UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD ECONOMÍA Y PLANIFICACIÓN

"IDENTIFICACIÓN DE CLIENTES QUE REALIZARON FUGA DE EQUIPOS MÓVILES EN UNA EMPRESA DE TELECOMUNICACIONES UTILIZANDO EL ALGORITMO RANDOM FOREST"

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TITULO DE INGENIERO ESTADÍSTICO E INFORMÁTICO

FRANCISCO MARQUEZ MEZA

Lima - Perú

2020

UNIVERSIDAD NACIONAL AGRARIA

LA MOLINA

FACULTAD ECONOMÍA Y PLANIFICACIÓN

“IDENTIFICACIÓN DE CLIENTES QUE REALIZARON FUGA DE EQUIPOS MÓVILES EN UNA EMPRESA DE TELECOMUNICACIONES UTILIZANDO EL ALGORITMO RANDOM FOREST”

PRESENTADO POR

FRANCISCO MARQUEZ MEZA

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL
TITULO DE INGENIERO ESTADÍSTICO E INFORMÁTICO

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO

……..…..…….....	.....…….....
M.A. Fernando René Rosas Villena Presidente	Dr. Jorge Chue Gallardo Asesor
……..…..…….....	.....…..…….....
Mg. Iván Dennys Soto Rodríguez Miembro	Mg. Ana Cecilia Vargas Paredes Miembro

Lima - Perú

2020

Dedicatoria

A Dios por iluminar mi camino y vigilar mis
pasos en cada momento de mi vida, y bendecirme
con las hermosas personas que conozco.

A mis padres por su amor, aliento y ejemplo
a lo largo de toda mi vida,
y por su apoyo incondicional en todos los proyectos
personales y profesionales que me he propuesto.

A mis hermanos
por su compañía, ejemplo y cariño
que me transmiten en cada aventura.

A mis amigos Luis, Elvis, Jorge y Cinthia
por su amistad incondicional e
inspirarme a seguir creciendo cada día.

Agradecimientos

A todos mis profesores que guiaron mi camino en la universidad.

En especial, a la profesora María Ines Nuñez por su cariño, apoyo y amistad, y al profesor Jorge Chue por acompañarme en la realización de este trabajo.

Índice del contenido

1. PRESENTACIÓN 1
2. INTRODUCCIÓN 2
3. OBJETIVOS 4
3.1. Objetivo General 4
3.2 Objetivo Específico 4
4. CUERPO DEL TRABAJO 4
4.1.Funciones Desempeñadas 4
4.2.Puesta en práctica de lo aprendido en la carrera 5
4.3.Contribución en la solución de situaciones problemáticas 30
4.4.Análisis de la contribución en términos de competencia y habilidades 31
4.5.Nivel de beneficio obtenido por el centro laboral 31
5. CONCLUSIONES Y RECOMENDACIONES 32
6. REFERENCIAS BIBLIOGRÁFICAS 34
7. ANEXOS 35

Índice de tablas

Tabla 1: Visualización de predicciones a través de la Matriz de Confusión. 11
Tabla 2: Desempeño del modelo propuesto versus otros algoritmos en el conjunto de datos Orange telecom. 12
Tabla 3: Desempeño del modelo propuesto versus otros algoritmos en el conjunto de datos Cell2Cell. 13
Tabla 4: Conjunto de las 10 variables más importante en la muestra de entrenamiento. 13
Tabla 5: Indicadores de desempeño del modelo propuesto y sus comparaciones. 14
Tabla 6: Variables independientes del conjunto de datos 15
Tabla 7: Distribución de altas nuevas por canal de venta. 17
Tabla 8: Evolutivo de la tasa de fuga del Canal 1. 18
Tabla 9: Evaluación de umbrales para reducir la ventana de monitoreo. 19
Tabla 10: Tasa de fuga en la muestra de modelamiento 21
Tabla 11: Descripción de variables disponibles. 21
Tabla 12: Resultados del análisis univariado de las variables cuantitativas 22
Tabla 13: Matriz de correlaciones entre pares de variables cuantitativas. 25
Tabla 14: Variables seleccionadas para la fase de modelado. 28
Tabla 15: Performance del modelo desarrollado. 28
Tabla 16: Performance del modelo en la muestra testing. 29

Índice de figuras

Figura 1: Esquema del funcionamiento de los árboles de clasificación. 6
Figura 2: Esquema del proceso de funcionamiento del Random Forest. 9
Figura 3: Esquema del monitoreo actual y deseado. 17
Figura 4: Esquema del diseño muestral. 20
Figura 5: División en muestras Training y Testing. 24
Figura 6: Importancia de variables según el criterio del "IV". 27
Figura 7: Beneficios asociados a la implementación de solución planteada. 32

Índice de anexos

Anexo 1: Análisis bivariado de la variable Departamento. 35
Anexo 2: Análisis bivariado de la variables Gama..... 35
Anexo 3: Análisis bivariado de la variables Marca 36
Anexo 4: Limpieza de valores extremos de la variable Precio. 36
Anexo 5: Limpieza de valores extremos de la variable Costo_Equipo 36
Anexo 6: Limpieza de valores extremos de la variable prom_llam_sal. 37
Anexo 7: Categorización de la variable Precio. 37
Anexo 8: Categorización de la variable Costo_Equipo. 37
Anexo 9: Categorización de la variable prom_llam_sal. 38
Anexo 10: Recategorización de la variable Marca. 38
Anexo 11: Variable Gama recategorizada. 38
Anexo 12: Recategorización de la variable Gama. 38
Anexo 13: Variable Gama recategorizada 39
Anexo 14: Código de procesamiento. 39

1. PRESENTACIÓN

Actualmente, el autor del presente trabajo lleva 2 años laborando profesionalmente en la División de Riesgos de una reconocida empresa del rubro bancario. Asimismo, sus principales funciones radican tanto en validar modelos de riesgo crediticio de productos retail como en desarrollar modelos de riesgo crediticio de portafolios non retail. Como parte de su rol de validador, se encarga de realizar una evaluación crítica e independiente de las herramientas predictivo-analíticas desarrolladas por la Unidad de Modelos, así como una revisión integral de la documentación correspondiente, para finalmente brindar una opinión objetiva sobre la recomendación o no del uso de la herramienta. Además de lo anterior, dentro del Gobierno de Modelos, se encarga de gestionar el proceso de seguimiento de las recomendaciones brindadas al culminar los proyectos de validación.

En su experiencia anterior, entre julio del 2016 y mayo del 2017, laboró en una prestigiosa empresa del sector de telecomunicaciones construyendo modelos predictivos que permitan la detección de clientes que realizaron fuga de equipos telefónicos móviles, comprados a través de algún descuento promocional pero que posteriormente eran vendidos en el mercado informal en lugar de generar tráfico telefónico; y así, generando pérdidas millonarias a la empresa. Además de lo anterior, su labor consistió en asegurar la implementación automática de los modelos construidos de la mano con la construcción de un reporte de métricas de performance que se actualizaba en tiempo real conforme se realizaban las predicciones diarias.

Previo a lo anterior, el autor trabajó en una de las mayores empresas del rubro de Call Center del país, en la cual desarrolló proyectos de Minería de Datos que permitieron agregar el componente analítico a la gestión de ventas de las diversas campañas, incrementando las ventas y mejorando los indicadores claves de los negocios como lo son la contactabilidad y efectividad.

En su primera experiencia laboral, tuvo la oportunidad de integrarse a una empresa del rubro de Investigación de Mercados, en la cual pudo participar en todas las etapas del desarrollo de distintos proyectos. De este modo, colaboró en le desarrollo y programación de encuestas, diseño del estudio de mercado, capacitación de encuestadores, supervisión del trabajo de campo, análisis de los datos del estudio, elaboración de informé técnico del estudio y participación en la presentación de resultados ante el cliente.

En resumen, el autor cuenta con más de 5 años de experiencia profesional en el análisis de datos, desarrollando proyectos analítico-predictivos que permiten extraer insights accionables que generen valor para las empresas en las que laboro, y en consecuencia generar mejoras en la forma en como estas operan. Por otro lado, el autor desea compartir el sentimiento de realización y satisfacción profesional y personal que considera ha alcanzado a través de la aplicación de los conocimientos y formación adquiridos en su paso por la Universidad Nacional Agraria La Molina. Asimismo, compartir el deseo de continuar con su formación profesional a través de una Maestría en el campo de la Ciencia de Datos el próximo año (2020).

2. INTRODUCCIÓN

Las empresas del sector de telecomunicaciones se enfrentan a diversos retos al momento de intentar maximizar los beneficios que puedan obtener de sus actividades comerciales. En este contexto, una de las problemáticas que enfrentaba la Empresa de telecomunicaciones en la que el autor laboró era la de identificar a aquellos clientes que, luego de adquirir un equipo móvil cuyo precio fue subvencionado a través de alguna promoción, nunca generaban el tráfico telefónico esperado; y en consecuencia, nunca realizaban consumos que le permitiecen a la Empresa generar beneficios futuros. Es decir, una vez que estas personas compraban los equipos con descuentos, no los utilizaban para realizar actividades telefónicas comúnes de cualquier usuario, sino, revendían los equipos adquiridos a un precio mayor en el mercado informal, impactando negativamente en los beneficios de la Empresa. A dicha actividad la Empresa la denomina como *fuga de equipos móviles*. Asimismo, en adelante a los clientes que realicen *fuga de equipos móviles* se les denominará *fugadores*.

Por otro lado, debido a que la Empresa gestionaba sus fuerzas de venta a través de cuatro canales de ventas diferenciados por sus ubicaciones geográficas, a los cuales denominaremos Canal 1, Canal 2, Canal 3 y Canal 4; la tarea que realizaba la Empresa para identificar a aquellos clientes que hicieron *fuga de equipos móviles* se realizaba de forma independiente en cada uno de los cuatro canales de venta mencionados anteriormente.

En este sentido, las tareas de monitoreo que permitían identificar a los fugadores eran llevadas a cabo por un equipo de trabajo, al cual se le denominará como el “Equipo”, y estaban conformadas por un conjunto de reglas de negocio duras. Dichas reglas fueron determinadas por un grupo de expertos de acuerdo a su conocimiento del negocio y a requerimientos de clientes internos. Asimismo, es importante indicar que el Equipo llevaba

a cabo dos monitoreos en periodos de tiempo distintos posteriores a la fecha en la que los equipos móviles eran adquiridos. El primer monitoreo se realizaba 15 días después de que se efectúe la compra del equipo móvil, mientras que el segundo monitoreo se realizaba a los 45 días posteriores del momento de la adquisición del producto. Posteriormente, una vez que el Equipo realizaba los monitoreos e identificaba a los fugadores, la Empresa realizaba una evaluación de los casos más resaltantes y, luego de ello, llevaba a cabo un proceso de auditoría interna el cual le permitía conocer de forma precisa los factores comerciales asociados a las casuísticas detectadas. Es decir, con dicho proceso de auditoría interna se podían conocer información asociada a los casos de los fugadores como: punto de venta, nombre del vendedor, modelos de equipos móviles, pérdidas asociadas al margen del descuento con el cual fue comprado el equipo telefónico, etc. Una vez conocida la información anterior, la Empresa tomaba acciones comerciales concretas sobre los puntos de venta que permitiesen mitigar el riesgo o futuros casos de *fuga de equipos móviles* o recuperar parte de los beneficios perdidos por dichos casos.

