UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MEJORAMIENTO DE UN MODELO DE TARGETING DE CLIENTES DE TELEFONÍA MÓVIL USANDO ANÁLISIS DE REDES SOCIALES Y MINERÍA DE DATOS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

GONZALO IGNACIO HERMOSILLA MARTELLI

PROFESOR GUÍA:
SEBASTIÁN A. RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:
MARCEL GOIC FIGUEROA
FELIPE AGUILERA VALENZUELA

SANTIAGO DE CHILE
2015

RESUMEN DE LA TESIS
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL
POR: GONZALO HERMOSILLA M.
FECHA: 16/01/2015
PROF. GUÍA: SEBASTIÁN A. RÍOS

MEJORAMIENTO DE UN MODELO DE TARGETING DE CLIENTES DE TELEFONÍA MÓVIL USANDO ANÁLISIS DE REDES SOCIALES Y MINERÍA DE DATOS

En los últimos años, la industria de las telecomunicaciones se ha ido desarrollando en un escenario muy competitivo, lo que ha llevado a las compañías a enfocarse en lograr una relación rentable y de largo plazo con sus clientes. El problema surge cuando se quiere decidir con qué clientes construir dicha relación, cuya solución se basa en el concepto de *targeting*, el cual tiene por objetivo identificar a los clientes sobre quienes se realizarán acciones para retener e incrementar su valor. En este caso se quiere estudiar, en términos de adopción y rentabilidad, el desempeño de un modelo que selecciona el conjunto de clientes a quienes, a través de una campaña telefónica, se les ofrece un producto de telefonía móvil.

Las compañías han utilizado el enfoque de selección basándose en los atributos sociodemográficos y comerciales de sus clientes, sin considerar el efecto que podrían tener sobre las decisiones de éstos sus amigos, familiares o cercanos. Es por esto que se plantea un modelo de targeting que incorpore atributos sociales extraídos de la red de teléfonos móviles de cada cliente. Adicionalmente, se propone estudiar la influencia que podrían tener adopciones previas de sus amigos sobre la adopción propia del cliente en estudio.

Los modelos de targeting social fueron construidos en base a diversas técnicas de clasificación y diferentes configuraciones del conjunto de entrenamiento, estructuras que son probadas a través de una serie de experimentos que permiten comparar los resultados y establecer cuál es el modelo con la mayor capacidad para resolver este problema. La calidad de dichos modelos se evalúa en dos etapas diferentes. En una primera instancia se comparan los resultados obtenidos con los entregados por el modelo base que no incorpora atributos sociales, comparación que se realiza a través del número de aciertos acumulados que logra cada modelo en los cortes del ranking de clientes. Por otro lado, en una segunda fase se busca identificar la técnica de clasificación utilizada que mejores resultados entrega y la configuración del conjunto de entrenamiento que resulta en la mejor capacidad de predicción de adopciones.

El hecho de que la incorporación de los atributos sociales de los clientes no mejore por sí mismo el poder de predicción de los modelos de targeting, pero que sí lo haga la combinación de los resultados del modelo social con los del modelo base de comparación, resultó ser el mayor descubrimiento de este trabajo. Esto deriva en que para lograr modelos de selección efectivos es necesario combinar algunos de ellos que estén construidos en base a diferentes técnicas de clasificación, ya que éstas permiten identificar clientes de diversos perfiles, elevando la capacidad de predicción de los mismos. En este caso, la agregación de datos en el conjunto de entrenamiento y la combinación de aciertos de los modelos permiten incrementar en promedio en un 8 % el desempeño del modelo, alcanzando un nivel de aciertos de un 89 %.

A los mismos de siempre...

Agradecimientos

Con este trabajo se cierra una de las etapas más lindas y enriquecedoras que probablemente me tocará vivir a lo largo de mi vida, una etapa llena de logros, éxitos y alegrías, pero también de caídas, desencantos y tropiezos. A lo largo de estos años pasé por todo tipo de experiencias que significaron un mundo de aprendizaje, por lo que tengo la certeza de que éstas me han preparado para enfrentar con energía y sabiduría los nuevos desafíos que se vienen por delante. En este camino me acompañaron un sin número de personas a las cuales me gustaría dedicar unas palabras.

En primer lugar me gustaría agradecer a quienes siempre estuvieron a mi lado en este camino de crecimiento impulsándome a ser quién soy: mi familia y mi Valentina. A ellos agradezco la confianza que siempre depositaron en mi y las palabras de aliento que me regalaron cada vez que el partido se ponía cuesta arriba. Gracias a mis padres por darme la oportunidad de recibir la mejor formación y por darme las herramientas que me hicieran cada día más fácil. Gracias mi polola por todas las horas, que no fueron pocas, en que trabajamos juntos y por el apoyo incondicional que me entregó para que pudiera cumplir todas mis metas.

