC y sensibilidad, los valores fueron 74.26% y 82.89% respectivamente.
2. Dado la determinación del modelo de referencia se plantearon estrategias para la toma de decisiones en la gestión de la base de datos.
3. La priorización de los *leads* en la gestión de base de datos se da en base a los que el modelo predice que obtendrán el producto y los más propensos a ser contactados.
4. En base al perfil de la gestión, se comprobó que la mayoría de los *leads* que adquirieron los productos financieros tuvieron mayores indicadores en los índices comerciales y esto se da con el cumplimiento del protocolo de venta.

5.2 Recomendaciones

1. Se recomienda gestionar en base a los resultados del modelo para generar eficiencias en la gestión de la base de datos.
2. Siempre monitorear los ratios intermedios de contacto efectivo y aceptación en la campaña vigente.
3. Sería provechoso para el negocio contar con un buen modelo de propensión al contacto basado en la mejor hora para llamar al cliente potencial conocido en inglés como *best time to call*, pues creará mayor oportunidad para la aceptación del producto que se le oferta. Esto impactará en la priorización de los *leads* al momento de gestionarlos.

4. Se debe fomentar el uso de las categorías y replicar las buenas prácticas que tienen los ejecutivos de ventas más productivos en los ejecutivos con menor productividad.
5. Sería interesante conocer todas las retroalimentaciones (*feedback*) que tiene el cliente potencial al realizarle la llamada pues sólo tenemos por parte del ejecutivo de ventas la forma que este gestionó la llamada mediante el uso de categorías. Esto ayudará a conocer las necesidades del cliente y/o mejorar su experiencia al contactarlo.
6. Los resultados obtenidos podrían ser mejorados al incluir nuevas variables de gestión o contextuales a la gestión de *leads*.

REFERENCIAS BIBLIOGRÁFICAS

6.1 Referencias bibliográficas

Chen, Tianqi; Guestrin, Carlos;. (2016). *XGBoost: A Scalable Tree Boosting System.*

Flach, P. (2012). *Machine learning: The Art and science of algorithms that make sense of data.* Cambridge: Cambridge University Press.

Friedman, J. H. (Octubre de 2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics, 29*(5), 1189-1232.

Guido, Sarah; Mueller, Andreas;. (2016). *Introduction to Machine Learning with Python.* O'Reilly Media.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome;. (2009). *The Elements of Statistical Learning.* Berlin: Springer.

Hossin, Mohammad; Sulaiman, Nasir ;. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.2, March 2015*.

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert;. (2014). *An Introduction to Statistical Learning.* Berlin: Springer.

Kelleher, John D.; Namee, Brian Mac; Aoife, D'Arcy;. (2015). *Fundamentals of machine learning for predictive.* Massachusetts Institute of Technology.

Kotler, P. T., & Keller, K. L. (2016). *Framework for Marketing Management.* Pearson.

Kriesel, D. (2007). *A Brief Introduction to Neural Networks.* Obtenido de http://www.dkriesel.com/_media/science/neuronalenetze-en-zeta2-2col-dkrieselcom.pdf

Li, Xiao-Lin; Yu, Zhong;. (2012). An overview of personal credit scoring: techniques and. *International Journal of Intelligence Science*, 181-189.

McKinsey Global Institute. (2016). *The age of analytics: competing in a data-driven world.* USA.

Mehrotra, Ankit ; Agarwal, Reeti ;. (13 de Julio de 2009). Classifying customers on the basis of their attitudes towards telemarketing. *Journal of Targeting, Measurement and Analysis for Marketing*. Palgrave Macmillan.

Mitchell, T. (1997). *Machine Learning.* McGraw-Hill.

Moro, Sérgio; Cortez, Paulo; Rita, Paulo;. (2014). A data-driven approach to predict the success of bank telemarketing. *Elsevier*, 22-31.

Moro, Sérgio; Cortez, Paulo; Rita, Paulo;. (Enero de 2015). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. *Neural Computing & Applications*. Springer.

Moro, Sérgio; Laureano, Raul; Cortez, Paulo;. (2011). Using data mining for bank direct marketing: an application of the crisp-dm methodology. *Proceedings of the European Simulation and Modelling Conference - ESM'2011*, (págs. 117-121). Guimaraes.

Myazina, E. (2017). Machine Learning for Credit Scoring. *(Tesis de maestría).* Charles University, Praga.

Pinedo, Michael; Seshadri, Sridhar; Shanthikumar, J. George;. (2000). Call Centers in Financial Services: Strategies, Technologies, and Operations. En *Creating Value in Financial Services* (págs. 357-388). Boston: Springer.