Considerando el contexto antes explicado, la Empresa decidió apostar por una iniciativa con un enfoque analítico-predictivo que permitiese mejorar los procesos de monitoreo, pasando así de un enfoque basado en reglas de negocio duras hacia uno que permitiese detectar a los fugadores mediante el análisis de sus patrones de tráfico telefónico. Asimismo, dicho nuevo enfoque debía permitir identificar a los casos de *fuga de equipos móviles* en periodos de tiempo menores a los que existían en ese momento. Finalmente, el nuevo proceso de monitoreo debería implementarse en el sistema de la empresa de forma automatizada, de modo que la identificación de los *fugadores* sea desarrollo de forma automática. En conjunto, de cumplirse ambos objetivos le permitirían a la Empresa minimizar sus pérdidas como consecuencia de contar con un proceso de monitoreo más robusto que identifica los casos de fugadores de forma anticipada.

De acuerdo a lo expuesto anteriormente, el presente trabajo tiene como objetivo ilustrar el proceso de construcción del modelo predictivo que permitió detectar los casos *fuga de equipos móviles* en la Empresa, de forma anticipada al monitoreo existente, y en concreto para el canal de venta Canal 1 y para la ventana de monitoreo de 45 días. Lo anterior, mediante la utilización del algoritmo de aprendizaje automático denominado Random Forest.

3. OBJETIVOS

3.1. Objetivo General

Diseñar e implementar un proceso de monitoreo automatizado que permita identificar, de forma anticipada al monitoreo actual, a los clientes que hicieron *fuga* de equipos móviles en una empresa de telecomunicaciones del Perú utilizando el algoritmo Random Forest.

3.2 Objetivo Específico

- Reducir la cantidad de días del monitoreo actual de 45 días de forma que no se pierda información relevante sobre el tráfico telefónico de los clientes al momento de identificar a aquellos que hicieron *fuga*.
- Identificar las principales variables que permiten la identificación de los clientes que hicieron *fuga* de equipos móviles en la nueva ventana de monitoreo.

4. CUERPO DEL TRABAJO

4.1. Funciones Desempeñadas

Durante su estancia en la empresa de telecomunicaciones, el autor desarrolló principalmente las siguientes actividades:

a. Desarrollo y documentación de modelos predictivos que permitan identificar a los clientes más propensos a haber realizado *fuga de equipos móviles*: Respecto a esta función, se lograron desarrollar ocho modelos predictivos que identificaban a los clientes más propensos a haber realizado *fuga de equipos móviles*, uno para cada canal (4 canales) y por cada ventana de monitoreo (15 y 45 días). Cabe indicar que los modelos fueron desarrollados utilizando el software R y siguiendo cada una de las fases sugeridas por la metodología CRISP. Asimismo, cada modelo fue debidamente documentado haciendo uso de la herramienta R Markdown.

b. Implementación automatizada de los modelos predictivos construidos: se asumió la responsabilidad de garantizar que los modelos predictivos sean implementados de forma automatizada y libres de errores materiales. Esta tarea fue desarrollada a través de la integración de los softwares R y SQL Server Intregation Services (SSIS), y del uso de procesos batcheros (archivos .bat). En el software R se programó el código

que puntuaba a cada cliente (que obtenía su probabilidad de fuga). Luego, se crearon archivos .bat que permitiesen compilar los script de puntuación de R de forma automática en la consola de comandos de Window (CMD). Finalmente, con el software SSIS se crearon ETLs que permitían calcular las covariables requeridas por el modelo y ejecutar de forma periódica (diariamente) el archivo .bat.

c. Implementación de indicadores y monitoreo de resultados de las predicciones dadas por los modelos predictivos desarrollados: Respecto a esta función, se definieron cuatro indicadores para monitorear los resultados de las predicciones diarias de los modelos. Dichos indicadores fueron: Tasa de Correcta Clasificación (TCC), Sensibilidad y el Valor Predictivo Positivo (VPP). Asimismo, se implementó un reporte de seguimiento que permitía visualizar los indicadores antes mencionados, y otros más. Este reporte se actualizaba en tiempo real al ser abierto. Dicho reporte fue construido medianto la integración de las herramientas Microsoft SQL Server Managment Studio y Microsoft Excel.
d. Gestión y coordinación de los proyectos de Analítica y BI con clientes externos: se asumió la resaponsabilidad directa de elaborar presentaciones de alto impacto exponiendo los proyectos de analítica predictiva desarrollados, coordinar con otras unidades de la compañía, y finalmente captar clientes externos para brindarles soluciones basadas en tecnologías de Analítica y BI.

4.2. Puesta en práctica de lo aprendido en la carrera

Como se indicó en secciones anteriores, la problemática existente en la Empresa de telecomunicaciones en la que laboré radicaba en cambiar la forma tradicional que se utilizaba para identificar a aquellos clientes que hicieron *fiuga de equipos móviles*. No obstante, dicho cambio tendría que ser anticipado respecto a los monitoreos de 15 y 45 días que en ese entonces estaban implementados y ya realizaban dicha identificación.

De acuerdo a lo anterior, a continuación se expondrá el proceso seguido para el desarrollo de la alternativa de solución propuesta para afrontar la problemática identificada, descrito a través de las tres próximas subsecciones.

4.2.1 Descripción de las técnicas estadísticas

Para poder comprender mejor los aspectos teóricos del algoritmo Random Forest es necesario conocer el funcionamiento de los algoritmos de árboles de decisión. Esto debido a que los árboles de decisión son una de las componentes que utiliza el algoritmo Random Forest dentro de su proceso de funcionamiento. De este modo, a continuación se explicarán los aspectos generales de los árboles de decisión, y posteriormente la metodología del Random Forest.

Árboles de decisión:

Dentro del cojunto de técnicas para resolver problemas de clasificación y regresión se encuentran las familias de algoritmos basados en árboles. No obstante, el presente trabajo se centrará en el uso de estas técnicas para problemas de clasificación (árboles de clasificación). Esta familia de algoritmos pertenecen al campo del Aprendizaje de Máquinas *(Machine Learning)*, y su funcionamiento de basa principalmente en dividir el espacio de las variables independientes en regiones distintas y no superpuestas, para lo cual examina de forma recursiva cada una de las variables predictoras y elije aquellas particiones que permitan separar mejor a las clases de la variable dependiente. Asimismo, cabe indicar que las variables independientes pueden ser cualitativas o cuantitativas. A continuación la Figura 1 muestra en un esquema desarrollado por (Orellana Alvear, 2019) que explica el funcionamiento del mecanismo utilizado por los árboles de decisión.

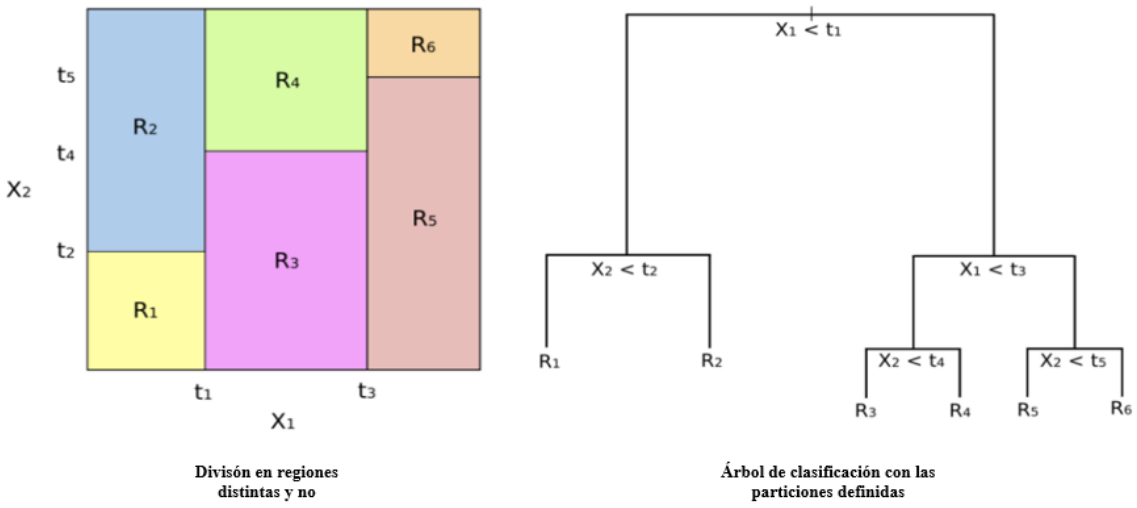


Figura 1: Esquema del funcionamiento de los árboles de clasificación.

Fuente: Johanna Orellana Alvear - bookdown.org/content/2031/

Por otro lado, los árboles de clasificación tienen un elemento denominado "nodo" que representa aquellas preguntas formuladas por el árbol antes de llevar a cabo cada partición. Así, se distinguen los siguientes tipos de nodos:

Nodo raíz: Representa a todo el conjunto de datos (muestra de entrenamiento) que luego será dividida en múltiples particiones o subconjuntos. En la Figura 1 es el punto asociado a la partición "X1 < t1".

Nodo de decisión o padre: Es aquel nodo que se divide en 2 o más nodos. En la Figura 1 son los puntos asociados a las particiones "X1<t1", "X2<t2", "X1<t3", "X2<t4" y "X2<t5".

Nodo terminal: Es aquel nodo que ya no se divide en más nodos o subconjuntos. En la Figura 1 son los puntos representados por "R1", "R2", "R3", "R4", "R5" y "R6".

Adicionalmente, es importante entender que los árboles de clasificación hacen uso de dos indicadores denominados Entropía y Ganancia de Información para decidir si un nodo será divido o no, de forma que se garantice que las observaciones de un mismo nodo sean lo más parecidas entre sí, y lo más diferentes a las de otros nodos.

Entropía: Es un indicador que mide el grado de homogeneidad entre los elementos de un mismo nodo y fluctúa entre 0 y 1. Donde una entropía de 0 indica que los elementos del nodo son muy homogéneos (nodo muy puro), mientras que un valor de 1 indica que los elementos del nodo son muy heterogéneos entre sí.

Ganancia de información: Es un indicador que mide la disminución de la entropía al realizar una nueva partición, lo cual sirve de apoyo para decidir si es conveniente o no seguir dividiendo un nodo.

Finalmente, entre las principales ventajas de los árboles de clasificación encontramos que son fáciles de entender (intuitivos), son robustos ante la presencia de valores extremos (outliers) y datos faltantes, y no requieren suposiciones sobre la distribución de la variable dependiente debido a que son una técnica no paramétrica. Asimismo, entre las principales desventajas tenemos se tiene que: los árboles pueden sufrir de problemas de sobreajuste si no son adecuadamente delimitados (lo cual se logra a través de un concepto denominado poda); hay una pérdida de información si las variables independientes de naturaleza contínua son categorizadas; e inestabilidad asociada a que pequeños cambios en el conjunto de datos puede cambiar la estructura del árbol definido.

Random Forest:

A pesar de la potencia de los árboles de decisión y sus ventajas, el algoritmo Random Forest nace como una alternativa para reducir la tasa de error y los problemas de inestabilidad de los árboles. Random Forest es un método de ensamble que combina predictores de tipo árboles de decisión de modo que se logre alcanzar una mejor precisión y estabilidad en las estimaciones, también conocido como Bosque Aleatorio. Así, entiéndase que múltiples modelos "débiles" basados en árboles de decisión se combinan generando un modelo más robusto. Similar a los árboles decisión, la técnica de Random Forest puede ser utilizada para resolver problemas de clasificación y regresión. La presente explicación se enfocará en el problema de clasificación, le cual hace uso de los árboles de clasificación anteriormente explicados.