Agradezco a todo el Centro de Inteligencia de Negocios, en particular al profesor Sebastián Ríos por entregarme la responsabilidad de realizar un trabajo tan desafiante como éste, y a Lautaro e Iván por guiarme desde el inicio con toda la disposición y dedicación que les fuera posible.

Mención especial también para mis amigos y compañeros, con quienes siempre tuve la cuota de diversión y alegría necesaria para sobrellevar de mejor manera el día a día, las extenuantes clases y las largas horas de estudio.

Finalmente me gustaría agradecer a la Universidad de Chile y a su Escuela de Ingeniería por cumplir con el rol más difícil e importante que se le puede encomendar a una institución: educar y formar a los jóvenes que en un futuro no muy lejano serán los responsables de hacer de nuestro país un mejor lugar para vivir, siendo intachables personas y virtuosos ciudadanos.

Tabla de Contenido

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

En los últimos años, especialmente desde el año 2010, la economía chilena ha estado creciendo en torno al 5 %, marcando el 2013 un incremento del 4,1 % [23]. En este contexto, el sector Telecomunicaciones ha presentado una evolución importante, alcanzando un crecimiento de un 4,9 % durante el año 2013, explicado principalmente por el desempeño de los mercados de Internet y telefonia móvil, donde este último logró una penetración de 134,2 %, es decir, alrededor de 135 teléfonos por cada 100 habitantes [80]. Así es como este sector se ha ido desarrollando en medio de un escenario muy competitivo, marcado por la reciente puesta en marcha de la portabilidad numérica[1] y la entrada de nuevos operadores al mercado[2], dinamismo que le ha permitido alcanzar una mayor relevancia dentro de la estructura económica chilena, representando el 2 % del Producto Interno Bruto del 2013 (US$267.000 millones) [23].

Todo lo anterior trae consigo un desafío para las compañías de telecomunicaciones, ya que éstas deben ser capaces de encontrar nuevas estrategias que les permitan diferenciarse de sus competidores, sobretodo en lo que respecta a la relación que logran establecer con sus clientes. Esta memoria busca encontrar nuevas maneras de resolver un problema muy común en la industria de las telecomunicaciones, el cual será descrito en la próxima sección.

1.1. El Problema

El problema que esta memoria busca resolver consiste en cómo mejorar el modelo de selección de clientes que formarán parte del segmento objetivo de una campaña de marketing *directo* relativa a un producto de telefonía móvil. Generalmente las compañías de telecomunicaciones han resuelto este problema basándonse tanto en la información sociodemográfica de sus clientes como en su comportamiento comercial y telefónico. Bajo esta mirada clásica,

[1] Ley que establece que los usuarios de telefonía fija y móvil son dueños de sus números telefónicos. Esto les permite cambiar de compañía conservando su número telefónico sin tener que enfrentar multas ni trabas contractuales.

[2] La cantidad de empresas participantes ha ido en aumento, pasando de tres a nueve operadores vigentes en el mercado.

al seleccionar un cliente sólo se considera la propensión individual del cliente a aceptar una oferta por un producto, sin incluir en el análisis el efecto que puede tener sobre su decisión la red social telefónica de la cual participa, lo que trae consigo una pérdida de información importante. Este trabajo apunta a elegir como objetivo de campaña a aquellos clientes más valiosos y con mayor probabilidad de adoptar el producto ofrecido, tomando en consideración el valor individual y el *valor social* que cada uno de ellos entrega a la compañía. En los capítulos venideros se mostrará el enfoque propuesto respecto de este problema.

1.2. Motivación

A lo largo de los años las compañías han conseguido implementar un conjunto de procesos y sistemas que les han permitido levantar diversas estrategias de negocios orientadas a construir una relación rentable y de largo plazo con ciertos clientes específicos denomidada *Customer Relationship Managment* o Gestión de Relación con el Cliente (CRM) [60]. Gracias al desarrollo de Internet y de nuevas tecnologías, el CRM se ha enfocado en el *Business Intelligence* o Inteligencia de Negocios como herramienta para lograr adquirir y retener clientes que le permitan maximizar el valor que cada uno de ellos entrega a la organización [67].

En una industria tan competitiva como la de las telecomunicaciones, el enfoque CRM aparece como una de las herramientas fundamentales para que las compañías logren diferenciarse unas de otras en un mercado que se caracteriza por ser muy dinámico. Tal como se menciona en [72], en general es más costoso conseguir un nuevo cliente que retener uno ya existente, por lo que los esfuerzos deben enfocarse en retener e incrementar el valor de éstos, estrategia que coloca como eje principal de la gestión a los procesos de *valorización* y *selección* de clientes.

