UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"PREDICCIÓN DE ADQUISICIÓN DE UN PRÉSTAMO PERSONAL BANCARIO A TRAVÉS DEL CANAL DE TELEVENTAS UTILIZANDO EL ALGORITMO RANDOM FOREST"

TRABAJO DE SUFICIENCIA PROFESIONAL
PARA OPTAR EL TÍTULO DE
INGENIERA EN ESTADÍSTICA E INFORMÁTICA

FIORELLA PAMELA DE LA CRUZ FLORES

LIMA – PERÚ
2020

UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"PREDICCIÓN DE ADQUISIÓN DE UN PRÉSTAMO PERSONAL BANCARIO A TRAVÉS DEL CANAL DE TELEVENTAS UTILIZANDO EL ALGORITMO RANDOM FOREST"

Presentado por:

FIORELLA PAMELA DE LA CRUZ FLORES

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL TÍTULO DE INGENIERA ESTADÍSTICA E INFORMÁTICA

SUSTENTADO Y APROBADO ANTE EL SIGUIENTE JURADO:

Dr César Higinio Menacho Chiok	Mg. Jesús Walter Salinas Flores
PRESIDENTE	**ASESOR**
MS. Grimaldo José Febres Huamán	Dr. Jorge Chue Gallardo
MIEMBRO	**MIEMBRO**

LIMA – PERÚ

2020

DEDICATORIA

A mi papá, por enseñarme a luchar por mis sueños.
Siempre estarás en mi corazón.
Un beso hasta el cielo.

AGRADECIMIENTO

A mi mamá, mi papá y mis hermanos, por apoyarme siempre en mis decisiones e impulsarme a seguir creciendo profesionalmente.
A mi enamorado, por motivarme en los momentos de cansancio e impulsarme a culminar la tesis.
Finalmente, al profesor Jesús Salinas, por asesorarme y aconsejarme durante la elaboración de este trabajo.

¡Muchas gracias a cada uno por su apoyo incondicional!

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

I. PRESENTACIÓN

La entidad donde se realizó el presente trabajo es una de las empresas de Contac Center líder en Latinoamérica. Dentro de sus principales subcontrataciones se encuentran los servicios de: atención al cliente, soporte técnico, gestión de cobranzas y gestión de ventas.

El presente trabajo describe la elaboración de un modelo de clasificación para el servicio de ventas del Contac Center que permitió predecir a los clientes que aceptan un préstamo personal en una entidad bancaría, que ya cuentan con una tarjeta de crédito, luego de ofrecerles el producto mediante llamadas telefónicas.

Esto fue planteado debido a que los resultados de las ventas no eran los esperados por el cliente interno; ya que, la gestión de llamadas no se estaba trabajando de la manera más eficiente, lo que ocasionaba que el coste de la operación para esta campaña sea más alto con respecto a otras similares que se trabajaban en la empresa. En consecuencia, para cumplir con los objetivos, los asesores y supervisores tenían que trabajar más horas de las debidas, generando mayores costos que con una correcta gestión se podían reducir significativamente.

Los resultados obtenidos permitieron gestionar de forma eficiente la cartera de clientes en función a la predicción, logrando incrementar el volumen de ventas del producto. Además, se optimizaron los recursos utilizados por la empresa, minimizando los costos operativos.

II. INTRODUCCIÓN

Como parte del crecimiento económico de las entidades financieras, continuamente buscan captar nuevos clientes y fidelizar a los que ya tienen en cartera. Para lograr este objetivo, los bancos se han visto en la necesidad de explotar los datos de su cartera de clientes de forma tal que puedan encontrar patrones para detectar a los clientes potenciales, para así poder ofrecerles un producto personalizado con la finalidad de atender sus distintas necesidades.

Uno de los productos más demandados en el sistema financiero, es el préstamo personal, definido como la entrega de una suma de dinero a una persona específica con la condición de devolver la suma prestada con una tasa de interés y un plazo definidos previamente con la entidad bancaria. Para el caso de este trabajo, el producto es ofrecido solo a los clientes que ya cuentan con una o más tarjetas de crédito en la entidad financiera.

Para poder ofrecer el préstamo a sus clientes, la empresa contrata los servicios de otra entidad, especializada en brindar servicios integrales de servicio al cliente, soporte técnico y gestión de ventas, conocida como Call Center. Esta empresa se encarga de la gestión de la cartera que comparte el cliente, garantizando vender el producto a sus consumidores potenciales, generando así ingresos a la empresa que contrato el servicio a un menor costo.

Cada quince días el banco comparte con el área de operaciones de la campaña de préstamos, la base de datos con la cartera de clientes a trabajar en ese periodo. Una vez procesada la base, el supervisor de la operación establece estrategias de segmentación de clientes, gestionando las llamadas entre sus asesores telefónicos buscando cumplir con los objetivos establecidos por el cliente. Sin embargo, estos objetivos no se cumplían de manera eficiente, debido a que esta segmentación no estaba funcionando de la manera adecuada.

Por esta razón, el área de Business Intelligence, donde desempeñé labores como analista de Business Analytics, se encarga de brindar soporte a la operación de manera tal, que se logre optimizar los resultados trabajando de manera más eficiente con la cartera de clientes y recursos del Call Center, generando así mayores ingresos a la empresa.

Por tal motivo, el objetivo de este trabajo es desarrollar un modelo de clasificación binaria para predecir si el cliente va a aceptar o rechazar el préstamo ofrecido por los asesores al contactarlos vía telefónica. Para ello, se utilizó el algoritmo Random Forest que permitió predecir a los clientes con mayor probabilidad a adquirir el producto y así, gestionarlos óptimamente para priorizar su venta.

Para el desarrollo del modelo se usó una muestra de seis meses (desde marzo 2017 hasta agosto 2017) de datos de los clientes que cuentan como mínimo con una tarjeta de crédito y tienen un préstamo pre- aprobado con la entidad bancaria. En total la base contó con 991 619 registros de clientes.

III. OBJETIVOS

3.1. Objetivo general

Predecir la adquisición de un préstamo personal bancario a través del canal de televentas utilizando el algoritmo Random Forest.

3.2. Objetivos específicos

- Aplicar el algoritmo Random Forest para predecir a los clientes que adquirirán un préstamo personal en una entidad financiera.
- Evaluar la capacidad de predicción del modelo de clasificación usando como indicadores la sensibilidad y la especificidad.
- Comparar las técnicas de balanceo de datos Undersampling y SMOTE para obtener un mejor resultado en la predicción del modelo.

IV. CUERPO DEL TRABAJO

4.1. Funciones desempeñadas

Agencia de medios y publicidad (agosto 2016 a marzo 2017)

Se desempeñaron labores específicamente en el área de publicidad digital con el cargo de "Professional Intern" por aproximadamente nueve meses. Entre las principales funciones estaban: implementar campañas en redes sociales, buscadores web y plataformas de streaming como YouTube. Posterior a la implementación, se realizaban reportes de seguimiento semanales de los principales indicadores de campañas como: número de vistas del anuncio, número de clics al enlace adjunto y compras directas.

Call-Center (abril 2017 a febrero 2018)

Se desempeñaron labores como "Analista de Business Analytics" en el área de "Business Intelligence", donde una de las funciones principales era implementar modelos predictivos para las campañas de ventas, refinanciamiento y telefonía. En función a los resultados del modelo, se realizaba la segmentación de la cartera de clientes en base a la probabilidad predicha de adquisición del producto. Entre otras funciones que se desempeñaron estaban: elaboración de reportes de los indicadores de desempeño de gestión (KPI'S) y proyección de ventas mensuales de las campañas asignadas.

Agencia de medios y publicidad (marzo 2018 a mayo 2019)

Se desempeñaron labores como "Analista de Data Business Intelligence" en el área digital de "Data & Analytics" de una agencia de medios. Lo aprendido anteriormente, permitió ser parte del proyecto de "transformación digital" que tenía como reto la empresa el año 2018.
Este proyecto tenía como objetivo la automatización de los procesos de: extracción, consolidación y limpieza de los datos no estructurados extraídos de las distintas fuentes, tanto para digital (redes sociales, webs y plataformas de programática) como para medios tradicionales (televisión, radio, revistas, etc.) mediante APIs en lenguaje PHP.