Para empezar, es necesario conocer la técnica de ensamble denominada Bagging (Bootstrap aggregation) la cual permite reducir la varianza de las predicciones de un modelo a través de la combinación de varios clasificadores, cada uno de ellos modelados a través de diferentes subconjuntos de observaciones. A continuación se detalla el proceso seguido para construir cada uno de los árboles de clasificación que son combinados por el proceso Bagging:

a. Dado el conjunto de datos con N elementos, se extraen B muestras aleatorias con reemplazado de tamaño N.
b. Dadas M variables independientes en el conjunto de datos original, se define un número de variables m que serán extraídas aleatoriamente del total M para cada una de las B muestras.

A continuación, la Figura 2 muestra un esquema desarrollado por (Kashyap, 2019) que muestra el proceso interno desarrollado por el algoritmo Random Forest a través del cual logra obtener sus predicciones.

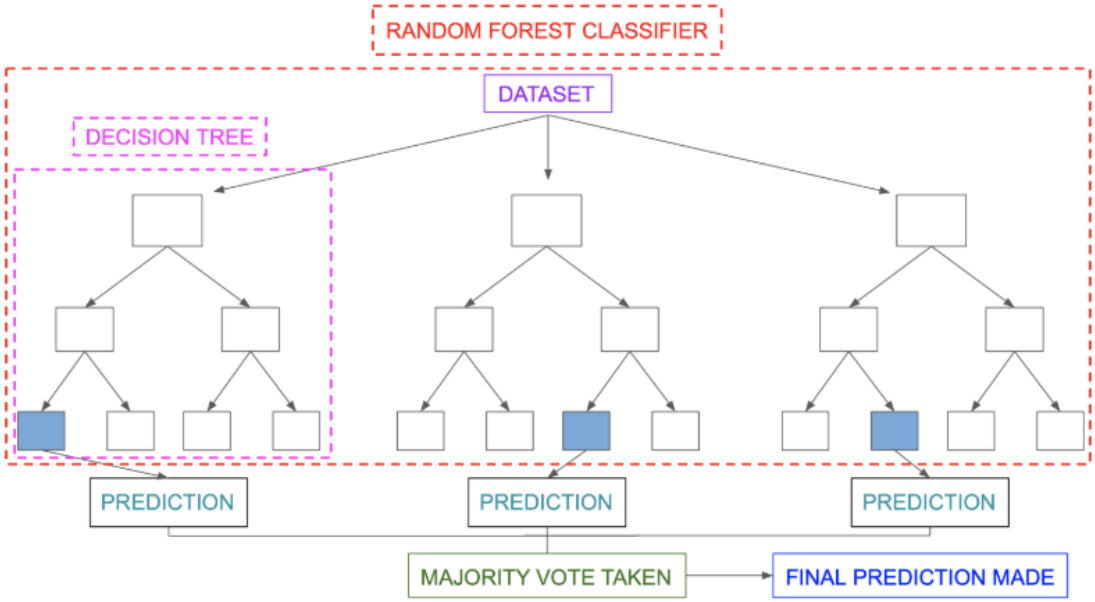


Figura 2: Esquema del proceso de funcionamiento del Random Forest.

Fuente: Karan Kashyap. https://medium.com/analytics-vidhya/

Como se puede ver, en este esquema el número de árboles construidos es 3 (B), y en cada uno de ellos el conjunto de datos corresponde a una muestra aleatoria extraída con reemplazo del conjunto de datos original. Asimismo, se puede apreciar que la moda (el mayor voto o predicción más votada) de la predicción de cada uno de los árboles es igual a la predicción final del Random Forest.

Otro concepto importante, explicado por (Breiman, 2001), y que utiliza el algoritmo Random Forest es el de Out of Bag (OOB), el cual hace referencia al conjunto de observaciones que no fueron consideradas en cada uno de los B árboles de clasificación generados por el proceso Bagging. Asimismo, el conjunto de observaciones del conjunto OOB de cada árbol es utilizado para calcular el error de generalización del modelo.

Respecto a los parámetros que el Random Forest requiere se especifiquen para su funcionamiento, se tiene que los principales son:

ntree: Número de árboles que se generarán en el proceso bagging. Se espera que a mayor número de árboles generados los resultados serán más estables; no obstante, un número muy grande de este parámetro puede generar que el proceso computacional tenga un coste demasiado alto.

mtry: Número de variables independientes (seleccionadas aleatoriamente) candidatas en cada partición.

sampsize: Tamaño de la muestra aleatoria con la que se entrenará cada árbol de clasificación. El valor por defecto es 63.25%.

nodesize: Mínimo número de observaciones en los nodos terminales. El valor por defecto para problemas de clasificación es de 1.

maxnodes: Número máximo de nodos terminales que los árboles del bosque pueden tener. Si no se especifica un valor los árboles crecerán hasta lo máximo posible.

Un aspecto relevante en cuanto al algoritmo Random Forest es la forma en la que calcula la importancia de las variables. Debido a que el algoritmo genera múltiples árboles que son entrenados con un subconjunto de atributos diferentes (aleatorios), ello conlleva a que cada árbol esté poco relacionado con otro, lo cual aumenta la precisión global debido a que cada árbol comete un error distinto. Los indicadores utilizados por el Random Forest para medir la importancia de las variables son Mean Decrease Gini (MDG) y Mean Decrease Accuracy (MDA). El MDG es la Ganancia de Información de cada variable en cada árbol del bosque promediada. Mientras, el MDA mide el impacto de cada atributo en la precisión del modelo, para lo cual se permutan los valores de cada atributo y se mide la disminución del modelo debido a dicha permutación; así, si el atributo es importante se espera que la permutación de sus valores impacte negativamente y de forma significativa en la precisión del modelo.

Finalmente, entre las ventajas del Random Forest se tiene que es muy eficiente manejando conjuntos de datos con cantidades ingentes de variables independientes e identificar a las más significativas, hace uso de métodos efectivos para estimar los valores faltantes, reduce la estabilidad de las predicciones debido al enfoque ensamblador (bagging). Entre las principales desventajas se tiene que existe una pérdida en la interpretabilidad y que el algoritmo tiende a evidenciar problemas de sobreajuste cuando el conjunto de datos incluye mucho ruido.

Indicadores de evaluación de performance:

Considerando un problema de clasificación en donde el evento de interés (éxito) es la "compra" de un determinado producto, y asumiendo que se tiene un modelo predictivo para predecir dicho evento, es posible medir el desempeño de las predicciones de dicho modelo a través de los siguientes indicadores.

Matriz de confusión: Permite visualizar el desempeño del algoritmo de clasificación mediante el contraste de los valores predecidos y observados.

Tabla 1: Visualización de predicciones a través de la Matriz de Confusión

Predicción	Valor real	
	No compra	**Compra**
No compra	Verdaderos negativos	Falsos negativos
Compra	Falsos positivos	Verdaderos positivos

Fuente: Elaboración propia.

De acuerdo a la Figura 3, los Verdaderos positivos (VP) representan a las compras reales que fueron predecidas correctamente, y los Falsos negativos (FN) represetan a las compras reales que fueron clasificadas incorrectamente. De forma similar, los Verdaderos negativos (VN) y Falsos positivos (FP) representan lo mismo que el VP y FN, respectivamente, pero para la clase de "No compra".

Sensibilidad: Es la proporción de los casos que fueron predecidos como "Compra" entre el total de casos que realmente compraron.

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

Valor predictivo posiivo (VPP): Es la proporción de los casos que fueron predecidos como "compra" entre el total de casos predecidos como compra.

$$\boldsymbol{VPP = \frac{VP}{VP + FP}}$$

Especificidad: Es la proporción de los casos que fueron predecidos como "no compra" entre el total de casos que realmente no compraron.

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

Valor predictivo negativo (VPN): Es la proporción de los casos que fueron predecidos como "no compra" entre el total de casos predecidos como "no compra".

$$\boldsymbol{VPN = \frac{VN}{VN + FN}}$$

Tasa de correcta clasificación (TCC): Es la proporción de los casos que fueron predecidos correctamente predecidos como "compra" o "no compra" entre el total de casos.

$$\boldsymbol{TCC = \frac{VP + VN}{VP + VN + FP + FN}}$$

4.2.2 Revisión de artículos científicos

En (Idris, Iftikhar, & ur Rehman, 2019), se integraron las capacidades de búsqueda de la Programación Genética (GP) y las capacidades del algoritmo de aprendizaje automático AdaBoost con el fin de desarrollar un sistema de predicción de fuga de alto rendimiento para la identificación de clientes fugadores en una empresa de telecomunicaciones. Asimismo, el desarrollo del sistema de identificación de fugadores fue motivado debido a que la feroz competencia entre empresas del rubro generó la necesidad de contar con una sistema que indentifique de manera eficiente a los clientes que potencialmente vayan a prescindir de los servicios de la empresa, es decir, los clientes que van a fugar. Además de lo anterior, el sistema desarrollado tiene por objetivo interpretar el comportamiento de fuga de los clientes. Por otro lado, se detectó que los conjuntos de datos de la empresa de telecomunicaciones estaba desbalanceado, lo cual fue tratado medianto el método de submuestreo basado en el algoritmo Particle Swarm Optimization (PSO, por sus siglas en inglés). El método de submuestreo basado en PSO en combinación con GP-AdaBoost da como resultado un sistema de predicción de la fuga (GP-GPAB), que ofrece un mejor aprendizaje de los fugadores y también identifica los factores subyacentes responsables del comportamiento de fuga de los clientes. Los conjuntos de datos para evaluar y comparar el sistema propuesto fueron Orange telcom y Cell2Cell y están compuestos por 50,000 y 40,000 registros, e incluían 260 y 76 variables respectivamente. También se observó que el conjunto de datos Orange está desbalanceado (7.3% de fugadores), mientras el Cell2Cell sí está balanceado. Para resolver el desbalanceo de la clase minoritaria en los datos de Orange telecom se utilizó la técnica PSO, con la cual se logró un equilibro de las clases en el conjunto de datos en mención (50% no fugadores y 50% fugadores). Una vez construido el modelo predictivo, los resultados en obtenidos en los conjuntos de datos fueron los siguientes:

Tabla 2: Desempeño del modelo propuesto versus otros algoritmos en el conjunto de datos Orange telecom

	Sensitivity	Specificity	AUC
Random forest	0.0049	0.9991	0.571
Rotation forest	0.0026	0.9998	0.583
RotBoost	0.0291	0.7212	0.601
GP-AdaBoost	0.3120	0.7501	0.631

Fuente: Idris, Iftikhar, & ur Rehman, 2019

Tabla 3: Desempeño del modelo propuesto versus otros algoritmos en el conjunto de datos Cell2Cell

	Sensitivity	Specificity	AUC
Random forest	0.690	0.601	0.592
Rotation forest	0.646	0.666	0.610
RotBoost	0.664	0.632	0.699
GP-AdaBoost	0.87	0.891	0.910

Fuente: Idris, Iftikhar, & ur Rehman, 2019

Como se observa en las Tablas 2 y 3, el sistema propuesto (ChP-GPAB) muestra un desempeño superior al de otros algoritmos, consiguiendo producir un AUC de 0.631 y 0.910 en los conjuntos de datos Orange telecom y Cell2Cell respectivamente.

En el trabajo de (Yildiz & Albayrak, 2015) se propone un modelo para estimar a los fugadores de una empresa de telecomunicaciones, mostrando cómo se incrementa la eficacia de la predicción al balancear los datos con el submuestreo y la clasificación por el método de rotación forestal. Además, compara el rendmiento de la técnica indicada con Antminer y el árbol de decisión C4.5. Las comparaciones se realizan utilizando el conjunto de datos de American Telecommunication Company y se utilizan la precisión, sensibilidad y especificidad de los criterios de rendimiento. El conjunto de datos utilizado consta de información de 5,000 clientes, 21 variables independientes y no faltan datos (no valores perdidos). A continuación la Table 4 muestra a las 10 mejores variables de acuerdo al criterio de Ganancia de información.