Rokach, Lior; Maimon, Oded. (2014). *Data mining with decision trees.* Singapore: World Scientific.

Suthaharan, S. (2016). *Machine learning models and algorithms for big data classification.* Greensboro: Springer.

Tsiptsis, Konstantinos; Chorianopoulos, Antonios;. (2010). *Data Mining Techniques in CRM_ Inside Customer Segmentation.* Wiley.

Zhou, Z.-H. (2012). *Ensemble Methods.* Chapman & Hall / CRC Press.

6.2 Referencias electrónicas consultadas

Analytics Vidhya. (Abril de 2016). Obtenido de https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

Diario Gestión. (18 de Noviembre de 2017). Obtenido de https://gestion.pe/empresas/atento-peru-lanzara-noviembre-su-servicio-digital-y-apunta-crecimiento-mayor-al-8-anual-2205039?href=mas_leidas

H2O.ai. (10 de Junio de 2019). Obtenido de H2O.ai: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/gbm.html

ANEXOS

ANEXO 1: Declaración de Autenticidad

ANEXO 2: Autorización de Consentimiento para realizar la investigación

ANEXO 3: Script de análisis descriptivo

ANEXO 4: Script de modelos en R

```
########################################
####Código Tesis en R ###################
####Frank Rivera ########################
########################################
#### 1. Descriptivo de variables ####
########################################
rm(list=ls())
library(dplyr)
library(tidyverse)
library(dplyr)
library(ggplot2)
datos=read.csv("datos_tesis.csv",header = T)
colnames(datos)[1]="X1"
colnames(datos)[2]="X2"
colnames(datos)[3]="X3"
colnames(datos)[4]="X4"
colnames(datos)[5]="X5"
colnames(datos)[6]="X6"
colnames(datos)[7]="X7"
colnames(datos)[8]="X8"
colnames(datos)[9]="X9"
colnames(datos)[10]="X10"
colnames(datos)[11]="X11"
colnames(datos)[12]="X12"
colnames(datos)[13]="X13"
colnames(datos)[14]="X14"
colnames(datos)[15]="X15"
```

```
colnames(datos)[16]="X16"
colnames(datos)[17]="X17"
colnames(datos)[18]="X18"
colnames(datos)[19]="X19"
colnames(datos)[20]="X20"
colnames(datos)[21]="X21"
colnames(datos)[22]="X22"
colnames(datos)[23]="X23"
colnames(datos)[24]="X24"

library(Hmisc)
describe(datos)

####Boxplot de mis variables numéricas####
datos_num=select_if(datos, is.numeric)#selecciono solo numericas
names(datos_num)

#Para variables:X10,X12,X14
dat.g1 <-  datos_num %>%select(1,2,3)%>%
  tidyr::gather(key = variable, value = valor)
dat.g1 <-dat.g1%>% filter(!is.na(valor))
ggplot(data = dat.g1, aes(x=variable, y=valor)) + geom_boxplot(colour =
"#1c2649",outlier.colour = "red", outlier.shape = 1)+
  facet_wrap(~variable,scales = "free")+theme_bw()+ggtitle("Distribución de los datos",subtitle
= "Boxplot")

#Para variables X20,X21,X22,X23
dat.g1 <-  datos_num %>%select(4,5,6,7)%>%
  tidyr::gather(key = variable, value = valor)
dat.g1 <-dat.g1%>% filter(!is.na(valor))
```

```
ggplot(data = dat.g1, aes(x=variable, y=valor)) + geom_boxplot(colour =
"#1c2649",outlier.colour = "red", outlier.shape = 1)+
  facet_wrap(~variable,scales = "free")+theme_bw()+ggtitle("Distribución de los datos",subtitle
= "Boxplot")


####BoxPlot para X20,X21,X22,X23 pero cuando X24=1 ####
datos_num_1=datos[,c(20,21,22,23,24)]
names(datos_num_1)
dat.g2 <-  datos_num_1 %>%filter(X24=="1")%>%select(1,2,3,4)%>%
  tidyr::gather(key = variable, value = valor)
dat.g2 <-dat.g2%>% filter(!is.na(valor))
ggplot(data = dat.g2, aes(x=variable, y=valor)) + geom_boxplot(colour =
"#1c2649",outlier.colour = "red", outlier.shape = 1)+
  facet_wrap(~variable,scales = "free")+theme_bw()+ggtitle("Distribución de los datos para la
clase de interés",subtitle = "Boxplot")


########################################
#### 2.Modelos predictivos ####
########################################
rm(list=ls())
library(h2o)
library(dplyr)
library(tidyverse)
library(ggplot2)
library(tictoc)
# Creación de un cluster local con todos los cores disponibles.
h2o.init(ip = "localhost",
         # -1 indica que se empleen todos los cores disponibles.
         nthreads = -1,
         # Máxima memoria disponible para el cluster.
         max_mem_size = "16g")
```

```
datos <- read.csv("../input/datos_tesis.csv")
str(datos)

datos$X24=as.factor(datos$ X24)
datos=datos[,-1]
names(datos)

datos=datos[,c(7,9,4,3,8,6,24)]
names(datos)
datos_h2o=as.h2o(datos)

particiones <- h2o.splitFrame(data = datos_h2o, ratios = c(0.7, 0.10),
                              seed = 123)#123