Posteriormente, se utilizaron los datos para elaborar dashboards en Power Bi con los indicadores de seguimiento de las campañas de manera que permitan a los clientes internos tomar decisiones en sus campañas basándose en los resultados obtenidos un día anterior. Se aplicaron también algunas técnicas estadísticas para la estimación de los principales indicadores mostrados en los reportes.

Entidad bancaria (junio 2019 a la actualidad)

Se desempeñan labores en el área de "Riesgos de una entidad bancaria" con el cargo de "Analista de estrategia Analytics", en donde gracias a mi experiencia laboral, se aporta con iniciativas como modelamiento predictivo para mejorar los otorgamientos de líneas de crédito a los clientes del banco, montos de préstamos bancarios, entre otros productos. Además, de dar soporte a otras áreas, respondiendo las preguntas del negocio en base a análisis de las diversas fuentes de datos y así dar soluciones de manera óptima al negocio.

4.2. Puesta en práctica de lo aprendido en la carrera

Con la disponibilidad de gran volumen de datos históricos de las campañas que se manejan en la empresa, se busca responder las necesidades del negocio en función al análisis de estas bases de datos. Según lo aprendido durante la carrera, mis actividades estaban más relacionadas a la aplicación de técnicas estadísticas y manejo de base de datos. Para dar solución a la problemática de este caso en particular, se hizo uso de alguna técnica o algoritmo de minería de datos.

4.2.1. Algoritmo de Machine Learning Random Forest

Una de las metodologías de Minería de datos empleadas es el Machine Learning, una rama de la inteligencia artificial. Brange (2013) cuyo objetivo es utilizar los datos para realizar predicciones a situaciones ya conocidas.

El problema por tratar es de aprendizaje supervisado; pues se conocen los valores de la variable a predecir, es decir, se busca saber si un cliente contactado mediante una llamada telefónica va a "aceptar" o "rechazar" la venta del producto. Específicamente, el algoritmo que se usó fue Random Forest; ya que, al ser un método de ensamblaje que permite tener una mayor precisión y estabilidad en el modelo (Cichosz, 2015).

A continuación, se describirán las técnicas aplicadas en la solución de la problemática en la empresa.

a. Balanceo de datos

Uno de los problemas más comunes al realizar una técnica de aprendizaje supervisado para clasificación, es que las clases de la variable dependiente no están representadas de manera equitativa, es decir, una clase supera en gran medida a la otra clase. Según Arrieta & Mera (2015) la falta de balanceo de datos puede influir de manera negativa al rendimiento del modelo; ya que, puede provocar un sobreajuste o la omisión de información relevante.

Para poder balancear la cantidad de datos de cada clase, existen métodos de muestreo como:

- **Undersampling**: Según Arnejo (2017), consiste en eliminar muestras al azar de la clase con mayor cantidad de datos (Figura 1). A pesar de tener la ventaja de su bajo costo, emplearlo puede ser riesgoso; ya que, se puede perder información relevante para el modelo al eliminar información.

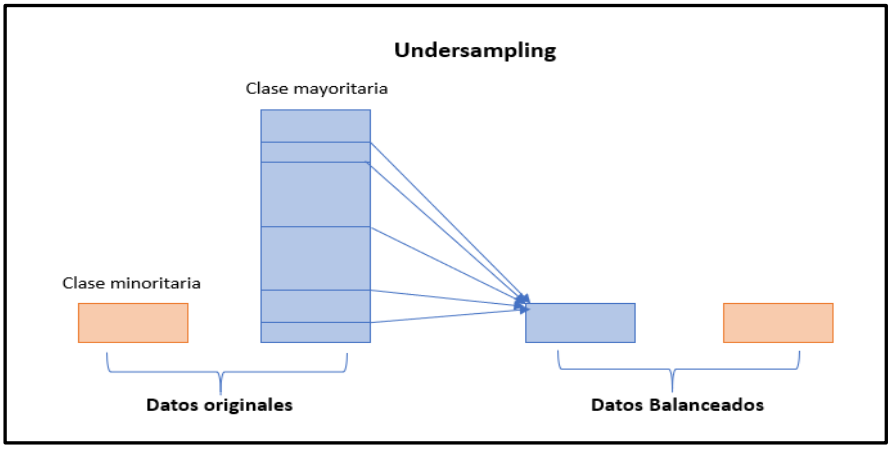

***Figura 1*: Método de balanceo Undersampling**

FUENTE: Elaboración propia

- **Oversampling:** Según Analytics Vidhya (2016) lo opuesto al primer método mencionado es el sobre muestreo. El método "oversampling" duplica al azar

muestras de la clase con menor cantidad de datos, y se agrega al conjunto de datos. Este método tiene un costo muy elevado y puede ocasionar un sobreajuste al modelo.

Otra forma que balancear los datos que funciona de forma más robusta con respecto a las mencionadas líneas arriba es el método SMOTE.

- **SMOTE (Synthetic Minority Oversampling Technique):** Chawla (2002), lo define como un método consiste en crear datos artificiales para la clase minoritaria interpolando los valores reales de los datos más cercanos de esa misma clase. Según Analytics Vidhya (2016), esto se hace para que el sesgo de aprendizaje se incline hacia la clase minoritaria. En la Figura 2 se puede apreciar como funciona el método.

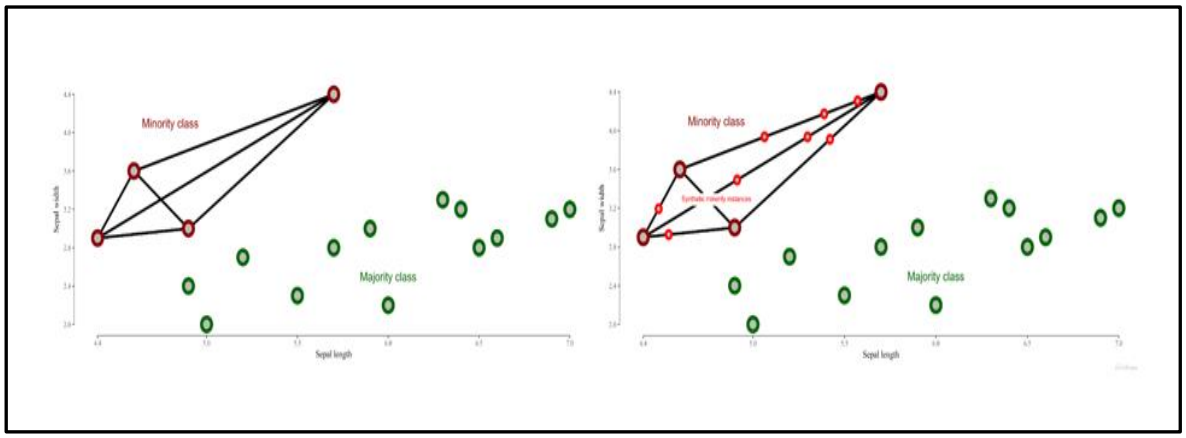

***Figura 2*: Método SMOTE**

FUENTE: Rikunert (2017)

b. Random Forest

Cichosz (2015) lo define como un algoritmo de aprendizaje supervisado que combina dos enfoques para la construcción del modelo: la técnica bagging que es usada para reducir la varianza de las predicciones atraves de la eleccion de distintas muestras con reemplazo para la construcción de los distintos árboles de decisión. Y por otro lado, la aleatoriedad, pues para la construcción de cada árbol las variables y observaciones son elegidas aleatoriamente.

Lateef (2019) indica que cuando mayor sea el número de árboles a usar en el algoritmo, mayor será la precisión de los resultados del algoritmo. Por otra parte, al ser un modelo de ensamblaje, aplica "trade-off" que es el equilibrio el error de sesgo y varianza.