Tabla 4: Conjunto de las 10 variables más importante en la muestra de entrenamiento

Feature Name	Feature Description	Value
international_plan	International call usage	Yes/No
total_day_minutes	Daily total talk time	Minutes
number_customer_service_calls	Number of call to customer service	
voice_mail_plan	Voice mail usage	Yes/No
total_eve_minutes	Total talk time in evening	Minutes
state	Living place	
total_day_charge	Daily Total spent credits	
number_vmail_messages	Number of voice messages	
total_intl_calls	Total number of international call	
total_intl_charge	Total spent credits of international calls	

Fuente: Yildiz & Albayrak, 2015

El algoritmo de Rotation Forest comenzó a ser utilizado en la literatura en los últimos años y fue presentado como la nueva generación de algoritmos de aprendizaje se basa en la

formación de un conjunto clasificador mediante el uso de análisis de componentes principales, que es una técnica de extracción de características. El principio básico de funcionamiento del algoritmo de Rotation Forest es similar al del Random Forest y se utilizan más de un árbol. Sin embargo, el conjunto de datos que se utiliza en la capacitación de cada árbol de decisión en el bosque se determina por el análisis de los componentes principales.

A continuación, se presentan los indicadores de Tasa de Correcta Clasificación (TCC), Sensibilidad y Especificidad del modelo planteado por (Yildiz & Albayrak, 2015) basado en el Rotation Forest y los algoritmos AntMiner+ y C4.5 (árbol de decisión).

Tabla 5: Indicadores de desempeño del modelo propuesto y sus comparaciones

		Tasa Correcta Clasificación (%)	Sensibilidad (%)	Especificidad (%)
Conjunto de datos original	Rotation Forest	95.68	73.4	99.49
	AntMiner+ [6]	90.85	37.09	99.71
	C4.5 [6]	93.59	64.93	98.34
Down Sampling	Rotation Forest (Subsets Average)	92.49	84.57	96.46
Oversampling	AntMiner+ [6]	93.15	65.76	97.72
	C4.5 [6]	91.66	80.82	93.45

Fuente: Yildiz & Albayrak, 2015

De acuerdo con los resultados anteriores, Rotation Forest es mejor que el árbol de decisión C4.5 y Antminer+ porque el aumento de la predicción real de la tasa de fuga de clientes es más importante. La diferencia entre los algoritmos de Rotation Forest y Antminer+ es del 36,31% en el conjunto de datos original para la tasa de sensibilidad. Los datos de balance se incrementan en todas las tasas de sensibilidad. De acuerdo con estos resultados, el método de Rotation Forest es el mejor algoritmo y un 18,81% más exitoso que Antminer+ en términos de sensibilidad.

(Jadhav & Pawar, 2011) desarrollaron un sistema de apoyo a la toma de decisiones utilizando tecnología de minería de datos para la predicción de fuga en una compañía de telecomunicaciones BSNL Satara (India). Además, el modelo propuesto es capaz de predecir el comportamiento de fuga de los clientes con bastante antelación, lo cual es sumamente valioso para la compañía debido a que puede tomar acciones comerciales para lograr retener a los clientes que presenten un comportamiento asociado a una posible fuga. A continuación, se presentan las variables utilizadas para el desarrollo del trabajo agrupadas de acuerdo al tipo de fuente de datos de la cual provienen.

Tabla 6: Variables independientes del conjunto de datos

Variable	Descripción	Fuente
Late_pay	Cantidad de cuentas atrasadas	Facturación
Extra_charges	Cuentas con cargas adicionales	Facturación
Max_dur	Máxima duración total de llamadas	Tráfico telefónico
Min_dur	Mínima duración total de llamadas	Tráfico telefónico
Max_count	Máximo número de llamadas	Tráfico telefónico
Min_count	Mínimo número de llamadas	Tráfico telefónico
Max_dif	N° máx. de números telefónicos llamados por semana	Tráfico telefónico
Min_dif	N° mín. de números telefónicos llamados por semana	Tráfico telefónico

Fuente: (Jadhav & Pawar, 2011)

Como se indica, se consideraron variables asociadas al tráfico telefónico de los clientes, así como otras relacionadas a fuents comerciales, puntualmente, información de pagos. Además, durante el desarrollo se indicó que se excluyeron aquellas observaciones asociadas a valores perdidos en las variables de tráfico telefónico. También se indicó que para evitar sesgos por la naturaleza o tipo de cliente, estos se extrajeron de la siguiente manera: 3.9% pertenecían al sector público, 82.1% al sector privado, y el 14% restante eran clientes asociados a una empresa. Por otro lado, se indicó el 17.7% del total de clientes estaban estaban asociados a un comportamiento de fuga y el 82.3% no; con lo cual se conformó la variable objetivo. Finalmente se indicó que se utilizó el algoritmo de clasificación de Redes Neuronales, implementado a través del software METALAB. Asimismo, se indicó que el conjunto de datos fue dividido en tres subconjuntos de desarrollo (modelamiento), validación (para monitorear el error durante el proceso de modelamiento) y test, el cual fue utilizado para verificar el performance del modelo construido.

4.2.3 Propuesta de alternativa de solución a la situación problemática

En esta subsección se expondrá el proceso seguido para el desarrollo de la alternativa de solución propuesta para resolver la problemática de la Empresa. Asimismo, es importante indicar que, debido a que la alternativa de solución fue desarrollada siguiendo los estándares que de la metodología CRISP, las explicaciones serán presentadas a través de las seis etapas que dicha metodología contempla. No obstante, antes de ahondar en la explicación de la alternativa de solución propuesta, se presentará la definición de la problemática identificada en la Empresa y la alternativa de solución que ya existía.

Definición de la problemática:

La Empresa de telecomuniciones tiene la necesidad de reemplazar los monitoreos existentes, que se basan en reglas duras para identificar a los clientes *fugadores*, por unos nuevos que se basen en el análisis de los patrones de tráfico telefónico, y a su vez permitan reducir el tiempo de la detección.

Solución existente:

La Empresa ya cuenta con procesos de monitoreo que identifican a los clientes *fugadores*, no obstante, estos monitoreos se basan en la validación de un conjunto de reglas duras, las cuales fueron determinadas por un grupo de expertos de acuerdo a su conocimiento del negocio y a requerimientos de clientes internos. Cabe indicar que los monitores en mención se llevan a cabo en ventanas de 15 y 45 días luego de realizada el alta del equipo, y de forma independiente para cada uno de los cuatro canales de venta (Canal 1, Canal 2, Canal 3 y Canal 4).

a) Conocimiento del negocio

La Empresa de telecomunicacones cuenta con una gerencia que tiene a su disposición un equipo de profesionales dedicados a identificar a aquellos clientes que, luego de adquirir un equipo móvil cuyo precio fue subvencionado a través de alguna promoción, nunca generaban el tráfico esperado; y en consecuencia, nunca realizaban consumos que le permitiecen a la Empresa generar beneficios futuros. En vez de ello, dichos clientes revendían los equipos adquiridos a un precio mayor en el mercado informal, impactando negativamente en los beneficios de la Empresa. A dicha actividad la Empresa la denominaba como *fuga de equipos móviles*. Asimismo, en adelante a los clientes que realicen *fuga de equipos móviles* se les denominará fugadores.

Para cumplir con la tarea mencionada, el equipo referido realiza monitoreos en dos momentos diferentes. El primer monitoreo se llevaba a cabo 15 días posteriores a la fecha del alta del nuevo equipo, mientras que el segundo se realizaba 45 días después de la fecha del alta. Asimismo, dichos monitoreos se realizaban de forma independiente para cada uno de los cuatro canales de venta de la Empresa. A continuación se muestra el porcentaje de ventas que concentra cada uno de los cuatro canales de ventas.

Tabla 7: Distribución de altas nuevas por canal de venta

Canal de venta	Ventas (%)
Canal 1	6
Canal 2	60
Canal 3	24
Canal 4	10

Fuente: Elaboración propia.

Como se puede ver, la distribución de las altas nuevas no es homogénea entre los canales de venta existentes. Por el contrario, se observa que el Canal 2 concentra la mayor cantidad de ventas (60%), lo cual se debe a que dicho canal contempla la mayor cantidad de puntos de venta de la Empresa. Asimismo, se aprecia que el Canal 3 y Canal 4 contienen el 24% y 10% de las ventas respectivamente. Finalmente, se puede observar que el Canal 1 solo concentra un 6% del total de altas nuevas.

Delimitación del alcance del presente trabajo

Por factores de limitación de información y confidencialidad de datos, el presente trabajo se enfocará en describir el proceso de desarrollo de la alternativa de solución para el Canal 1 y para el monitoreo de 45 días. Dicho esto, a continuación se muestra un esquema del proceso de monitoreo de 45 días que se realizaba al Canal 1 (Monitoreo Tradicional) y del proceso de monitoreo que se espera lograr con la alternativa de solución planteada (Monitoreo con Modelo Predictivo).

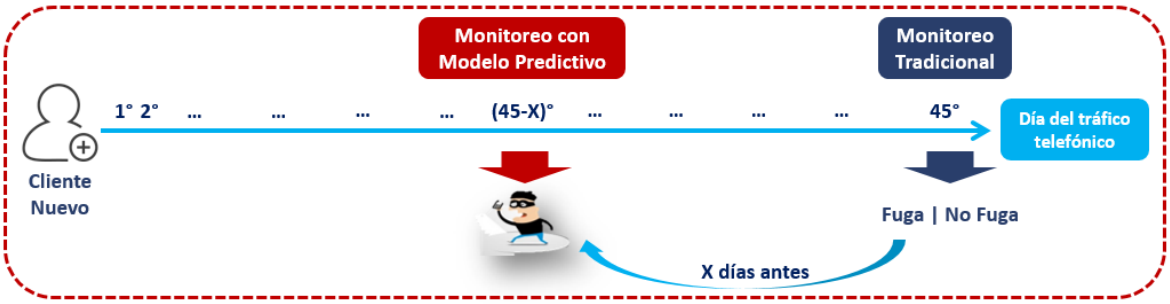


Figura 3: Esquema del monitoreo actual y deseado**.**

Fuente: Elaboración propia.

Como se puede ver, el monitoreo tradicional permite identificar a los clientes que *fugaron* al momento de adquirir el equipo telefónico 45 días posteriores a la fecha en que adquirieron el equipo. Asimismo, se observa un prototipo de lo que se espera del

monitoreo cuando la alternativa de solución sea implementada. Es decir, la identificación de los clientes *fugadores* de forma anticipada al monitoreo tradicional; de acuerdo al esquema, la anticipación será de X días antes de los 45 días.

Por otro lado, es importante conocer cual es el evolutivo del indicador de *fuga* del canal de ventas a analizar. De acuerdo a esto, a continuación se presenta el evolutivo del ratio de *fuga* del Canal 1 de los periodos comprendidos entre dic-2015 y may-2016.

Tabla 8: Evolutivo de la tasa de fuga del Canal 1

Periodo	dic-15	ene-16	feb-16	mar-16	abr-16	may-16
Ratio de Fuga	35.9%	39.2%	39.9%	37.4%	32.7%	33.9%

Fuente: Elaboración propia.