datos_train_h2o <- h2o.assign(data = particiones[[1]], key = "datos_train_H2O")
datos_val_h2o   <- h2o.assign(data = particiones[[2]], key = "datos_val_H2O")
datos_test_h2o  <- h2o.assign(data = particiones[[3]], key = "datos_test_H2O")

dim(datos_train_h2o)[1]
dim(datos_val_h2o)[1]
dim(datos_test_h2o)[1]

predictores=names(datos_h2o[-c(7)])
var_respuesta=names(datos_h2o[c(7)])

##voy a balancear mi data de entrenamiento

library(ROSE)
#h2o.table(datos_train_h2o$ X24)[1,2]
datos_b=as.data.frame(datos_train_h2o)
data_balanced_over <- ovun.sample(X24 ~ ., data = datos_b,
```

```
                    method = "over", N = table(datos_b$ X24)[1]*2, seed = 123)$data
table(data_balanced_over$ X24)
data_balanced_over_h2o=as.h2o(data_balanced_over)

#### modelo gbm ####
modelo_gbm <- h2o.gbm(
  # Tipo de distribución (clasificación binaria)
  distribution = "bernoulli",
  # Variable respuesta y predictores.
  y = var_respuesta,
  x = predictores,
  # Datos de entrenamiento.
  training_frame = data_balanced_over_h2o,#datos_train_h2o,
  #balanceo
  #balance_classes = T,
  #kfolds
  nfolds = 5,
  # Datos de validación para estimar el error.
  validation_frame = datos_val_h2o,
  # Número de árboles.
  ntrees = 1000,
  col_sample_rate_per_tree = 0.3,
  col_sample_rate =0.8,
  # Complejidad de los árboles
  max_depth = 3,#3
  min_rows = 10,#10
  # Aprendizaje
  learn_rate = 0.003,
  # Detención temprana
  score_tree_interval = 5,
```

```
  stopping_rounds = 3,
  stopping_metric = "AUC",
  stopping_tolerance = 0.001,
  model_id = "modelo_gbm",
  seed = 123)#123

predicciones <- h2o.predict(object = modelo_gbm, newdata = datos_test_h2o)
predicciones

tabla=h2o.table(predicciones[,1], datos_test_h2o[,7])
tabla
summary(datos_test_h2o[,7],exact_quantiles=TRUE)
# AUC de test
h2o.performance(model = modelo_gbm, newdata = datos_test_h2o)@metrics$AUC#0.731077
sensibilidad=tabla[4,3]/(tabla[2,3]+tabla[4,3])
sensibilidad#0.5339384

#### Modelo Xgboost ####
modelo_xgboost <- h2o.xgboost(x = predictores,
                              y = var_respuesta,
                              training_frame = data_balanced_over_h2o,
                              distribution = "bernoulli",
                              ntrees = 1000,
                              col_sample_rate_per_tree = 0.4,
                              col_sample_rate =0.40,
                              max_depth = 5,
                              min_rows = 100,
                              learn_rate = 0.5,
                              nfolds = 5,
                              fold_assignment = "Modulo",
```

```
                    score_tree_interval = 5,
                    stopping_rounds = 3,
                    stopping_metric = "AUC",
                    stopping_tolerance = 0.001,
                    keep_cross_validation_predictions = TRUE,
                    model_id = "modelo_xgboost",
                    seed = 123)

predicciones2 <- h2o.predict(object = modelo_xgboost, newdata = datos_test_h2o)
predicciones2

tabla2=h2o.table(predicciones2[,1], datos_test_h2o[,7])#7
tabla2
summary(datos_test_h2o[,7],exact_quantiles=TRUE)
# AUC de test
h2o.performance(model = modelo_xgboost, newdata = datos_test_h2o)@metrics$AUC
sensibilidad2=tabla2[4,3]/(tabla2[2,3]+tabla2[4,3])
sensibilidad2 # 0.7426 de AUC y sensibilidad de 0.8289

#### Modelo Redes ####
set.seed(123)
tic()
m1 <- h2o.deeplearning(
  model_id="modelo_redes",
  training_frame=data_balanced_over_h2o,
  validation_frame=datos_val_h2o,
  x=predictores,
  y=var_respuesta,
  activation="RectifierWithDropout",
  input_dropout_ratio = 0.1,
```

```
  hidden_dropout_ratios = c(0.2,0.2,0.2,0.2,0.2,0.2),
  hidden=c(10,10,10,10,10,10),
  epochs=2000,
  balance_classes = T,
  seed=123,
  loss = "CrossEntropy",
  adaptive_rate=F,
  nfolds=5,
  rate=0.0001,
  variable_importances=T,
  stopping_rounds=2,
  stopping_metric="logloss",
  stopping_tolerance=0.001
)
toc()
summary(m1)

pred <- h2o.predict(m1, datos_test_h2o)
pred

tabla=h2o.table(pred[,1], datos_test_h2o[,7])#7
tabla
summary(datos_test_h2o[,7],exact_quantiles=TRUE)
# AUC de test
h2o.performance(model = m1, newdata = datos_test_h2o)@metrics$AUC #58.70%
sensibilidad=tabla[4,3]/(tabla[2,3]+tabla[4,3])
sensibilidad #55.52%

# Apagado del cluster H2O
h2o.shutdown(prompt = FALSE)
```