Random Forest selecciona muestras al azar, cada una con variables y registros diferentes entre si para crear distintos conjuntos de datos. Estos subconjuntos se usan para elaborar un árbol de decisión con cada muestra. Luego de crear los múltiples árboles de decisión, cada árbol vota por una clase de la variable a predecir, finalmente la clase con la mayoría de votos se escoge como la clase predicha (Figura 3).

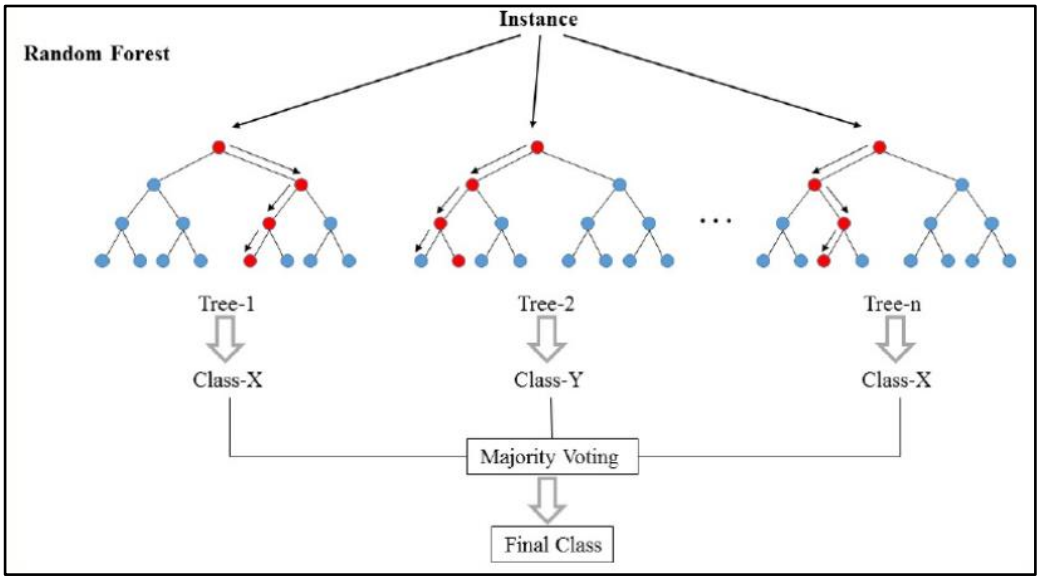

***Figura 3*: Proceso del algoritmo Random Forest**

FUENTE: Koehrsen (2017)

- **Estimación del error en Random Forest (OOB error)**

  Al momento de crear los árboles para el algoritmo Random Forest, existe una muestra que no es utilizada para modelar el árbol, este conjunto de datos es conocido como datos fuera de la bolsa OOB (Out Of Bag). Estos datos son aprovechados como muestras de validación para calcular el error de predicción de este algoritmo, denominado error fuera de la bolsa (OOB error) (Cichosz, 2015).

Lateef (2019) define la tasa de error OOB como la proporción de casos que fueron clasificados de manera erronea respecto a la clase real. Siendo un estimador insesgado del error de predicción.

- **Ventajas del algoritmo Random Forest**
  1. Como es un algoritmo de ensamblaje, la combinación de múltiples árboles hace que los resultados sean más precisos que otro método de ensamblaje. Por lo que puede ser utilizado en distintas áreas como medicina, banca, marketing, entre otros.
  2. Puede manejar gran volumen de variables de entrada e identificar las más significativas sin generar mucho costo computacional.
  3. Es posible usarlo como método no supervisado y detectar outliers.

- **Desventajas del algoritmo Random Forest**
  1. Perdida de la interpretación.
  2. No se tiene control en lo que se hace en el modelo.

c. Validación del modelo

Para garantizar el éxito de un modelo predictivo es necesario evaluar su desempeño. Según Torgo (2017), uno de los métodos más comunes para evaluar la predicción de un modelo es validación cruzada k-Fold.

Este método consiste en dividir los datos de entrenamiento en k subconjuntos, de los cuales un subconjunto es usado como prueba y los demás (k-1) son usados para datos de entrenamiento (Figura 4). Esta distribución se repite en k iteraciones, para finalmente calcular la precisión y error por cada modelo, para luego, promediar los resultados de cada iteración y obtener el resultado final. Fernández (2018), indica que al utilizar este método lograron reducir el sesgo y la variabilidad en la estimación del rendimiento del modelo.

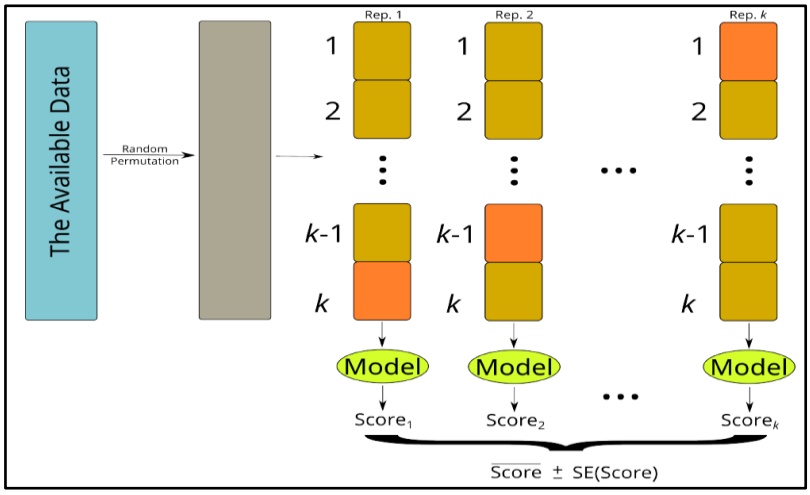

***Figura 4*: Proceso del método validación cruzada k-fold**

FUENTE: Torgo (2017)

d. Métricas de evaluación para el modelo de clasificación

Para la evaluación de un modelo de clasificación binaria principalmente se usa la matriz de confusión. Esta tabla cruzada contiene información de los valores reales comparados con los valores predichos por el método usado (Kuhn & Johnson, 2013). Dentro de la matriz de confusión tenemos cuatro posibles valores (Figura 5), tomando como ejemplo nuestro caso de "Venta" y "No venta": (1) si los valores predichos son clasificados como "Venta" cuando realmente son "Venta" se denominan verdaderos positivos (VP), (2) si los valores son clasificados como "Venta" cuando en realidad son "No Venta", se denominan, falsos positivos (FP), (3) si los valores son clasificados como "No Venta" cuando en efecto son "No Venta" se denominan, verdaderos negativos (VN) y (4) si los valores son clasificados como "Venta" cuando en realidad son "No venta", se denominan, falsos negativos (FN).

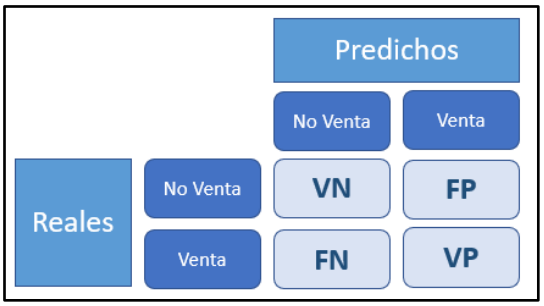

***Figura 5*: Matriz de confusión**

FUENTE: Elaboración propia

- **Sensibilidad:** Kuhn & Johnson (2013), lo define como la tasa verdadera positiva, pues mide la precisión a los casos positivos, para nuestro ejemplo la tasa de "Venta". Se calcula de la siguiente forma:

$$Sensibilidad = \frac{VP}{VP + FN}$$

- **Especificidad:** Según Kuhn & Johnson (2013), se define como la tasa de predicción de los casos negativos, para este ejemplo, la proporción de "No venta" sobre los casos que efectivamente son "No venta". Se calcula de la siguiente forma:

$$Especificidad = \frac{VN}{VN + FP}$$

4.2.2. Antecedentes de lo aplicado

Cubiles (2017) aplicó el algoritmo Random Forest para evaluar el riesgo crediticio en una entidad bancaria. Para este caso se compararon dos técnicas de aprendizaje supervisado, esto ayudaría a la empresa a evaluar que clientes tendrían mayor riesgo a morosidad que otros, anticipando a otorgarles el crédito a los clientes que no serán morosos y optimizando sus costos de operatividad. Los datos utilizados para este trabajo no estaban balanceados, por lo que se usó el método de muestro "undersampling" para balancearlos y luego aplicar los modelos con: (1) datos balanceados con Random Forest y (2) datos balanceados con árboles de clasificación. Los resultados obtenidos con el segundo modelo obtuvieron un 10% más de exactitud al predecir mejor a un cliente "riesgoso" o "normal", si se le otorgará un microcrédito en la entidad bancaria, con respecto al primer modelo.