Como se puede ver, el ratio de fuga del Canal 1 osciló entre 32.7% y 39.9% en los periodos mostrados. De estos resultados se puede concluir que el ratio de fuga ha sido relativamente estable en los meses vistos. Asimismo, es importante indicar que aunque el ratio de fuga del Canal 1 es relativamente alto, este canal concentra únicamente el 6% del total de altas nuevas de la Empresa en los periodos mencionados.

Una vez que se lleva a cabo el monitoreo de 45 días y se identifican a los clientes que hicieron *fuga* al momento de adquirir sus equipos móviles, la Empresa realiza una evaluación de los casos más resaltantes. Luego de ello, se lleva a cabo un proceso de auditoría interna, el cual le permite conocer de forma precisa los factores comerciales asociados a las casuísticas de fuga detectadas. Es decir, dicho proceso de auditoría permite conocer información asociada a los casos de los fugadores como: punto de venta, nombre del vendedor, modelos de los equipos móviles, pérdidas asociadas debido al descuento con el cual fue adquirido el equipo telefónico, etc. Una vez conocida la información anterior, la Empresa toma acciones comerciales concretas sobre los puntos de venta que permiten mitigar el riesgo o futuros casos de *fugadores*, o recuperar parte de los beneficios perdidos por dichos casos.

b) Comprensión de los datos:

Esta etapa será explicada en dos fases. En principio, se explicará el análisis que se llevó a cabo para concluir si era viable identificar a los clientes *fugadores* en una

ventana menor a los 45 días. Luego, se explicarán los análisis de las variables del negocio propios a la fase de Comprensión de los datos.

Fase 1: "Evaluación de la reducción de la ventana de monitoreo"

Este análisis fue desarrollado con el objetivo de evaluar en cúanto podría reducirse la ventana de monitoreo de 45 días de forma que no se pierda información relevante sobre el tráfico telefónico generado por los clientes. De este modo, se evaluaron distintos indicadores del negocio (y otros solicitados por los clientes internos) en conjunto con indicadores de tráfico telefónico. Entre los diversos indicadores evaluados se dio mayor relevancia al indicador F1, el cual cuantifica el porcentaje de tráfico telefónico de los clientes que se logra conservar al reducir la ventana de monitoreo actual de 45 días a una de menos días. Es importante indicar que este análisis fue desarrollado por un equipo comercial de la Empresa en conjunto con los integrantes del Equipo (expertos) que llevaba a cabo el monitoreo tradicional para la identificación de los clientes *fugadores*. A continuación, la Tabla 9 presenta los resultados del indicador F1 para distintos cortes de días (potenciales ventanas de monitoreo).

Tabla 9: Evaluación de umbrales para reducir la ventana de monitoreo

Monitoreos propuestos		F1	
		Tráfico generado	Tráfico por generar
Corte	15	74%	26%
	20	80%	20%
	25	**84%**	**16%**
	30	88%	13%
	35	90%	10%
	40	93%	7%

Fuente: Elaboración propia.

Como se puede ver, los resultados anteriores muestran que el porcentaje de tráfico que se logra conservar para un corte de 25 días corresponde al 84% del tráfico total. De ese modo, el comportamiento de tráfico telefónico que se sacrifica asciende al 16%. Dados los resultados anteriores, la recomendación dado por el Equipo de monitoreo actual fue desarrollar el modelo analítico considerando dicho corte (25 días) debido a que permite mantener un porcentaje representativo del tráfico total. Además, se llegó a un consenso entre el equipo comercial, la Gerencia y el equipo

desarrollador (conformado por el autor del presente trabajo) de optar por desarrollar la herramienta analítico-predictiva considerando una ventana de monitoreo de 25 días, evaluar sus resultados y, en caso no se tengan resultados satisfactorios plantear futuras mejoras y/o recomendaciones.

De acuerdo con lo anterior, la ventana de monitoreo que se consideró para el desarrollo del modelo en el presente trabajo fue de 25 días posteriores a la fecha de alta del equipo.

Fase 2: "Análisis de variables del negocio"

Para el desarrollo de esta fase se considerará el resultado de la fasa anterior. De este modo los valores de las variables serán medidos en una profundidad de 25 días, los cuales están comprendidos entre la fecha de alta y la fecha en la que se realizará el monitoreo con el modelo predictivo que se está elaborando.

Población de desarrollo:

La población de desarrollo está compuesta por todas las altas de equipos móviles adquiridas en la Empresa a través del Canal 1 en los periodos comprendidos entre jun-2016 y set-2016. A continuación, se presenta un esquema que muestra el diseño muestral del modelo.

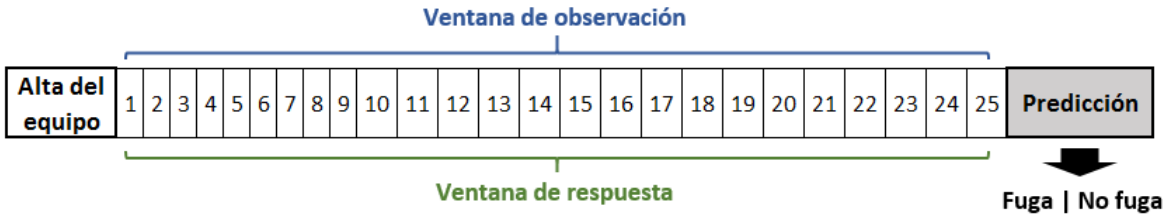


Figura 4: Esquema del diseño muestral.
Fuente: Elaboración propia.

Como se puede ver, la ventana de observación está compuesta por los 25 días de tráfico telefónico posteriores a la fecha de alta del equipo adquirido. En consecuencia, las variables del modelo serán construidas utilizando data transaccional del tráfico telefónico correspondiente al periodo indicado. Asimismo, se visualiza que la ventana de respuesta corresponde a los mismos 25 días posteriores a la fecha de alta del equipo. Esto se debe a que el evento de *fuga del equipo móvil* se da en el momento en el que este es adquirido en los puntos de venta. De este modo,

lo que se persigue con los monitoreos (tradicional y analítico-predictivo) es identificar a los clientes más propensos a haber realizado *fuga de equipo móviles* lo antes posible, y tomar medidas correctivas para controlar o mitigar esta problemática a futuro. A continuación se muestra la distribución de la variable objetivo dentro de la muestra de modelamiento.

Tabla 10: Tasa de fuga en la muestra de modelamiento

Target	Clientes (#)	Clientes (%)
Fuga	5,763	34.3
No fuga	11,065	65.8
Total	**16,828**	**100.0**

Fuente: Elaboración propia.

Como se observa, la tasa de fuga en la población de desarrollo es del 34.3%. Cabe indicar que aunque el valor de la tasa de fuga observada es alto, las ventas de este canal (Canal 1) solo representan el 6% del total de altas de equipo en la Empresa. Asimismo, es importante indicar que el valor de la tasa de fuga de la población de desarrollo es consecuente con los valores de la tasa de fuga mostrados en el apartado de Comprensión del Negocio. Finalmente, la tasa de fuga observada sugiere que no es necesario aplicar ninguna técnica de balanceo de los casos de éxito.

Descripción de los datos:

A continuación se muestra una descripción de la metadata de las variables disponibles recopiladas para el desarrollo del presente modelo.

Tabla 11: Descripción de variables disponibles

Origen de datos	Variable	Tipo de Variable	Tipo de Dato	Descripción
Fuente demográfica	Departamento	Cualitativa	Nominal	Departamento del punto de venta
Fuentes comerciales	Costo_Equipo	Cuantitativa	Numérica	Costo del equipo al ser adquirido por la Empresa
Fuentes comerciales	Gama	Cualitativa	Ordinal	Gama del equipo telefónico móvil
Fuentes comerciales	Marca	Cualitativa	Nominal	Marca del equipo telefónico móvil
Fuentes comerciales	Precio	Cuantitativa	Numérica	Precio de venta del equipo telefónico móvil
Tráfico telefónico	prom_llam_ent	Cuantitativa	Numérica	Número promedio de las llamadas entrantes
Tráfico telefónico	prom_llam_sal	Cuantitativa	Numérica	Número promedio de las llamadas salientes
Tráfico telefónico	prom_min_ent	Cuantitativa	Numérica	Número promedio de los minutos entrantes
Tráfico telefónico	prom_min_sal	Cuantitativa	Numérica	Número promedio de los minutos salientes
Tráfico telefónico	prom_sms	Cuantitativa	Numérica	Número promedio de los mensajes de texto salientes
Fuentes de monitoreo	riesgo_neto_imei	Cualitativa	Ordinal	Flag *fuga* (1:Fuga | 0:No fuga)

Fuente: Elaboración propia.

Como se observa, el conjunto de variables disponibles está conformado por diez potenciales variables independientes y la variable dependiente denominada

riesgo_neto_imei. Asimismo, se aprecia que cuatro variables independientes provienen de fuentes comerciales, una de fuente de información demográfica, y las cinco restantes de las fuentes del tráfico telefónico generado por los clientes.

Exploración de los datos:

El análisis exploratorio será presentado en dos partes. La primera mostrará los resultados de la exploración de las variables cuantitativas, y la segunda los resultados de las variables cuantitativas.

A continuación, la Tabla 12 muestra los resultados de la exploración de las variables cuantitativas a través del análisis univariado.

Tabla 12: Resultados del análisis univariado de las variables cuantitativas

Variable	Q1	Mediana	Q3	Media	Mínimo	Máximo	Valores perdidos	Desv. Estándar	Asimetría
Costo_Equipo	48.71	49.18	120.87	94.41	0	2010.17	0	71.75	3.78
Precio	59	59	99	93.89	49	2059	0	69.89	4.65
prom_llam_sal	0	0.32	1.14	0.95	0	34.16	0	1.76	4.91
prom_min_sal	0	0.34	1.28	1.46	0	103.51	0	3.99	8.53
prom_llam_ent	0	0.17	0.76	0.6	0	21.77	0	1.08	4.53
prom_min_ent	0	0.18	1.15	1.27	0	64.54	0	3.32	7.16
prom_sms	0	0.05	0.27	0.32	0	17.96	0	0.86	7.27

Fuente: Elaboración propia.

Como se puede ver, el valor mínimo de todas las variables de tráfico telefónico (*prom_llam_sal, prom_min_sal, prom_llam_ent, prom_min_ent* y *prom_sms*) es igual a cero, lo cual es esperado debido a que los clientes analizados han adquirido su equipo telefónico hace pocos días, lo cual hace más probable que alguno de ellos no hayan generado tráfico telefónico en alguna de las variables en mención. Además, se puede apreciar que los clientes del canal analizado (Canal 1), en promedio, realizan más llamadas salientes (media=0.95) de las que reciben (media=0.6). Pese a lo anterior, se observa que el promedio de minutos hablados salientes y entrantes (1.46 y 1.27 respectivamente) no se diferencian tan notoriamente. Por otro lado, las variables asociadas a las fuentes comerciales (*Costo_Equipo* y *Precio*) presentan una tendencia similar considerando las medidas de resumen mostradas, lo cual es esperado debido a que tienen una misma naturaleza (ambas están asociadas al equivalente en dinero del equipo telefónico). En términos generales, se observa que ninguna variable presenta valores perdidos, que los valores del coeficiente de asimetría indican que todas las variables presentan un sesgo hacia la derecha, y que las diferencias entre los valores máximos y los percentiles 75 (Q3 de cada variable)

sugieren la presencia de valores extremos (outliers). De acuerdo a ello, en el siguiente apartado se realizará la depuración de dichos valores extremos.