Generalmente son tres aspectos los necesarios para lograr un incremento eficaz del valor de los clientes: *up-selling*, *cross-selling* y retención [75]. Up-selling consiste en vender al cliente productos similares a los que haya comprado anteriormente pero de mayor valor; y cross-selling se refiere a la venta de productos que el cliente nunca haya comprado, es decir, un nuevo tipo de productos; mientras que la retención corresponde al esfuerzo de mantener a los clientes al interior de la compañía. Dado esto, ahora se presenta el problema de cómo seleccionar a los clientes adecuados sobre quiénes se desea realizar una campaña de marketing directo, logrando incrementar el valor que éstos dejan a la compañía.

El problema de selección lo trata de resolver el enfoque de *targeting* de clientes, el cual busca construir una estrategia cuyo principal objetivo es identificar a clientes específicos que se ajusten a los intereses de los tomadores de decisión y sobre los cuales se desean establecer planes de marketing directo [87]. Bajo esta mirada, el CRM busca seleccionar a aquellos clientes que potencialmente podrían incrementar su valor basándose en las herramientas que entrega Inteligencia de Negocios y Minería de Datos, las cuales, a través de la construcción de modelos de clasificación basados en técnicas como Árboles de Decisión [77], Redes Neuronaloes [66] y Support Vector Machines [13], son capaces de identificar a los clientes más

propensos o con mayor probabilidad de adoptar un producto ya conocido a través de un up-selling (e.g, incrementar el valor de su plan telefónico móvil) o uno nuevo a partir de un cross-selling (e.g, comprar una bolsa de Internet).

Un nuevo enfoque se puede dar a la valorización y selección de clientes a través del *targeting social* de clientes. Éste apunta a que al momento de valorizar un cliente no sólo puede ser considerado el valor individual que éste tiene por el hecho de ser cliente, sino que también debiera ser tomado en cuenta el *valor social* que cada cliente posee dada la relación que éste tiene con todos los miembros que pertenecen a su *red social* [25]. Dado que la red de telefónos móviles de un cliente puede ser representada por una red social [27], la combinación de herramientas de Análisis de Redes Sociales y Minería de Datos será capaz de extraer información estructural de grafos, la que permitirá construir una serie de nuevos atributos sociales de manera de complementar los modelos de clasificación que se utilizan en la selección de clientes.

Este enfoque podría ejemplificarse con el caso de un cliente que por sí solo no califica como un cliente de alto valor, ya que no es propenso a realizar cross-selling de un nuevo producto (por lo tanto no es seleccionado), pero que posee un valor social alto, dado que al interior de su red social de teléfonos móviles tiene *amigos* que sí lo adoptaron, los que pueden ejercer *presión social* sobre él, incrementando la probabilidad de éste realice la compra cruzada, y así incluyéndolo en el conjunto de clientes seleccionados.

En lo que sigue, esta memoria describirá cómo a través de la utilización de técnicas de clasificación y de la inclusión de atributos sociales en los modelos tradicionales, es capaz de realizar una selección de clientes bajo un nuevo enfoque sobre los cuales se desea llevar a cabo una serie de acciones de marketing para que éstos incrementen su valor por medio de up-selling o cross-selling.

1.3. Significancia del Problema

Este estudio presenta un nuevo enfoque para resolver un importante problema para la industria de las telecomunicaciones. Incorporando mayor información a través de atributos sociales de clientes extraídos de su red social de teléfonos móviles, se busca mejorar un modelo de clasificación que tiene por objetivo seleccionar el conjunto de clientes más propensos a adoptar un producto por medio de un up-selling o un cross-selling. Esta metodología permite hacer más eficiente la selección de los clientes que son *target* de una campaña de marketing, a través de una elección de los clientes con mayor probabilidad de adquirir el producto, y que a la vez son aquellos que tienen un alto valor para la compañía.

1.4. Objetivos

En esta sección serán presentados los objetivos de esta memoria, tanto el objetivo general como aquellos específicos.

Objetivo General

Mejorar la predicción de un modelo de targeting de clientes para la adopción de productos a través de la inclusión de información extraída de una red social de teléfonos móviles, usando Análisis de Redes Sociales y Minería de Datos.

Objetivos Específicos

1. Revisar bibliografía relacionada con técnicas de clasificación, *targeting* de clientes, análisis de redes sociales y marketing basado en redes sociales.

2. Evaluar las mejoras en el poder de predicción de los modelos de targeting que incorporan atributos sociales extraídos de la red social de teléfonos móviles.

3. Llevar a cabo una serie de experimentos que varían la configuración del conjunto de entrenamiento del modelo de targeting con el fin de encontrar aquella con el mejor desempeño predictivo.

4. Realizar una comparación de los resultados obtenidos por los modelos de targeting social basados en tres técnicas de clasificación diferentes.

1.5. Resultados Esperados

1. Generación de un reporte (un capítulo) que resuma los antecedentes revisados relacionados con técnicas de clasificación, *targeting* de clientes y marketing basado en análisis de redes sociales.