Cárdenas (2019) usó el algoritmo Random Forest para clasificar si un cliente en campaña aceptará o rechazará adquirir una tarjeta de crédito en una entidad financiera. El área de CRM de dicho banco al aplicar este modelo buscaba mejores ganancias al reducir el costo de sus recursos; ya que, con la correcta gestión de cartera a través de sus canales de atención, priorizaron la atención de los clientes potenciales a través de sus mejores canales y dejando la atención de los clientes menos probables a adquirir la tarjeta de crédito a los canales con menor difusión. Se comparó los resultados de cuatro modelos aplicando el algoritmo, estos

fueron: (1) modelo sin balancear los datos, (2) modelo balanceando los datos mediante el método SMOTE, (3) modelo sin balancear los datos y utilizando un tuneo de parámetros para encontrar los mejores valores y (4) modelo con balanceo SMOTE y tuneo de parámetros. Los mejores resultados se obtuvieron con el último modelo, que presentó 79% de sensibilidad.

Hidalgo Ruiz-Capillas (2014) afirma que el algoritmo Random Forest es una buena solución al problema de clasificación de fraude. En su estudio se buscaba detectar si una transacción es fraudulenta en base a la información de estas y la detección de patrones de comportamiento anómalos de los clientes. Los datos para esta área también estaban desbalanceados, por ello se realizó el balanceo de datos mediante el método SMOTE, con los que se obtuvieron mejores resultados en comparación a los que se obtuvieron utilizando el método de oversampling.

Independientemente del algoritmo a usar al realizar un modelo predictivo de clasificación, si no se balancean los datos de entrenamiento no se lograrán obtener resultados óptimos que puedan usarse para predecir. Por esta razón, Pariona Huarhiachi (2017) utilizó el método SMOTE para balancear los datos de entrenamiento de un modelo de regresión logística para la clasificación de fuga de clientes en una entidad financiera. Para la variable dependiente, la "fuga" o "no fuga" de clientes contaba con una proporción de 91% para la clase "no fuga" y un 9% para la clase "fuga".

Si se mantienen estos valores los resultados del modelo van a tener un sesgo hacía la clase mayoritaria, lo que ocasionaría tener muy buenos resultados para los clientes con "no fuga" y resultados poco favorables para la clase "fuga", que es la que nos interesa predecir. Aplicando el método SMOTE se obtuvo un 56% para la clase "fuga" y un 44% para la clase "no fuga". Con estos datos se realizaron tres modelos: (1) modelo de regresión sin balancear los datos, (2) modelo de regresión usando el método de muestreo "undersampling" y (3) modelo de regresión logística usando SMOTE. Se comprobó con los resultados que el tercer modelo, obtuvo mejores resultados se sensibilidad y especificidad con respecto a los otros dos, a pesar de que el segundo modelo también uso balanceo de datos.

4.2.3. Propuesta de solución

a. Comprensión del negocio o problema

El objetivo mensual de la campaña que ofrecía un préstamo personal en la entidad bancaria era obtener mensualmente con las ventas del producto un monto de colocación de aproximadamente 3 millones de soles. Si este objetivo no se cumplía, la empresa Call Center obtenía un 2.5% menos del total de las ganancias obtenidas por las ventas.

En base a este problema, se tuvo la necesidad de implementar un modelo de clasificación que pueda predecir si un cliente en campaña, contactado vía telefónica por un asesor, va a aceptar el producto ofrecido, para este caso, el préstamo personal con la entidad bancaria.

De esta manera el área de operaciones podía priorizar la gestión de los clientes más probables a adquirir el préstamo bancario con los mejores asesores, contactándolos además en función a los horarios de mejor contactabilidad de forma que se pueda asegurar el cumplimiento del objetivo mensual planteado.

Para este proyecto se usó la metodología CRISP – DM (Cross Industry Standard Process for Data Mining) uno de los métodos más utilizados para orientar trabajos de minería de datos. En el siguiente gráfico (Figura 6) se describe los pasos que comprende esta metodología.

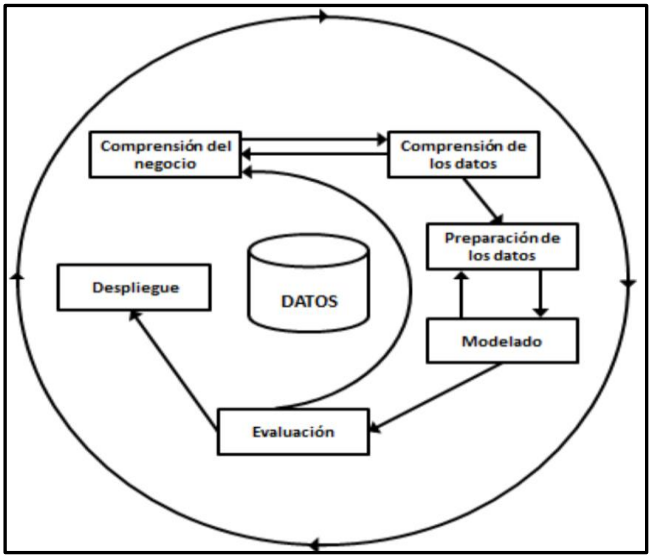

***Figura 6*: Etapas de la metodología CRISP-DM**

FUENTE: Research Gate (2014)

b. Comprensión de los datos

Para la aplicación del algoritmo se usaron los datos históricos de ocho meses de la campaña de ventas de préstamos personales en la entidad financiera que contrato los servicios del Call Center. Los datos usados contemplan los periodos de marzo del 2017 a agosto del mismo año, teniendo en total 991 619 registros.

La variable dependiente denominada "Venta" presenta dos clases:

- **Cliente acepta PP:** se tipifica a aquellos clientes que aceptan el préstamo personal al ser contactados telefónicamente. Se codifica con el valor "1" y se etiqueta como "Venta".
- **Cliente rechaza PP:** se tipifica a aquellos clientes que rechazan el préstamo personal al ser contactados telefónicamente por diversas razones. Se codifica con el valor "0" y se etiqueta como "No Venta".

La distribución y proporción de la variable se presenta en la tabla que se muestra a continuación (Tabla 1):

Tabla 1: ***Tabla de distribución de la variable dependiente***

Variable independiente	Nº de clientes	%
No venta	987,454	99.58%
Venta	4,165	0.42%
Total	**991,619**	**100%**

FUENTE: Elaboración propia

En cuanto a las variables predictoras, se contó con alrededor de 40 variables entre demográficas como "Edad" y "localidad"; financieras como "tipo de tarjeta de crédito", "número de entidades financieras en las que tiene un producto", "línea de crédito", "Saldo de TC en los bancos afiliados" y variables propias del producto como "tasa de interés", "plazo a pagar", "monto a ofrecer", "fecha de pago", entre otras.

c. Preparación de los datos

- Se realizó la exploración de los datos para poder detectar datos nulos. Se encontraron 3 variables con un porcentaje máximo de 1.4% de datos perdidos. Por su poca relevancia al modelo se procedió a la depuración de estos datos. Además, se encontró 3 variables con un promedio de 21% de datos perdidos. Estás variables fueron eliminadas; ya que, no discriminaban en función a la efectividad de ventas de la campaña.
- Se encontraron variables con variancia muy cercana a cero, se fueron eliminando. También, se analizó la multicolinealidad entre las variables predictoras, al calcular la correlación entre ellas se vio alta correlación entre el saldo de otros bancos.
- El siguiente paso fue realizar la selección de variables, para ello, se realizó el análisis de forma univariada y bivariada para identificar las variables que estén relacionadas con la variable "Ventas" y también las relaciones entre las variables independientes.
- Para el caso de las variables cualitativas, se realizó la prueba de chi cuadrado para cada uno con la variable dependiente, para saber cuáles podrían ser significativas al modelo. Con esto se obtuvo con alta significancia las variables: "Edad","Segtipotarjeta","SegtipoTasa", entre otras.