Respecto al análisis exploratorio de las variables cualitativas, se realizó un análisis univariado y bivariado para cada una de ellas (ver anexo 1). A continuación se exponen los principales hallazgos:

Departamento: Compuesto por 26 categorías, se observó que el ratio de fuga oscila entre 27.37% y 60%, no obstante, hay que tener en cuenta que 5 categorías ("1", "8", "10", "12" y "19") no contienen una cantidad representativa de casos. De acuerdo a lo anterior, se sugiere realizar una recategorización a fin de generar categorías mejor representadas, de mayor estabilidad en el tiempo y que ayuden a construir un modelo más parsimonioso.

Gama: Compuesta por 5 categorías, se observó que el ratio de fuga oscila entre 32.41% y 38.68% en las categorías bien representadas ("2", "3" y "5"). Asimismo, se observó que las categorías "1" y "4" no se encuentran adecuadamente representadas y podrían generar problemas de estabilidad de datos en el tiempo. Se sugiere recategorizar la variable.

Marca: Compuesta por 11 categorías, se observó que las categorías "2", "6", "7" y "10" no se encuentran bien representadas y podrían sufrir problemas de estabilidad en el tiempo. Asimismo, se observó que el ratio de fuga oscila entre 25.45% y 50.83% en las categorías bien representadas. Se sugiere recategorizar la variable con la finalidad de obtener categorías más sólidaz, libres de potenciales problemas de estabilidad y que ayude en la obtensión de un modelo más parsimonioso.

c) Preparación de los datos

En esta etapa se explicarán los procesos realizados con la finalidad de determinar el subconjunto de datos que será utilizado en el proceso de modelamiento. De este modo, se cubrirán aspectos como la limpieza de valores extremos, análisis de correlación, construcción de nuevas variables (también conocido como *ingeniería de variables*) y selección de variables. No obstante, como primer paso se procederá a dividir el conjunto de datos (inicial) en los muestras de traininig y testing.

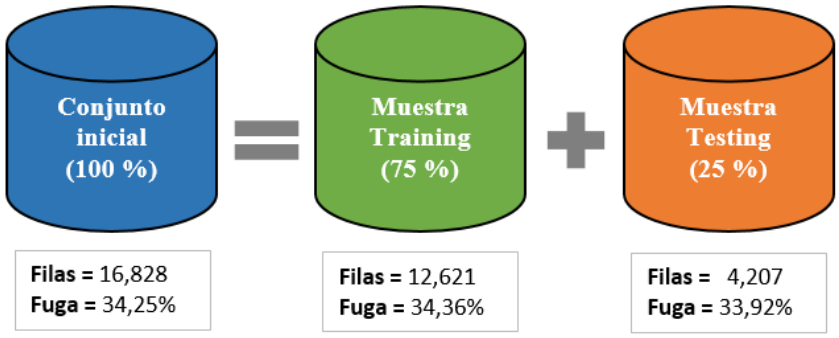


Figura 5: División en muestras Training y Testing.
Fuente: Elaboración propia.

De acuerdo a la figura anterior, la muestra training concentra el 75% del total de registros, mientras la muestra testing concentra el 25% restante. Asimismo, es importante indicar que los proceso mencionados en el párrafo anterior serán realizados únicamente sobre la muestra training. De ese modo, la muestra testing permanecerá inalterada, con el propósito de probar el desempeño del modelo obtenido en un conjunto de datos con valores naturales, es decir, tal y como vendrán cuando el modelo sea desplegado.

Limpieza de valores extremos

Para llevar a cabo esta tarea se hizo uso del criterio experto. De este modo, como se muestra en el anexo 2, primero se visualizó de forma univariada cada variable cuantitativa, y en función a su distribución se eligió un valor (umbral) a partir del cual los datos serían considerados como valores extremos. Asimismo, al elegir dicho umbral se verificó que la cantidad de casos excluidos no impacte significativamente en el conjunto de datos. Así, luego de limpiar los valores extremos de todas las variables cuantitativas la cantidad de registros de la muestra training pasó de 12,621 registros a 12,224 registros, lo cual representó una disminución del 3.15% del total de casos.

Pese a lo anterior, es importante indicar que previamente se evaluaron otros dos criterios de limpieza de valores extremos. El primero fue el basado en el comando "outlier()", el cual identifica al valor que esté más alejado a la media de la variable y lo define como umbral para identificar valores extremos. Mientras, el segundo criterio se basó en el análisis de los diagramas de cajas, el cual supone que los datos

mayores al “Q3 + 1.5*RIQ”, o menores al “Q1 - 1.5*RIQ” son considerados como valores extremos. No obstante, se determinó que estos dos últimos criterios no fueron eficientes al depurar a los valores extremos de la muestra training. El primer criterio realizaba una limpieza casi nula de los valores extremos, mientras que el segundo criterio ocasionaba una exclusión de un porcentaje significativo de casos.

Análisis de correlación

Como primer paso, se verificó la normalidad de las variables a fin de determinar qué coeficiente de correlación será el adecuado para evaluar la relación lineal entre las variables cuantitativas. De este modo, se realizó la prueba de normalidad de Anderson Darling y Kolmogorov Smirnov considerando un nivel de confianza del 0.05. La hipótesis de la prueba en mención es:

$$H_0: \textit{La variable sigue una distribución normal.}$$
$$H_1: \textit{La variable no sigue una distribución normal.}$$

De acuerdo a los valores del p-valor (aproximadamente cero), se concluyó que existe evidencia estadística suficiente para rechazar la hipótesis nula, con lo cual, se concluye que ninguna de las variables cuantitativas siguen una distribución normal. En consecuencia, se utilizó el coeficiente de correlación de Spearman el cual, al ser una prueba no paramétrica, no requiere del supuesto de normalidad para verificar si existen pares de variables con alta correlación lineal entre ellas. El siguiente recuadro muestra las correlaciones entre pares de variables cuantitativas:

Tabla 13: Matriz de correlaciones entre pares de variables cuantitativas

Correlación	Precio	Costo_Equipo	prom_llam_sal	prom_min_sal	prom_llam_ent	prom_min_ent	prom_sms
Precio		0.873	-0.0565	-0.0501	-0.0567	-0.0512	-0.0431
Costo_Equipo	0.873		-0.0512	-0.0446	-0.0575	-0.0513	-0.0294
prom_llam_sal	-0.0565	-0.0512		0.9791	0.9093	0.8802	0.875
prom_min_sal	-0.0501	-0.0446	0.9791		0.8876	0.8795	0.8685
prom_llam_ent	-0.0567	-0.0575	0.9093	0.8876		0.9672	0.7811
prom_min_ent	-0.0512	-0.0513	0.8802	0.8795	0.9672		0.775
prom_sms	-0.0431	-0.0294	0.875	0.8685	0.7811	0.775	

Fuente: Elaboración propia.

De acuerdo a los valores mostrados de la matriz de correlaciones, se observó una alta correlación entre varios pares de variables. Así, se observó que existe una alta correlación entre el par de variables "Precio" y "Costo_Equipo" (variables comerciales), y entre los distintos pares de variables de tráfico telefónico ("prom_llam_sal","prom_min_sal","prom_llam_ent","prom_min_ent","prom_sms "). En ese sentido, se considerarán estos resultados al momento de seleccionar las variables que entrarán al modelo.

Respecto a las variables cualitativas, se analizó la correlación entre cada una de ellas y la variable target (fuga) a través de la prueba de independencia de Chi-cuadrado. Es importante indicar que antes de llevar a cabo la prueba de hipótesis en mención se realizó una transformación de las variables cualitativas, las cuales se basaron en una recategorización que agrupó a las categorías con similar tasa de fuga (ver Anexo 3). La hipótesis de la prueba de independencia es la siguiente:

$$\boldsymbol{H_0}: \boldsymbol{Las\ variables\ son\ independientes}$$
$$\boldsymbol{H_1}: \boldsymbol{Las\ variables\ no\ son\ independientes}.$$

De acuerdo a los valores del p-valor (aproximadamente cero), y considerando el mismo nivel de confianza de la prueba anterior (0.05), se concluyó que existe suficiente evidencia estadística para rechazar la hipótesis nula. Es decir, se puede concluir que las variables cualitativas influye en la la variable target (*fuga*).

Selección de variables

Una vez realizados los análisis de anteriores, se procedió a evaluar la importancia de cada variable en término de cuánto aporta a explicar a la variable target (fuga). Para ello, se empleó el comando "IV" de la librería "InformationValue" del software R que calcula el Valor de Información (IV) de cada variable. La siguiente figura muestra los resultados de los IV de cada variable:

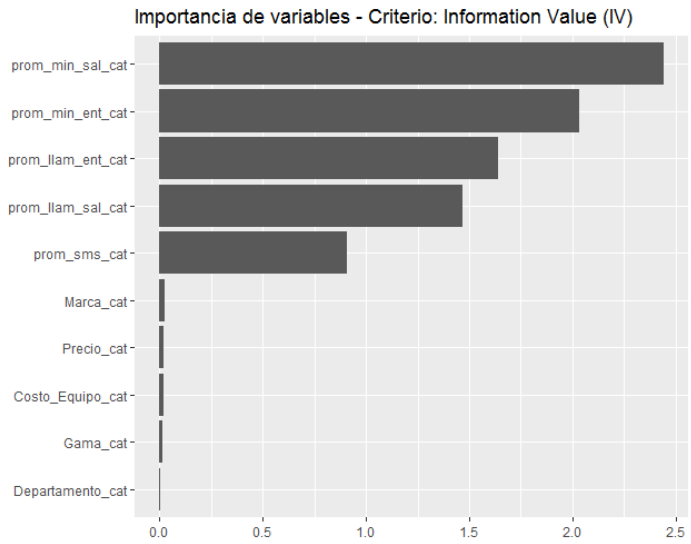


Figura 6: Importancia de variables según el criterio del "IV".
Fuente: Elaboración propia.

De acuerdo a los resultados mostrados, se seleccionaron las variables "prom_llam_ent_cat", "prom_min_sal_cat", "prom_sms_cat", "Precio_cat" y "Marca_cat" como las más importantes debido a las siguientes consideraciones:

1. La variable "Departamento_ca" fue descartada debido a que tiene el menor IV.
2. Pese a que mostraron un IV muy bajo se eligieron a las variables "Marca_cat" y "Precio_cat" debido a que permiten incluir información comercial del equipo. No obstante, se descartaron las variables "Gama_cat" y "Costo_Equipo_cat" (mismo tipo de información pero menor IV).
3. De las variables de tráfico telefónico, se seleccionó a la variables "prom_llam_ent" debido a que es la primera de su tipo que muestra una correlación media con otra variable de tráfico ("prom_sms") y a su vez permite incorporar a una variable de tráfico que mide el aspecto asociado a los minutos que el cliente ha consumido ("prom_min_sal").
4. De acuerdo a lo anterior, se descartaron el resto de variables de tráfico: "prom_min_ent" y "prom_llam_sal".

Pese a su alta correlación con el resto de variables de tráfico, es importante indicar que la inclusión de la variable "prom_min_sal" se forzó para incorporar el aspecto de los minutos hablados por los clientes, y para que el modelo no sea muy

dependiente del gran poder predictivo mostrado por la variable "prom_llam_ent"; la cual en caso sufra alguna cambio severo en su distribución no impacte tanto en el performance del modelo.