2. Medición de la calidad de los resultados obtenidos con el nuevo enfoque de selección, comparando con resultados entregados por los modelos que no incluyen atributos sociales.

3. Reconocimiento del diseño del conjunto de entrenamiento que permite a los modelos de targeting lograr el mejor desempeño predictivo.

4. Identificación de la técnica de clasificación que presenta el mayor poder de predicción para la adopción de productos de telefonía móvil.

1.6. Estructura de la Memoria

Para resolver el problema que esta memoria plantea es indispensable contar con un nivel de entendimiento profundo acerca de modelos de targeting, selección de clientes, técnicas y algoritmos de clasificación y análisis de redes sociales. De manera de resumir todos estos antecedentes, en el Capítulo 2 se presenta una síntesis de los aspectos más importantes de la literatura asociados a clasificación de instancias y las técnicas existentes, targeting de clientes, análisis y construcción de redes sociales y targeting asociado a SNA.

En el Capítulo 3 se presentará la metodología utilizada, la cual se basa en la combinación del proceso jerárquico CRISP-DM con SNA, donde se revisará en detalle cada uno de los pasos que la componen, abarcando tanto las fases originales de ella como las que se crearon para efectos de esta investigación. En particular, se describirán los pasos de la metodología asociados a la preparación de los datos, la configuración de la red social y la construcción de los modelos de targeting.

Una aplicación de la metodología sobre datos reales de una compañía de telecomunicaciones será presentada en el Capítulo 4. En él se da a conocer el diseño de cada uno de los experimentos llevados a cabo y se describe en detalle el preprocesamiento al cual fueron sometidos los datos de los clientes. Posteriormente se explica cómo se construyó la red social de teléfonos móviles y cómo a partir de ella se extrajeron los atributos sociales que serán incluidos en el conjunto de variables que será utilizado como base para el modelo de targeting social. Más adelante se intenta estudiar el efecto que la adopción social podría tener sobre la decisión del cliente, para luego presentar los resultados de todos los experimentos realizados bajo la utilización de las tres técnicas de clasificación. Finalmente las mejoras logradas en los modelos se traducen a cifras económicas a través de una evaluación financiera que permite dimensionar el monto de dinero extra que la compañía podría ingresar al usar el modelo de targeting propuesto.

Finalmente en el Capítulo 5 se comentarán las principales conclusiones extraídas a partir de los resultados obtenidos en la sección anterior, incluyendo los descubrimientos y aportes más relevantes de este trabajo de investigación. Adicionalmente se entregerán algunos lineamientos para el trabajo futuro que permitirán profundizar el estudio de este tipo de problemas, particularmente asociados al masivo mundo de las telecomunicaciones, y que darán pie para seguir perfeccionando la resolución de los mismos.

Capítulo 2

Revisión de Antecedentes

En este capítulo se revisará la bibliografía más relevante en relación a los métodos y técnicas utilizadas, con el fin de dar a los lectores algunas nociones sobre ciertos temas claves para entender este trabajo. En primer lugar se presentará una breve descripción del problema de clasificación y de las técnicas más utilizadas para clasificar instancias, para luego profundizar en el uso de este enfoque como herramienta para el *targeting* o selección de clientes. Posteriormente, se introducirán algunas nociones generales respecto de análisis de redes sociales, para finalmente entender cómo todo lo anterior puede complementarse para realizar un *targeting* más certero basado en las propias redes sociales que el cliente puede construir.

2.1. Clasificación

Machine Learning o Aprendizaje Automático refiere al desarrollo de sistemas automáticos que son capaces de procesar grandes volúmenes de datos con el fin de extraer información significativa que sea potencialmente utilizable para apoyar la toma de decisiones [62]. En general, este concepto permite enfrentar problemas que requieren de la identificación de patrones comunes a través de diferentes técnicas de minería de datos que permiten la formulación de modelos de asociación, estimación, clasificación y segmentación.

Clasificación consiste en el proceso de búsqueda de una función que describa y distinga clases de datos con el propósito de utilizar la información *aprendida* para predecir la etiqueta de los objetos cuya clase es desconocida, por lo que un modelo de clasificación correponde a un modelo de aprendizaje supervisado, ya que para cada observación se conoce previamente a qué clase pertenece o qué valor le corresponde.

Un modelo del clasificación debe ser capaz de mapear o clasificar una instancia caracterizada por un vector o conjunto de variables en una de varias clases predefinidas [79], las que generalmente son escogidas para ser disjuntas, por lo que cada observación es asignada a una

y sólo una de estas clases. El espacio donde viven estas observaciones puede ser dividido en *regiones de decisión* cuyos límites son llamados *límites de decisión* o *superficies de decisión*, permitiendo así separar estos *items* en diferentes grupos a través de estas fronteras [9]. La Figura 2.1 muestra un conjunto de datos que son claramente diferenciables y potencialmente separables en dos clases, mientras que la Figura 2.2 refleja como estas clases son separadas de manera simple a través de un límite lineal.