- Finalmente se realizó la recategorización de la variable “fecha de pago” en tres niveles en función al evolutivo de efectividad de la campaña (Figura 7).

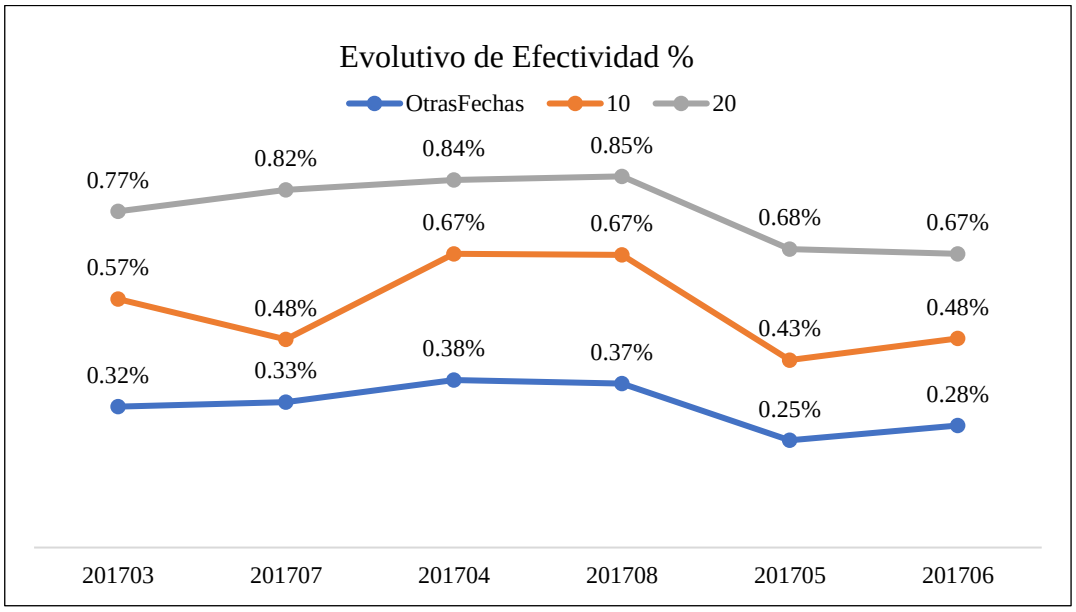

***Figura 7*: Efectividad por periodo y tipo de fecha de pago**

FUENTE: Elaboración propia

Finalmente, luego del procesamiento de los datos, se eligieron para el modelo las variables predictoras que se muestran en la siguiente tabla:

Tabla 2: *Variables predictoras a usar en el modelo*

Variable predictora	Descripción
LineaOferta	Línea a ofrecer en el producto PP
Edad	Edad del cliente en campaña
SegRangoMontos	Rango de Oferta
SegAntiguedad	Meses de antigüedad del cliente en campaña
SegtipoTarjeta	Tipo de tarjeta de crédito en el banco
SegtipoTasa	Tipo de tasa a ofrecer en el producto
SegBancos	Número de bancos en los que el cliente tiene otro producto
SegZona	Segmentación de la zona donde
SegAmbito	Segmentación de que ciudad es el cliente
Propension2	Segmento al que pertenece el cliente
FechaPago2	fechas de pago del producto

FUENTE: Elaboración propia

d. Modelamiento

Se dividieron los datos en 70% para los datos de entrenamiento y el 30% restante para los datos de prueba que ayudaron a evaluar el modelo. La partición de los datos no debe afectar la proporción de las clases en la variable Y, es decir, para ambos conjuntos de datos se seguirá manteniendo la proporción inicial de la variable Y (Figura 8).

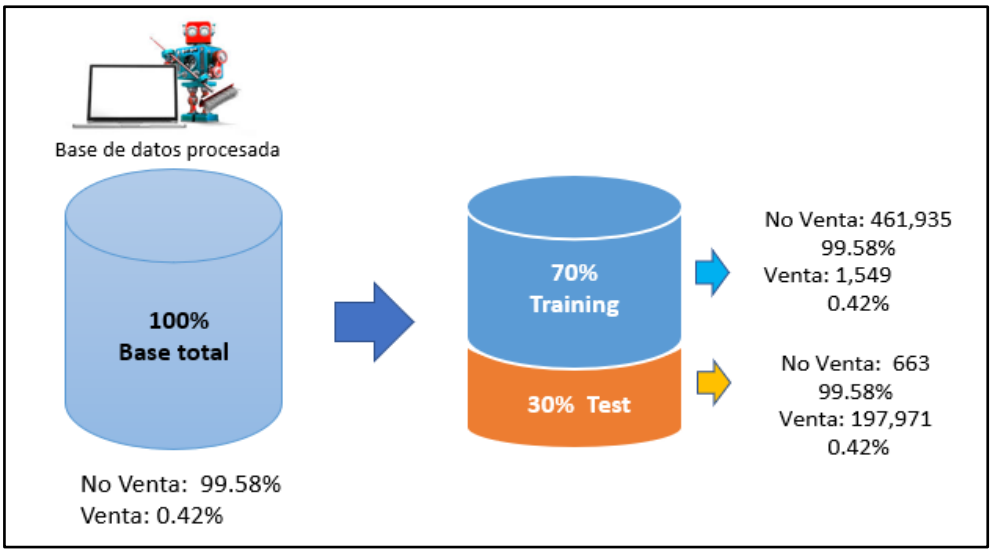

***Figura 8*: Selección de muestra Training y Test**

FUENTE: Elaboración propia

Para la base de datos "Training" se realizaron dos tipos de balanceo: (1) o Undersampling y (2) SMOTE. Con cada una de estas muestras se aplicó el algoritmo Random Forest utilizando validación cruzada k=10 para validar los modelos ejecutados y búsqueda de los hiperparámetros óptimos (Figura 9).

Se esta utilizando k=10, debido a que este valor ha tenido muy buenos resultados en "trade-off", es decir mantiene un equilibrio entre las estimaciones de los errores de sesgo y varianza (Martinez Gil, 2018).

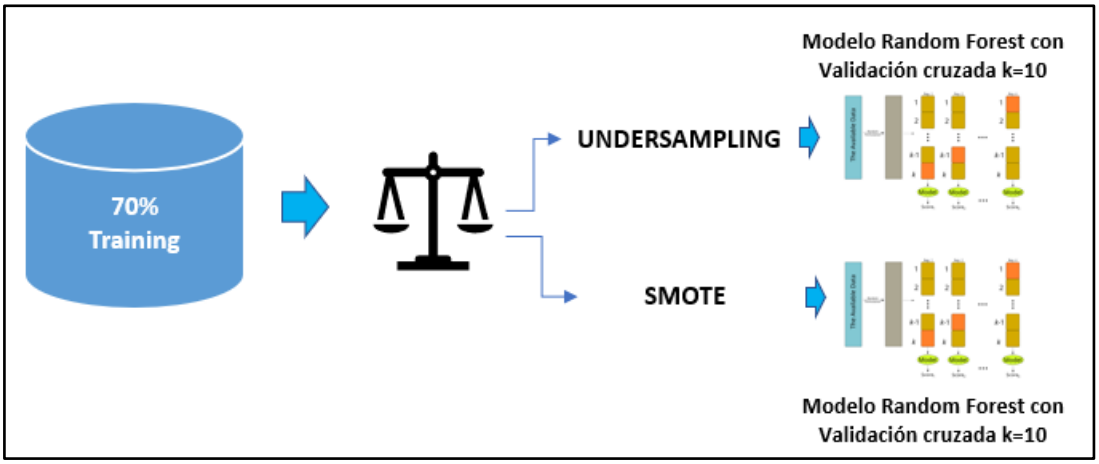

***Figura 9*: Proceso de balanceo de datos y modelamiento**

FUENTE: Elaboración propia

- Algoritmo Random Forest con balanceo Undersampling

  Con los datos ya procesados, se aplicó el algoritmo Random Forest con validación cruzada k=10 con los datos de entrenamiento balanceado mediante el método de Undersampling. Como se puede observar en la (Tabla 3) ahora se tiene un 49% para los datos de la clase "Venta" y 51% para la clase "No Venta".