Finalmente, es importante indicar que en la fase de modelado se trabajó con la versión contínua de las variables de tráfico seleccionadas, así como con dicha versión de la variable "Precio_cat". De este modo, las variables a incluir en el modelo son:

Tabla 14: Variables seleccionadas para la fase de modelado

Origen de datos	Variable	Tipo de Variable
Tráfico telefónico	prom_llam_ent	Cuantitativa
Tráfico telefónico	prom_min_sal	Cuantitativa
Tráfico telefónico	prom_sms	Cuantitativa
Fuentes comerciales	Marca_cat	Cualitativa
Fuentes comerciales	Precio	Cuantitativa

Fuente: Elaboración propia.

d) Modelado

Una vez obtenido el conjunto de datos con las variables finales, se procedió a entrenar modelos utilizando el algoritmo Random Forest mediante el comando "randomForest" de la librería del mismo nombre. De este modo, se procedió a ejecutar un proceso iterativo considerando diferentes valores para los principales parámetros del algoritmo. Este procedimiento también es conocido como tuneo de parámetros. Asimismo, los principales parámetros del algoritmo son:

ntree: **Número de árboles a entrenar**.

mtry: **Número de variables a seleccionar en cada árbol**.

Es importante indicar que en el proceso de tuneo de parámetros se consideraron los valores 100, 200 y 300 para el parámetro **ntree**, y los valores 2 y 3 para el parámetro **mtry**. Finalmente, los valores de los parámetros elegidos fueron de 400 y 2 para los parámetros **ntree** y **mtry** respectivamente, con los cuales se tuvieron los siguientes resultados:

Tabla 15: Performance del modelo desarrollado

ntree	mtry	Sens	Espe	VPP	VPN	TCC
400	2	95.08%	99.57%	99.18%	97.36%	97.98%

Fuente: Elaboración propia.

e) Evaluación

Una vez obtenido el modelo analítico se procedió a testear su performance en la muestra testing, obteniéndose los siguientes resultados:

Tabla 16: Performance del modelo en la muestra testing

Sens	Espe	VPP	VPN	TCC
96.01%	99.78%	99.56%	97.99%	98.50%

Fuente: Elaboración propia.

De acuerdo a los resultados mostrados en la tabla anterior podemos concluir que el modelo permitió:

Sens=96.01% → Identificar correctamente al 96.01% del total de *fugadores* reales.

Espe=99.78% → Identificar correctamente al 99.78% del total de *no fugadores* reales.

VPP=99.56% → Acertó en el 99.56% de casos predecidos como *fugadores*.

VPN=97.77% → Acertó en el 97.99% de casos predecido como *no fugadores*.

TCC=98.50% → Clasificó correctamente, como *fugadores* o *no fugadores* al 98.50% del total de clientes analizados.

f) Implementación

Una vez desarrollado el modelo analítico que permite identificar a los *fugadores* del Canal 1 de forma anticipada al monitoreo tradicional de 45 días (en 25 días), se procedió a implementar de forma automatizada en los sistemas de la Empresa. Asimismo, pese a que la implementación del modelo desarrollado no se encuentra dentro del alcance del presente trabajo, a continuación se presenta un esquema general del proceso de implementación.

1. **Modelo.Rdata**: El modelo desarrollado que fue asignado a un objeto en el software R se guarda como un objeto de extensión .Rdata en un directorio.
2. **Script_puntuacion.R**: Se elabora un script en R que recibe un conjunto de datos, le crea las variables input del modelo, le aplica el modelo, obtiene y guarda las probabilidades de *fuga* y la clasificación (*fuga, no fuga*) de los clientes del

conjunto de datos, y finalmente guarda los cambios en el repositorio de base de datos de donde inicialmente leyó el conjunto de datos.

3. **Archivo_puntuacion.bat:** Se construyó un archivo batchero que mediante un clic ejecute automáticamente el script_puntuacion.R.
4. **ETL:** Se construyó un ETL (Extract, Transform and Load) en el software SQL Server Integration Services (SSIS) que extraía los conjuntos de datos a ser puntuados por el modelo y compilaba el Archivo_puntuacion.bat. Este ETL generaba un archivo de extensión .dtsx (al que denominaremos "paquete.dtsx").
5. **Batchero_automatizar.bat:** Se creó otro archivo batchero que mediante un clic ejecutaba automáticamente el archivo "paquete.dtsx".
6. **Tarea_programada:** Finalmente, se creo una tarea programada en el programador de window que ejecutaba a cierta hora y de forma diaria el archivo Batchero_automatizar.bat.

Con lo expuesto en los puntos anteriores se logró implementar de forma automatizada el modelo analítico de identifiación de *fugadores* del Canal 1.

4.3. Contribución en la solución de situaciones problemáticas

El desarrollo del presente trabajo le permitió a la Empresa poder identificar a los clientes *fugadores* del Canal 1 con una anticipación de 15 días respecto al monitoreo tradicional. Es decir, mientras el monitoreo tradicional permitía identificar a los clientes *fugadores* 45 días luego de la fecha de alta del equipo, el modelo analítico construido permitió realizar dicha identificación en el día 25. Adicionalmente, la implementación de la solución expuesta permitió que la tarea de identificación se base en el análisis de variables de fuentes comerciales y, principalmente, en el análisis de los patrones de tráfico telefónico de los clientes. De este modo, el proceso de identificación de clientes *fugadores* es más robusto debido a que no se basa la verificación del cumplimiento de reglas de negocio duras que no evolucionan con el tiempo, sino, en el estudio de los patrones de tráfico telefónico que los clientes dejan cuando hacen uso de los equipos móviles.

Por otro lado, el desarrollo de la solución expuesta le permitió a la Empresa transmitirle a sus clientes internos y externos su capacidad de innovación y desarrollo respecto a los últimos avances tecnológicos como lo son las soluciones basadas en la analítica de los datos.

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

El desarrollo de la solución expuesta en le presente trabajo representó un crecimiento sustancial de las competencias técnicas y habilidades blandas del autor. De este modo, los conceptos técnicos adquiridos durante su formación profesional en la Universidad Nacional Agraria La Molina fueron puestos en práctica a través del desarrollo de los distintos análisis llevados a cabo a través de todas las fases del CRISP. Asimismo, dichas capacidades técnicas se ampliaron debido al aprendizaje de aspecto como: la metodología de trabajo CRISP, programación en el software R a través del desarrollo de soluciones adhoc, investigación de metodologías para la evaluación de la importancia de las variables y del algoritmo de Aprendizaje Automático denominado Random Forest, etc. Además, aunque en una menor escala, el desarrollo de la solución le permitió al autor adquirir conocimientos básicos-intermedios del desarrollo de ETLs para la implementación del modelo analítico.

Por otro lado, respecto a las habilidades blandas, el desarrollo del presente trabajo le permitió al autor mejorar capacidades como:

Comunicación: Habilidad indispensable para interactuar con otros equipo de trabajo tanto en entornos técnicos como de negocio.

Escucha: Habilidad clave para entender las necesidades de los clientes internos y externos, a través de las cuales se logró precisar los objetivos principales y secundarios del trabajo.

Proactividad: Habilidad para satisfacer las necesidades de los clientes de forma anticipada y lograr así desarrollar soluciones idóneas.

4.5. Nivel de beneficio obtenido por el centro laboral

La implementación de la solución presentada le significó a la Empresa un impacto directo en el desarrollo de su negocio. Entre los cuales se puede destacar:

- Reducción de un 55% en el tiempo de identificación de clientes *fugadores* (de 45 a 25 días).
- Medidas correctivas anticipadas en los puntos de venta y socios estratégicos.
- Reducción de pérdidas asociadas a la *fuga de equipos móviles*.

Asimismo, a continuación se presentan los beneficios indicados medidos a través de indicadores del negocio:

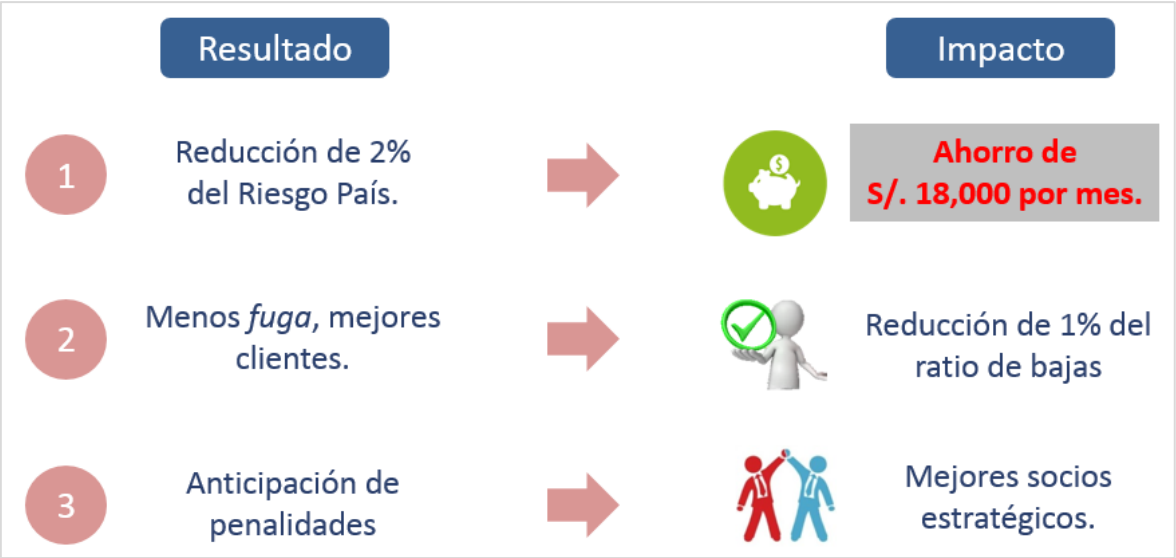


Figura 7: Beneficios asociados a la implementación de solución planteada.
Fuente: Elaboración propia.

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones:

Teniendo en cuenta los objetivos descritos en el presente trabajo se llegó a las siguientes conclusiones:

1. Se logró diseñar e implementar un proceso de monitoreo automatizado para identificar a los clientes que hicieron *fuga* de equipos móviles en una empresa de telecomunicaciones del Perú utilizando el algoritmo Random Forest.
2. Se logró reducir la cantidad de días del monitoreo actual de 45 días a 25 días sin perder información relevante sobre el tráfico telefónico de los clientes, al momento de identificar a aquellos que hicieron *fuga*.
3. Las principales variables que permiten identificar a los clientes *fugadores* en una ventana de 25 días son: “prom_min_sal”, “prom_llam_ent”, “prom_sms”, “Marca_cat” y “Precio”.

5.2. Recomendaciones:

A continuación, se presentan algunas recomendaciones del autor para futuros desarrollos que deseen complementar o mejorar la solución expuesta en este trabajo.

- Debido a los resultados *tan satisfactorios* del modelo, a pesar de la importante reducción de días del monitoreo, se sugiere una revisión de las reglas de negocio duras que regían el actual proceso de monitoreo pues podrían haberse *desfasado*.
- Comparar el desempeño del algoritmo Random Forest con el de otro algoritmo de clasificación como las Redes Neuronales, Regresión Logistica, XGBoost,etc.
- Analizar la inclusión de variables predictoras asociadas al uso paquetes de datos utilizados para la navegación por internet debido al cada vez mayor uso del internet.
- Monitorear periódicamente la calibración del modelo desarrollado con la finalidad de garantizar su óptimo rendimiento.
- Evaluar e implementar futuras actualizaciones al modelo analítico desarrollado en función a los resultados del monitoreo del desempeño del mismo.
- Implementar un proceso de Gobierno de Modelos que permita gestionar adecuadamente el modelo del trabajo presentado, y los futuros modelos analíticos que la Empresa desarrolle.

6. REFERENCIAS BIBLIOGRÁFICAS

Breiman, L. (2001). Random Forests. Statistics Department University of California Berkeley.

Idris, A., Iftikhar, A., & ur Rehman, Z. (2019). Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling. Cluster Computing - The Journal of Networks, Software Tools and Applications.