UNIVERSIDAD RICARDO PALMA – ESCUELA DE POSGRADO
MAESTRIA EN CIENCIAS DE LOS DATOS

ANEXO E: MATRIZ DE CONSISTENCIA INTERNA DEL PROYECTO DE INVESTIGACIÓN

PROBLEMA	OBJETIVO	HIPOTESIS	VARIABLE	DIMENSIÓN
GENERAL ¿De qué manera un modelo predictivo de referencia que determine la aceptación de un producto financiero afecta en la gestión de la base de datos para la adquisición de nuevos clientes vía telefónica en una campaña vigente de productos financieros?	GENERAL Determinar un modelo predictivo de referencia que determine la aceptación de un producto financiero basado en la gestión de la base de datos para la adquisición de nuevos clientes vía telefónica en una campaña vigente de productos financieros.	GENERAL Afecta de manera positiva el uso de un modelo predictivo de referencia que determine la aceptación de un producto financiero en la gestión de la base de datos para la adquisición de nuevos clientes vía telefónica en una campaña vigente de productos financieros.	INDEPENDIENTE Modelo predictivo de referencia.	
ESPECIFICOS 1. ¿Cómo afecta en el negocio de adquisición de clientes vía telefónica el uso del algoritmo *Gradient Boosting Machine* (GBM) en la determinación de la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales ? 2. ¿Cómo afecta en el negocio de adquisición de clientes vía telefónica el uso del algoritmo *Extreme Gradient Boosting* (GBM) en la determinación de la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales ?	ESPECIFICOS 1. Aplicar el algoritmo de aprendizaje automático de *Gradient Boosting Machine* (GBM) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros. 2. Aplicar el algoritmo de aprendizaje automático de *Extreme Gradient Boosting* (XGBOOST) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.	ESPECIFICOS Se podrá aplicar el algoritmo de aprendizaje automático de *Gradient Boosting Machine* (GBM) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros. Se podrá aplicar el algoritmo de aprendizaje automático de *Extreme Gradient Boosting* (XGBOOST) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.	DEPENDIENTE Algoritmos de aprendizaje automático.	Indicadores de desempeño del modelo: AUC y sensibilidad

3. ¿Cómo afecta en el negocio de adquisición de clientes vía telefónica el uso del algoritmo de Redes Neuronales Artificiales (RNA) en la determinación de la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales?	3. Aplicar el algoritmo de aprendizaje automático de Redes Neuronales Artificiales (RNA) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.	Se podrá aplicar el algoritmo de aprendizaje automático de Redes Neuronales Artificiales (RNA) para determinar la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros.		
4. ¿Qué diferencias existen al comparar los algoritmos de aprendizaje automático para la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros en base a los indicadores: AUC y sensibilidad?	4. Comparar los algoritmos de aprendizaje automático para la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros en base a los indicadores: AUC y sensibilidad.	Existen diferencias al comparar los algoritmos de aprendizaje automático para la aceptación de un producto financiero basado en la gestión de llamadas a clientes potenciales en una campaña vigente de productos financieros en base a los indicadores: AUC y sensibilidad.		