Tabla 3: *Balanceo de datos utilizando Undersampling*

Venta PP	Clientes	%
Venta	2911	49%
No venta	2986	51%
Total	5897	100%

FUENTE: Elaboración propia

Con estos datos se obtuvo los siguientes resultados en el modelo (Tabla 4).

Tabla 4: *Matriz de confusión Modelo Random Forest Undersampling*

	Predicho		
Real	**No venta**	**Venta**	**Total**
No venta	219,792	72,237	292,029
Venta	354	893	1,247
Total	220,146	73,130	293,276

FUENTE: Elaboración propia

- Algoritmo Random Forest con balanceo SMOTE

  Se aplicó el algoritmo Random Forest con validación cruzada k=10 con los datos de entrenamiento balanceados con el método SMOTE. Como se puede observar en la (Tabla 5) ahora se tiene un 48% para los datos de la clase "Venta" y 52% para la clase "No Venta".

Tabla 5: ***Balanceo de datos utilizando SMOTE***

Balanceo SMOTE		
Venta PP	**Clientes**	**%**
Venta	8733	48%
No venta	9315	52%
Total	18048	100%

FUENTE: Elaboración propia

Con estos datos, los resultados del algoritmo se muestran en la (Tabla 6).

Tabla 6: ***Matriz de confusión con el algoritmo Random Forest – SMOTE***

	Predicho		
Real	**No venta**	**Venta**	**Total**
No venta	244,737	47,292	292,029
Venta	578	669	1,247
Total	245,315	47,961	293,276

FUENTE: Elaboración propia

e. Evaluación del modelo

Para efecto del problema, se usaron la sensibilidad y especificidad como métricas de evaluación de los modelos.

Tabla 7: ***Métricas de evaluación de los modelos***

	RF Undersampling	**RF SMOTE**
Sensibilidad	71%	54%
Especificidad	75%	83%

FUENTE: Elaboración propia

Como muestran los resultados (Tabla 7), el modelo RF que usó la muestra de entrenamiento con datos balanceados con Undersampling obtuvo 71% de sensibilidad, lo que indica que predice mejor las “Ventas” en comparación al segundo modelo, el que usó el método de SMOTE como muestra de entrenamiento, que obtuvo un 54% de sensibilidad. Por el contrario, para la especificidad el segundo modelo obtuvo 83%, por lo que este modelo predice mejor las “No ventas” en comparación con el primer modelo que obtuvo 75% de especificidad.

f. Implementación

Al mes siguiente de terminado el modelo, se espera la nueva base de campaña proporcionada por el banco los últimos días del mes para iniciar su gestión el primer día del mes siguiente con la implementación del modelo.

El proceso a seguir, contempló los siguientes pasos (Figura 10):

- Luego de realizar el procesamiento de la base nueva, se carga la base al servidor de base de datos en SQL server.
- Ejecutar el modelo con la base nueva, agregándole los valores de predicción y segmentándolos por la probabilidad de cada registro con el campo “Rango”.
- Con esos campos nuevos, se genera la plantilla de carga, para almacenar los datos al sistema discador que llamadas.
- Con la base lista, se coordina con el área de operaciones las estrategias a utilizar con los asesores para iniciar la gestión al siguiente día de realizar la carga de base. Teniendo como prioridad, la gestión de los clientes con el mejor segmento (R1).

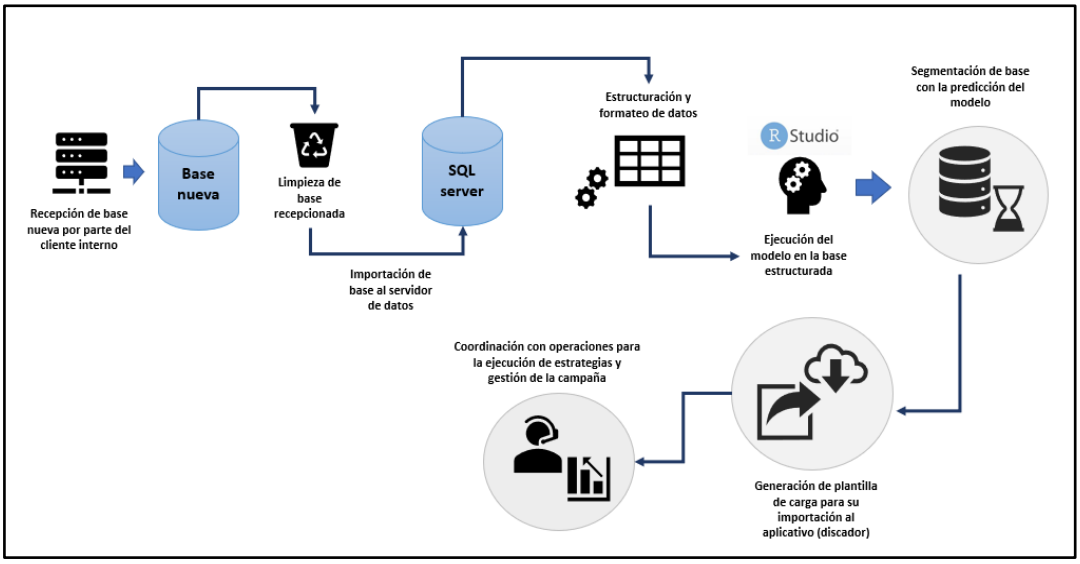

***Figura 10*: Proceso de implementación del modelo**

FUENTE: Elaboración propia

4.3. Contribución en la solución de situaciones problemáticas

El área de operaciones, que se encarga de gestionar las llamadas telefónicas a todos los clientes que tienen en cartera, tiene la necesidad de contar con un modelo predictivo que les permita identificar si el cliente en gestión, que cuenta con al menos una tarjeta de crédito, va a adquirir el producto que le van a ofrecer; ya que, antes de implementar el modelo en la campaña de préstamos, el área no tenía ningún criterio analítico para gestionar la cartera de clientes.

Por esta razón el área de Business Analytics se reunió con el área de operaciones de la campaña y se planteó la elaboración de un modelo de Machine Learning, que pueda predecir a los clientes potenciales a la compra del producto y así priorizar su gestión. El algoritmo Random Forest permite estimar la probabilidad de adquisición del producto por parte del cliente. Se evaluó la efectividad de venta en función a los rangos de probabilidad de adquisición del producto. Se observa que el rango “R1” (probabilidades mayores a 0.7) tienen efectividad de ventas mayores a 1%, el rango “R2” (probabilidades desde 0.7 hasta 0.4) tiene efectividad mayor a 0.48% y el último rango “R3” (probabilidades menores a 0.4) tienen las menores efectividades de venta. Por lo que, se definieron tres segmentos de venta para la gestión de la campaña (Figura 11).