Jadhav, R., & Pawar, U. (2011). Churn Prediction in Telecommunication Using Data Mining Technology. International Journal of Advanced Computer Science and Applications (IJACSA).

Kashyap, K. (2019). Machine Learning - Decision Trees and Random Forest Classifiers. Medium.

Orellana Alvear, J. (2019). Árboles de decisión y Random Forest. bookdown.

Yildiz, M., & Albayrak, S. (2015). Customer Churn Prediction in Telecommunication with Rotation Forest Method.

7. ANEXOS

Anexo 1: Análisis bivariado de la variable Departamento.

Variable	Categoría	N.Total	N.Fugas	Fugas (%)
Departamento	1	5	2	40
Departamento	2	225	72	32
Departamento	3	18	7	38,89
Departamento	4	762	241	31,63
Departamento	5	97	37	38,14
Departamento	6	221	64	28,96
Departamento	7	95	26	27,37
Departamento	8	5	3	60
Departamento	10	5	2	40
Departamento	11	168	55	32,74
Departamento	12	7	4	57,14
Departamento	13	485	177	36,49
Departamento	14	369	140	37,94
Departamento	15	720	235	32,64
Departamento	16	663	224	33,79
Departamento	17	6583	2260	34,33
Departamento	18	326	98	30,06
Departamento	19	9	4	44,44
Departamento	20	58	20	34,48
Departamento	21	17	5	29,41
Departamento	22	1008	349	34,62
Departamento	23	86	29	33,72
Departamento	24	161	62	38,51
Departamento	25	65	26	40
Departamento	26	171	61	35,67
Departamento	27	292	99	33,9

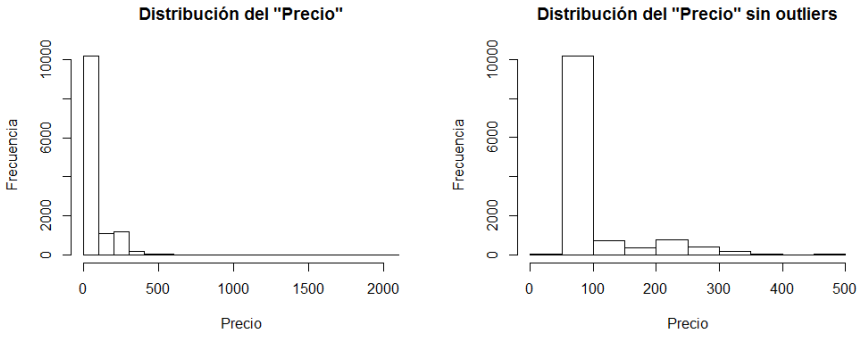
Anexo 2: Análisis bivariado de la variables Gama.

Variable	Categoría	N.Total	N.Fugas	Fugas (%)
Gama	1	23	4	17,39
Gama	2	3035	1174	38,68
Gama	3	414	159	38,41
Gama	4	2	0	0
Gama	5	9147	2965	32,41

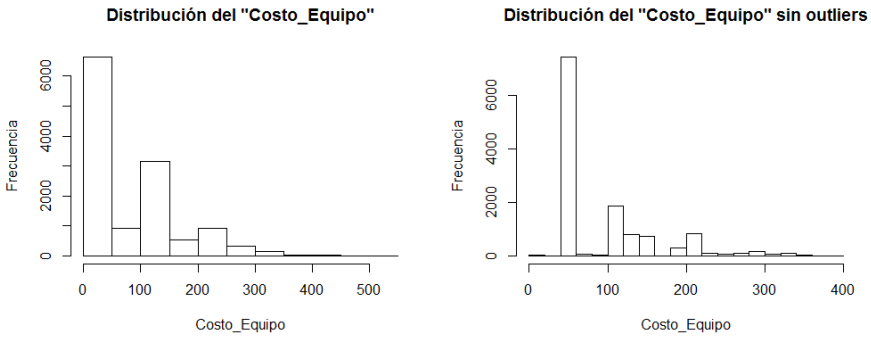
Anexo 3: Análisis bivariado de la variables Marca

Variable	Categoría	N.Total	N.Fugas	Fugas (%)
Marca	1	3930	1430	36,39
Marca	2	3	3	100
Marca	3	7950	2573	32,36
Marca	4	38	12	31,58
Marca	5	120	61	50,83
Marca	6	6	0	0
Marca	7	14	12	85,71
Marca	8	194	65	33,51
Marca	9	110	28	25,45
Marca	10	9	1	11,11
Marca	11	247	117	47,37

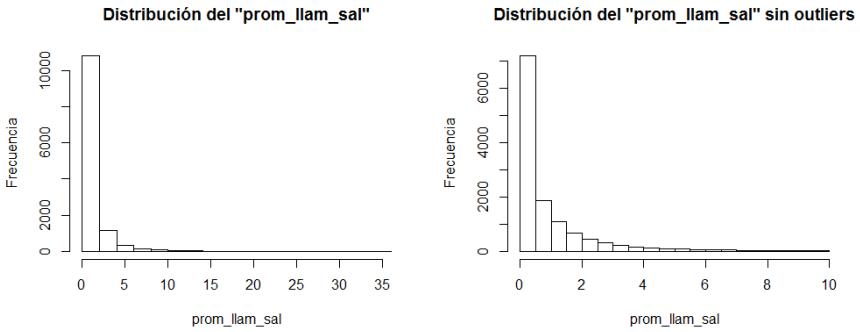
Anexo 4: Limpieza de valores extremos de la variable Precio.



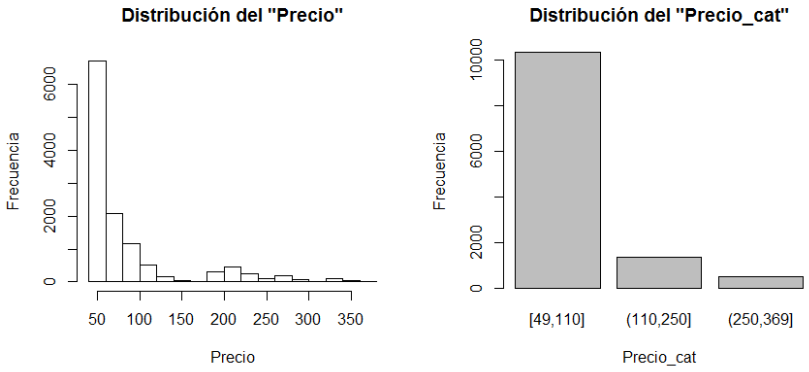
Anexo 5: Limpieza de valores extremos de la variable Costo_Equipo



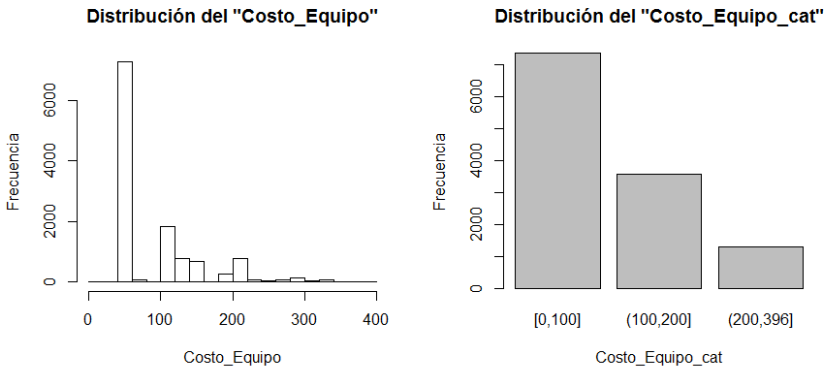
Anexo 6: Limpieza de valores extremos de la variable prom_llam_sal.



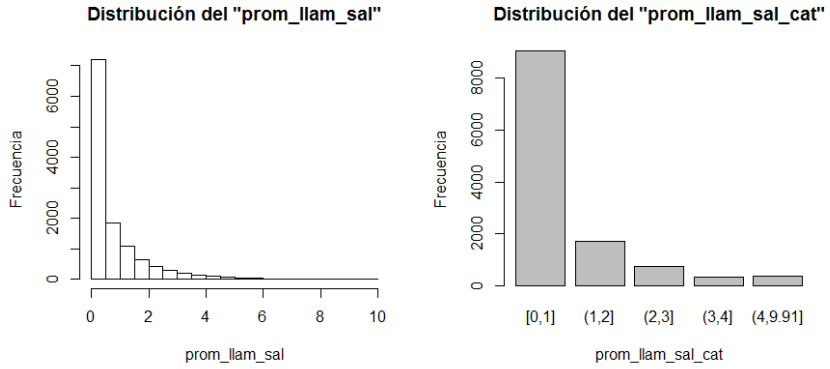
Anexo 7: Categorización de la variable Precio.



Anexo 8: Categorización de la variable Costo_Equipo.



Anexo 9: Categorización de la variable prom_llam_sal.



Anexo 10: Recategorización de la variable Marca.

Variable	Categoría	N.Total	N.Fugas	Fugas (%)
Marca	2	3	3	100
Marca	7	14	12	85,71
Marca	5	120	61	50,83
Marca	11	247	117	47,37
Marca	1	3930	1430	36,39
Marca	8	194	65	33,51
Marca	3	7950	2573	32,36
Marca	4	38	12	31,58
Marca	9	110	28	25,45
Marca	10	9	1	11,11
Marca	6	6	0	0

Anexo 11: Variable Gama recategorizada.

Variable	Categoría	N.Total	N.Fugas	Fugas (%)
Marca_cat	A	17	15	88,2%
Marca_cat	B	367	178	48,5%
Marca_cat	C	3930	1430	36,4%
Marca_cat	D	8182	2650	32,4%
Marca_cat	E	125	29	23,2%

Anexo 12: Recategorización de la variable Gama.

Variable	Categoría	N.Total	N.Fugas	Fugas (%)
Gama	2	3035	1174	38,68
Gama	3	414	159	38,41
Gama	5	9147	2965	32,41
Gama	1	23	4	17,39
Gama	4	2	0	0

Anexo 13: Variable Gama recategorizada

Variable	Categoría	N.Total	N.Fugas	Fugas (%)
Gama_cat	A	3449	1333	38,6%
Gama_cat	B	9147	2965	32,4%
Gama_cat	C	25	4	16,0%

Anexo 14: Código de procesamiento.

```
#------------------------------------------------------------------------
# 3) LIMPIEZA DE OUTLIERS - CRITERIO EXPERTO  -- By: Francisco Marquez Meza
cuantis
# [1] "Precio"       "Costo_Equipo"  "prom_llam_sal" "prom_min_sal"
# [5] "prom_llam_ent" "prom_min_ent"  "prom_sms"
dim_inicial <- dim(train) # DIMENSION INICIAL
dim_inicial
# [1] 12621    11

i=1 #Donde i "in" [1;7]
par(mfrow=c(1,2))
hist(train[,get(cuantis[i])],
    main=paste0('Distribución del "',cuantis[i],'"'),
    xlab = cuantis[i], ylab = 'Frecuencia')
unmbral=500 #Definir por cada variable
hist(train[get(cuantis[i])<unmbral,get(cuantis[i])],
    main=paste0('Distribución del "',cuantis[i],'" sin outliers'),
    xlab = cuantis[i], ylab = 'Frecuencia')
cbind(Porcentaje=round(100*dim(train[get(cuantis[i])>unmbral,])[1]/dim_inicial[1],2),
     Cantidad=dim(train[get(cuantis[i])>unmbral,])[1])
#------------------------------------------------------------------------
```