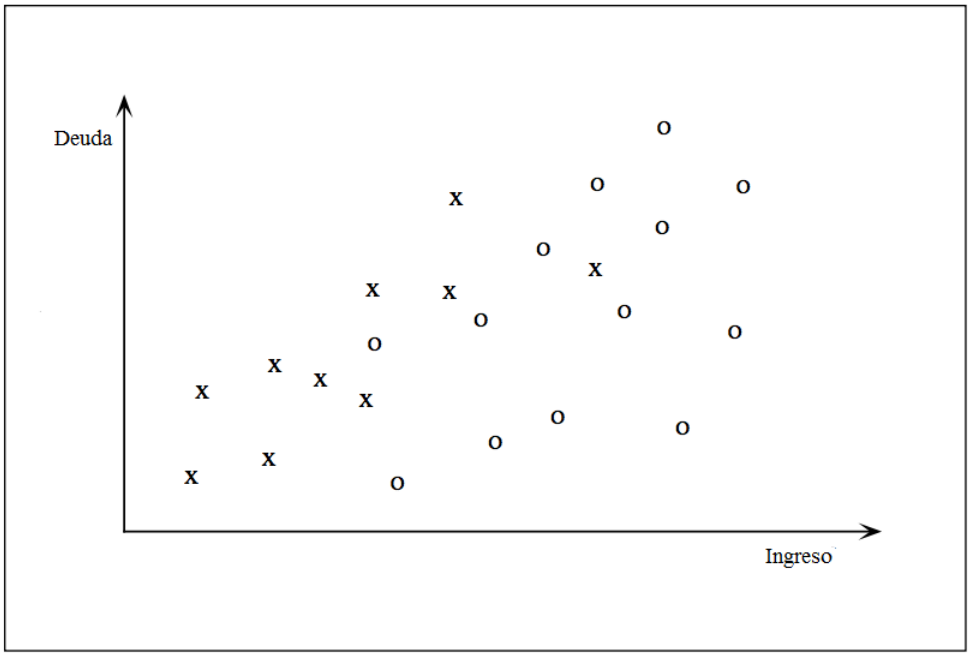


Figura 2.1: Conjunto de datos con dos clases diferenciables

Existe un sin número de problemas donde la aplicación de modelos de clasificación entrega importantes soluciones a industrias como el *retail*, medicina, banca y crédito, manufactura, telecomunicaciones, servicios públicos, transporte y seguros; basándonse en distintas fuentes de información como la demográfica, estilos de vida, comportamiento crediticio, características financieras, patrones de compra e información comercial o de marketing. Algunos ejemplos de estos problemas son la clasificación de postulantes a crédito como de bajo, medio o alto riesgo; la elección del contenido a desplegar en un sitio web; determinación de qué número de teléfono corresponden a máquinas de fax; detección de reclamos fraudulentos; y asignación de trabajos en base a descripción de trabajos de texto libre [7]. Las técnicas más comunes y típicamente utilizadas para resolver problemas de clasificación son Árboles de Decisión o *Decision Trees*, Aprendizaje basado en Reglas o *Based-rules Learning*, Redes Neuronales o *Artificial Neural Networks*, *Naives Bayes* y *Support Vector Machines*.