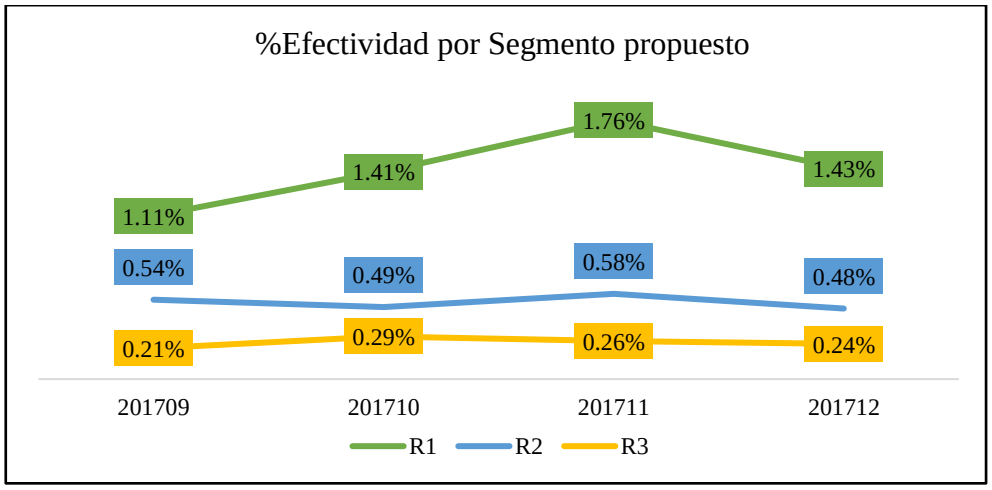

***Figura 11*: Efectividad de venta por segmento propuesto**

FUENTE: Elaboración propia

El resultado del modelo obtuvo un alto indicador de sensibilidad. Esto permitió que se pueda realizar la segmentación de la cartera y se pueda gestionar a los clientes en campaña de clientes de forma óptima al priorizar a los segmentos con mayor indicador de efectividad.

4.4. Análisis de la contribución en términos de competencia y habilidades

La malla curricular de la carrera de Estadística e Informática cuenta con cursos que permiten al estudiante, consolidar gran capacidad de análisis que son fundamentales para la toma de decisiones de las áreas de inteligencia comercial, marketing, banca y muchas más.

En toda empresa es imprescindible el manejo de base de datos, cuya base se proporcionan en los cursos de Técnicas de programación y Base de datos. El análisis de los datos permite responder muchas de las preguntas del negocio por parte de las áreas no especializadas en la explotación de datos. Además, es posible respaldar con las diferentes técnicas estadísticas las hipótesis que puedan surgir en el día a día de la empresa. Los cursos de Análisis de regresión, técnicas multivariadas, análisis de datos categóricos, análisis de series de tiempo, Modelos lineales y Técnicas de muestreo son fundamentales para alcanzar la capacidad analítica requerida por el mercado laboral.

Por otro lado, hoy en día es muy valorado en las empresas que el profesional, además de contar con conocimientos propios de la carrera de estadística e informática, desarrolle

habilidades blancas como el liderazgo, capacidad para entender el negocio, comunicación fluida con las diversas áreas de la empresa y, sobre todo, el resolver situaciones de conflicto con la rápida toma de decisiones. Por ello, es importante incluir en la malla algún curso de "liderazgo" y "metodologías ágiles" que permitan al profesional ser más competitivo y crecer más rápido profesionalmente.

4.5. Nivel de beneficio obtenido por el centro laboral

Luego de implementar el modelo para la gestión de llamadas para la campaña de préstamos se incrementaron las ventas, por ende, se logró llegar a la meta planteada por la operación. Con esto se logró aumentar los ingresos por la campaña, pues no había descuento por penalización del objetivo.

Como se puede observar en la (Figura 12), los montos colocados superan el objetivo y van en incremento desde que se implementó el modelo. Además, en la (Figura 13), se observa que el mejor segmento es el segmento que tiene mejor colocación respecto a los otros.

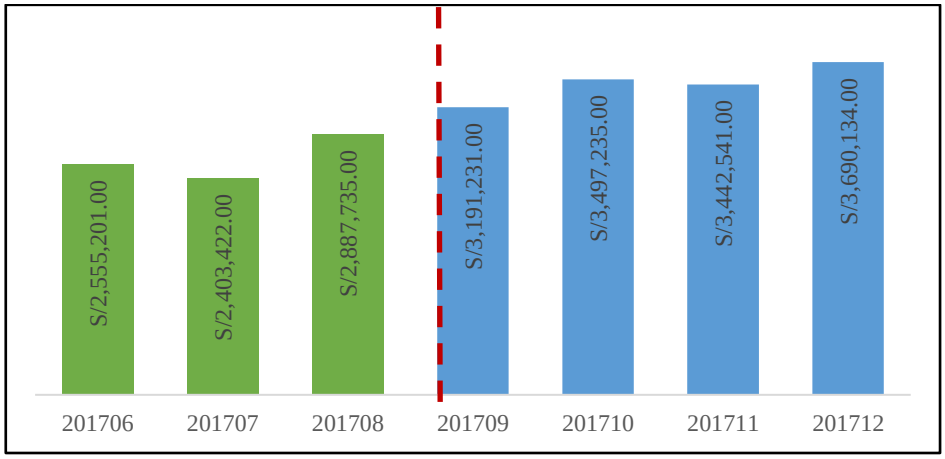

***Figura 12*: Evolutivo de campaña PP luego de la implementación**

FUENTE: Elaboración propia

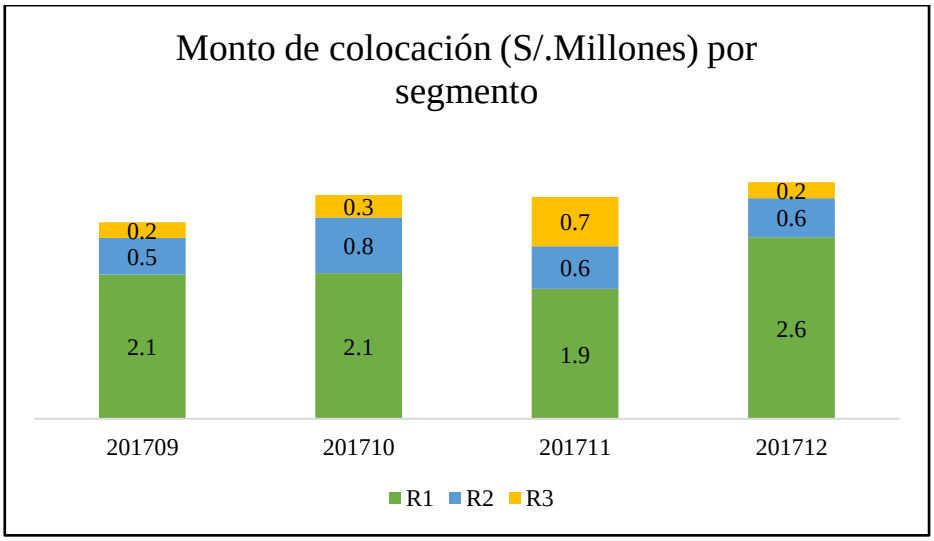

***Figura 13*: Monto de colocación por segmento**

FUENTE: Elaboración propia

V. CONCLUSIONES Y TRABAJO A FUTURO

5.1. Conclusiones

- El desarrollo del modelo de clasificación para predecir si un cliente va a adquirir o rechazar un préstamo personal bancario a través del canal de televentas utilizando el algoritmo Random Forest incrementó las ventas de campaña desde que se comenzó a gestionar las bases con el modelo. Con ello se logró cumplir las metas trazadas mensualmente.

- Con los resultados obtenidos de la evaluación de los modelos, se llegó a tener valores aceptables para la sensibilidad y especificidad para los datos de prueba y los datos globales. Por lo que la capacidad de predicción del modelo implementado garantiza un buen desempeño a futuro.

- Al comparar los dos modelos se obtuvieron mejores resultados con el modelo que utilizó el balanceo de datos Undersampling en comparación al segundo modelo que utilizó el balanceo SMOTE. El primero obtuvo 17% más de sensibilidad, que es el valor de interés que se tiene para el caso, pues la prioridad es predecir si los clientes verdaderamente son clasificados como “Venta”.

- Adicionalmente, se usaron variables que están fuertemente relacionadas con el ratio de efectividad de la campaña, por lo que en términos del negocio esto hace aún más aceptable de cara a su uso en la empresa.