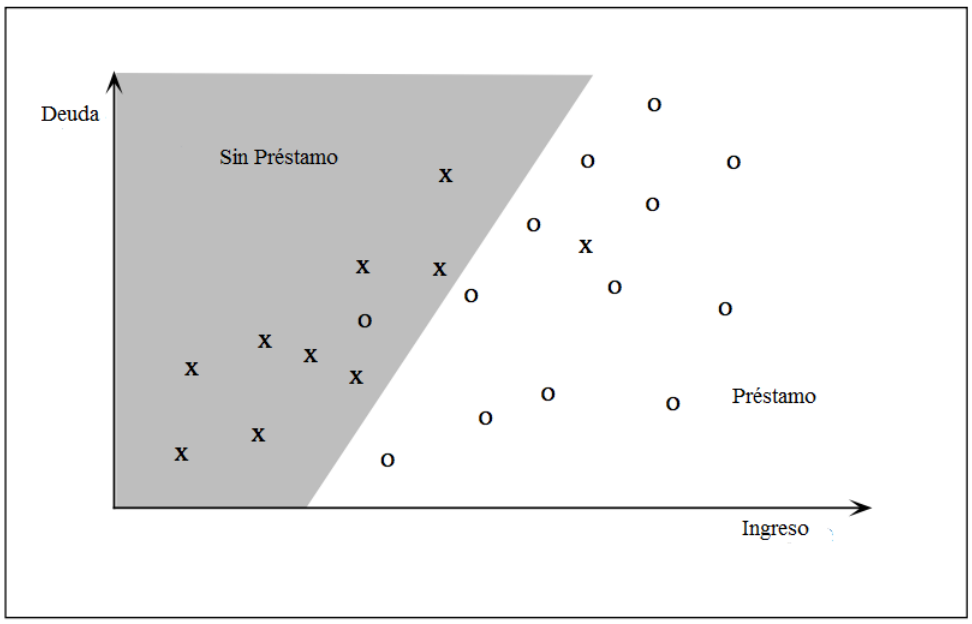


Figura 2.2: Límite de clasificación lineal simple

2.2. Técnicas de Clasificación

Para efectos de esta memoria sólo se entrará en detalle de las técnicas de clasificación utilizadas en el desarrollo de la investigación, es decir, Árboles de Decisión, Redes Neuronales y Support Vector Machines. Para Based-rules learning puede encontrarse mayor información en [37] y para Naives Bayes en [76].

2.2.1. Árboles de Decisión

Los árboles de decisión son árboles que clasifican instancias ordenándolas de acuerdo a las características de cada una de ellas. Un árbol de decisión consta de un nodo *raíz* que no tiene arcos entrantes, nodos internos que poseen arcos de entrada y de salida y nodos terminales o nodos *hojas* que no tienen arcos salientes, donde cada uno de éstos representa una característica de la instancia que será clasificada y cada rama o arco representa el valor que cada nodo puede asumir [77]. Las instancias son clasificadas en clases predefinidas, subdividiendo el árbol secuencialmente -comenzando por el nodo raíz- en base a las reglas de decisión definidas, asignando una etiqueta de clase a cada observación de acuerdo a al nodo final en la que ésta cae [32]. La Figura 2.3 es un ejemplo de un árbol de decisión para el conjunto de entrenamiento de la Tabla 2.1, donde la instancia [at1 = a1, at2 = b2, at3 = a3, at4 = b4] es asignada a los nodos at1, at2, y finalmente al at3, el que clasifica a la instancia como positiva (representada por el valor “Si”).

atributo1 (at1)	atributo2 (at2)	atributo3 (at3)	atributo4 (at4)	Clase
a1	a2	a3	a4	Si
a1	a2	a3	b4	Si
a1	b2	a3	a4	Si
a1	b2	b3	b4	No
a1	c2	a3	a4	Si
a1	c2	a3	b4	No
b1	b2	b3	b4	No
c1	b2	b3	b4	No

Tabla 2.1: Conjunto de entrenamiento para un árbol de decisión

El atributo que mejor divida el conjunto de entrenamiento debería ser el nodo raíz del árbol. Existen numerosos métodos que permiten encontrar dicha característica tales como *information gain* o ganancia de información [50] y el *índice de gini* [12], sin embargo, la mayoría de los estudios han concluido que no existe un sólo método que sea el mejor de todos.

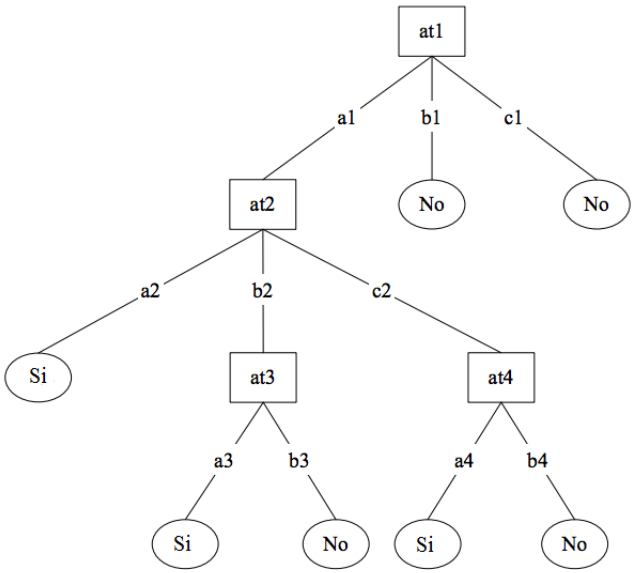


Figura 2.3: Árbol de Decisión

Los árboles de decisión representan hoy una herramienta altamente efectiva en áreas tales como *text mining* o minería de textos, extracción de información, machine learning y reconocimiento de patrones. Además, presentan una serie de beneficios, entre los cuales se destacan [77]:

- Versatilidad para una amplia gama de problemas de minería de datos, tales como clasificación, regresión, *clustering* y selección de atributos.

- Modelo autocontenido, donde la estructura de decisión es explícita, lo que lo hace fácil de entender e interpretar.

- Flexibilidad en el manejo de diferentes tipos de datos de entrada: nominales, numéricos o de texto.

- Adaptabilidad en el procesamiento de conjuntos de datos que poseen errores o datos faltantes.

Dada su versatilidad y flexibilidad, los árboles de decisión han sido ampliamente utilizados en un sin número de aplicaciones y problemas de minería de datos como por ejemplo la predicción de terremotos [91], donde se intenta predecir la concentración de un gas característico que se libera después de un movimiento telúrico en base a una serie de características del medio ambiente; o el apoyo en decisiones médicas [74] que incluyen casos como diagnóstico de infarto al miocardio [81], identificación de señales de una posible respuesta adversa a cierta droga [52] o apoyo en decisiones en el contexto de una campaña de vacunación contra el sarampión [68].

En un lado más comercial, se pueden destacar aplicaciones como el diseño de un sistema de reconocimiento y entendimiento de un lenguaje realizado en base a un árbol que clasifica semántincamente secuencias de palabras [57]; la creación de un sistema de recomendación de productos personalizado basado en el uso de la web por parte del cliente [21]; la predicción de fuga de clientes y gestión de *churn* desarrollada en una compañía de telecomunicaciones [4]; y la construcción de un modelo de *scoring* para apoyar decisiones de crédito en una institución financiera [6], entre muchas otras aplicaciones y usos que muchos autores han logrado dar a esta importante herramienta.

2.2.1.1. Variaciones de árboles de decisión

A continuación se presentarán técnicas que varían la configuración de los árboles de decisión, pero que los siguen utilizando como unidad básica.

Random Forest

La técnica *random forest* o *bosque aletorio* es una combinación de árboles de decisión tal que cada árbol depende de los valores de un vector seleccionado aleatoriamente con la misma distribución para todos los árboles en el bosque [12]. En la etapa de entrenamiento, esta técnica utiliza una muestra del conjunto de entrenamiento original, buscando sólo a lo largo de un subconjunto de los atributos de entrada escogido aleatoriamente para determinar la división en cada nodo. Para la clasificación, cada árbol del bosque entrega una respuesta respecto de a qué clase debiera pertenecer cada instancia que entró como input, siendo la decisión definitiva determinada por la respuesta más frecuente entre los árboles individuales [36].

Esta variación de los árboles de decisión puede manejar datos de mayor dimensionalidad y utilizar una mayor cantidad de árboles en conjunto. Esto, combinado con el hecho de que la selección aletoria de atributos busca minimizar el error, hace que los resultados obtenidos sean comparables con otras técnicas, pero siendo computacionalmente más ligero, lo que por supuesto representa una ventaja muy importante.

Alternating Decision Trees

Esta técnica de alternancia de árboles de decisión corresponde a una estructura similar a los árboles tradicionales que trabaja con una representación distinta de la regla de clasificación común [31]. En esta nueva representación cada nodo de decisión clásico es reemplazado por dos nodos: un *nodo de predicción* y un *nodo separador*. El nodo separador tiene la misma función que el nodo de decisión típico, mientras que el nodo de predicción está asociado a un valor real.

Tal como en un árbol de decisión tradicional, una instancia es ingresada a un camino que recorre el árbol desde la raíz hasta una de las hojas. Sin embargo, a diferencia de los árboles clásicos, la clasificación asociada al camino recorrido no es etiquetada en la hoja final, si no que depende del signo de la suma de las predicciones a lo largo del camino.

Un ejemplo de esto se muestra en la Figura 2.4. En (a) se presenta un árbol de decisión tradicional que clasifica la instancia $a = 0.5$ como clase -1, mientras que en (b) se tiene un árbol alternante que toma la misma instancia $a = 0.5$, suma sus predicciones como $signo$(0.5-0.7-0.2) = $signo$(-0.4), y dado que el signo es negativo, la clasifica como clase -1.

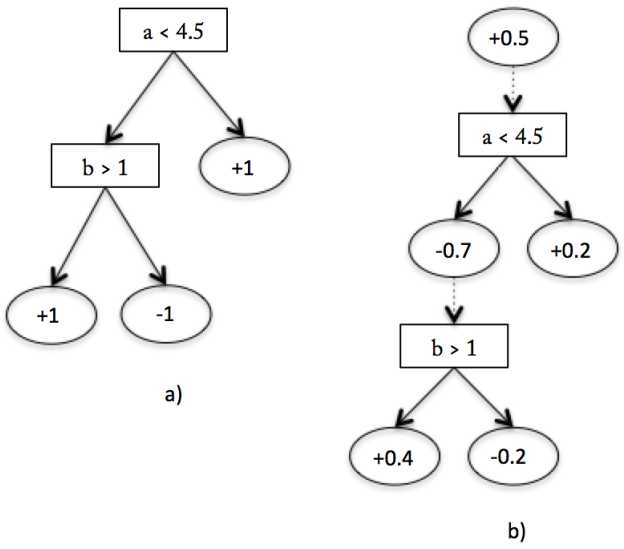


Figura 2.4: (a) Árbol de decisión tradicional, (b) Árbol de decisión alternante.