5.2. Trabajo a futuro

- Actualmente hay otros métodos de balanceo que podrían ser probados para obtener mejores resultados de predicción. Además, se podría probar realizar la transformación de algunas variables para evaluar si esto ayudaría a mejorar el rendimiento del modelo.

- Por otro lado, también sería bueno comparar si es que con otras técnicas o algoritmos de Machine Learning es posible obtener mejores resultados de los que se obtuvieron con el algoritmo Random Forest.

VI. REFERENCIAS BIBLIOGRÁFICAS

Analytics Vidhya. (28 de marzo de 2016). *Practical Guide to deal with Imbalanced Classification Problems in R*. Obtenido de Analytics Vidhya Learn about analytics: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

Arnejo Calviño, H. A. (2017). *Métodos para la mejora de predicciones en clases desbalanceadas en el estudio de bajas de clientes (CHURN).* Coriña: Universidad de Coruña.

Arrieta, J., & Mera, C. (2015). *https://www.researchgate.net. bookdown.org/arbol de decisión y random forest*. (s.f.).

Brange, Á. (7 de julio de 2013). *Álvaro Brange's Blog*. Obtenido de http://brange.me/2013/07/07/apuntes-de-machine-learning/

Cárdenas Garro, J. A. (2019). *Clasificación de aceptación de campaña para una entidad financiera, usando random Forest con datos balanceados y datos no balanceados.* Lima, Perú.

Chawla, N. (2002). SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artifcial Intelligence Research 16*, 326-327.

Cichosz, P. (2015). *Data mining algorithms : explained using R* . WILEY.

Fernández Félix, B. M. (2018). *Validación interna de modelos predictivos de regresión logística Comando validation STATA.* Madrid, España: Universidad Complutense de Madrid.

Furnkranz, J., Camberger, D., & Lavrac, N. (2012). Machine learning and data mining. En J. Furnkranz, D. Camberger, & N. Lavrac, *Fundation and rules Learning* (pág. 16).

Hidalgo Ruiz-Capillas, S. (2014). *Random Forests para la detección de fraude en medios de pago.* Madrid, España.

Koehrsen, W. (27 de diciembre de 2017). *Random Forest Simple Explanation*. Obtenido de https://medium.com/: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

Kuhn, & Max. (2010). Variable Selection Using The caret Package. *The journal R* .

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling.* Connecticut: Springer.

Lateef, Z. (2019, Mayo 22). *A Comprehensive Guide To Random Forest In R.* Retrieved from eudereka.co: https://www.edureka.co/blog/random-forest-classifier/

Martinez Gil, C. (Mayo de 2018). *rpubs*. Obtenido de MÉTODOS DE REMUESTREO Y VALIDACIÓN DE MODELOS: VALIDACIÓN CRUZADA Y BOOTSTRAP: https://github.com/CristinaGil/Estadistica_machine_learning_R

Meyer, D. (2014). Support Vector Machine: The Interface to libsvm in package e1071. *FH Technikum Wien, Austria*.

Pariona Huarhiachi, J. C. (2017). *CLASIFICACIÓN DE FUGA DE CLIENTES EN UNA ENTIDAD FINANCIERA UTILIZANDO EL ALGORITMO SMOTE PARA DATOS DESBALANCEADOS EN UNA REGRESIÓN LOGÍSTICA.* Lima, Perú.

Pino Cubiles, P. (2017). *Evaluación del riesgo crediticio mediante árboles de clasificación y bosques aleatorios.* Sevilla, España.

Rikunert. (06 de Noviembre de 2017). *SMOTE explained for noobs - Synthetic Minority Over-sampling TEchnique line by line.* Obtenido de http://rikunert.com: http://rikunert.com/SMOTE_explained

Santana, & Emmanuel. (22 de noviembre de 2014). *Data Mining with R.* Obtenido de Examples of Data mining with R: http://apuntes-r.blogspot.pe/search/label/Validacion%20Cruzada

Torgo, L. (2017). *Data Mining with R learning with cases studies.* Portugal: Taylor & Francis Group.

Vega Alaluna, J. A. (2019). *Modelo de random forest aplicado a ventas cruzadas en un e-commerce de telefonía movil para la predicción de compra o no compra de productos.* Lima, Perú.

VII. ANEXOS

Anexo 1: Código en R para pre-procesamiento, modelado y validación de los datos

```
# Para ver las variables con valores perdidos
which(colSums(is.na(datos))! =0)

datos.r=na.omit(datos)

# Para ver las variables con valores perdidos
which(colSums(is.na(datos.r))!=0)

Sum (is.na(datos.r)) # no hay datos perdidos

# Identificando variables con variancia cero o casi cero

library(caret)
sv <- nearZeroVar(datos.r, saveMetrics= TRUE)
sv
sv[sv$nzv==T,]

#  Identificando predictores correlacionados

descrCor <- cor(datos.r[,c(1:2,4:11,20)])
descrCor

summary(descrCor[upper.tri(descrCor)])
altaCorr <- findCorrelation(descrCor, cutoff = .50, names=TRUE)
altaCorr


#Separar los datos en training y test
library(caret)
set.seed(123)
index <- createDataPartition(data_clean$Venta, p=0.7, list=FALSE)
training <- data_clean[ index, ]  #datos de entrenamiento
testing <-  data_clean[-index, ]  #datos de prueba

# Aplicando el modelo con Validación Cruzada
     library(caret)
modelLookup(model='rf')

library(caret)
ctrl <- trainControl(method="cv", number=10)
```

MODELO UNDERSAMPLING

```
#balanceo de datos con undersampling

train1<- ovun.sample(Venta~., data =training,
method = "under",p = 0.49998, seed = 1994)$data

addmargins(table(train1$Venta))

# modelo undersampling

set.seed(1994)
modelo_train1 <- train(Venta ~ .,
                       data = train1,
                       method = "rf",
                       trControl = ctrl,
                       tuneLength = 11,
                       metric="Accuracy")
modelo_train1

# Predicción de la clase y probabilidad con RANDOM FOREST UNDERSAMPLING
CLASE.RF <- predict(modelo_train1,newdata =testing )
head(CLASE.RF_rose)

PROBA.RF <- predict(modelo_train1,newdata = testing, type="prob")
PROBA.RF <- PROBA.RF[,2]
head(PROBA.RF)


# Evaluando la performance del modelo RANDOM FOREST con UNDERSAMPLING

# Tabla de clasificación


library(gmodels)
CrossTable(x = testing$Venta, y = CLASE.RF,
           prop.t=FALSE, prop.c=FALSE, prop.chisq = FALSE)

#evaluación del modelo
caret:confusionMatrix( CLASE.RF,testing$Venta,positive="Venta")
```

Modelo SMOTE

```
# Balanceo de datos SMOTE

library(DMwR)
set.seed(123)
smote_train <- SMOTE(Venta ~ .,
                     data=training,
                     perc.over = 200,
                     perc.under=160)

addmargins(table(smote_train$Venta))

# modelo SMOTE

 set.seed(1994)
 modelo_SMOTE<- train(Venta ~ . ,
                       data = smote_train,
                       method = "rf",
                       trControl = ctrl,
                      tuneLength = 11,
                       metric="Accuracy")
modelo_SMOTE


# Predicción de la clase y probabilidad con RANDOM FOREST CON SMOTE
```

```
CLASE.RF.smote<- predict(modelo_SMOTE,newdata =testing )
head(CLASE.RF.smote)

PROBA.RF.smote<- predict(modelo_SMOTE,newdata = testing, type="prob")
PROBA.RF.smote<- PROBA.RF.smote[,2]
head(PROBA.RF.smote)
caret::confusionMatrix(CLASE.RF.smote,testing$Venta,positive="Venta")


# Evaluando la performance del modelo RANDOM FOREST con SMOTE

# Tabla de clasificación
library(gmodels)
CrossTable(x = testing$Venta, y = CLASE.RF.smote,
           prop.t=FALSE, prop.c=FALSE, prop.chisq = FALSE)

# Resumen
caret::confusionMatrix(CLASE.RF.smote,testing$Venta,positive="Venta")
```