2.2.2. Redes Neuronales

Una red neuronal artificial puede considerarse como un modelo simplificado de una red neuronal biológica, cuya estructura consta de una malla de unidades de procesamiento o *neuronas* conectadas entre sí de acuerdo a una tipología que busca reconocer y aprender patrones a partir de un conjunto de datos que servirá como entrada al modelo [89]. Como muestra la Figura 2.5, cada neurona consiste en una parte de suma que y una parte de salida. La etapa de suma recibe N valores de entrada o *inputs*, para luego asignar un ponderador a cada uno de éstos, y ejecutar una suma ponderada, cuyo resultado se conoce como *valor de activación*; mientras que la etapa de salida produce una señal de salida o *output* a partir del valor de activación.

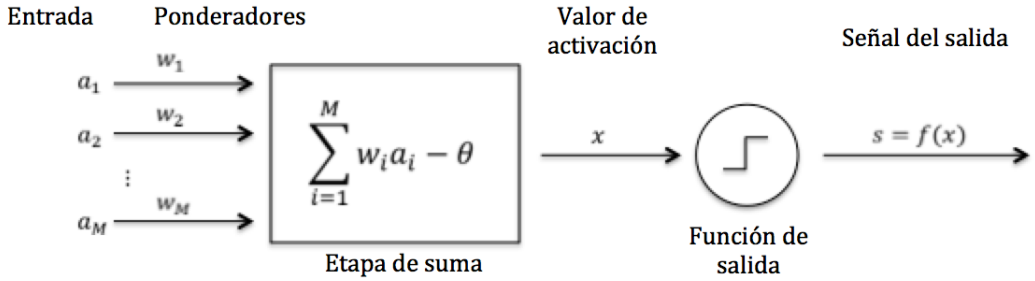


Figura 2.5: Estructura de una red neuronal artificial

El funcionamiento de una red neuronal comienza cuando cada unidad de la malla recibe inputs desde otras unidades conectadas o desde fuentes externas, generando la suma ponderada de estos valores que representará el valor de activación, el que a su vez determinará la señal de salida que será emitida por la unidad de salida. El proceso continúa gracias a que estos valores de salida y otros inputs establecen la activación de otras neuronas y la generación de sus respectivos outputs, y así sucesivamente hasta que el modelo logra encontrar el patrón buscado en los datos.

El *aprendizaje* se genera gracias a una combinación particular de neuronas, interconexiones y ponderadores apropiados, la que determina la *función de memoria* de la red, responsable de aprender y almacenar los patrones al interior de ésta. Los ponderadores apropiados para la búsqueda del patrón de interés se calculan a partir de múltiples ajustes de éstos, modificaciones que se llevan a cabo a partir de varias repeticiones del proceso de funcionamiento de la red descrito anteriormente. Los signos de los ponderadores de cada uno de los inputs determinarán si éstos son de carácter *excitatorio* (ponderador positivo) o *inhibitorio* (ponderador negativo), estableciendo su aporte a la señal de salida que será enviada a la neurona de destino.

Para problemas de clasificación generalmente se utiliza el modelo de *perceptrón* construido por Rosenblatt [78], el que consiste en outputs provenientes de unidades sensoriales y que van a un conjunto de unidades de asociación previo a la etapa de suma. Estas unidades de asociación son las que entregan la principal propiedad a este tipo de modelo, ya que a diferencia de otros modelos que trabajan con ponderadores fijos [89], permiten a las unidades aprender a través del ajuste de dichos ponderadores.

Este modelo de Rosenblatt permite resolver problemas de clasificación donde las instancias son tanto linealmente separables, como aquellos problemas donde no lo son [56]. Para el caso donde el conjunto es linealmente separable, es decir, donde existe un línea recta o un hiperplano que pueda separar las instancias en sus categorías correctas, basta con que la red neuronal conste de una sóla capa compuesta por las unidades de entrada y por las unidades de salida, por lo que reciben el nombre de *perceptrones de una capa*. Para el caso donde las instancias no se puede separar linealmente, es necesario construir una red neuronal que conste de múltiples capas de neuronas perceptrones, ya que con una sola capa sería incapaz de clasificar las instancias correctamente. Esta estructura conocida como *perceptrones multicapa*, consiste en una serie de neuronas conectadas de manera de formar tres tipos de capas: unidades de entrada, que reciben la información a ser procesada; unidades de salida, donde los resultados del procesamiento son obtenidos; y unidades intermedias, conocidas como *unidades escondidas* [56].

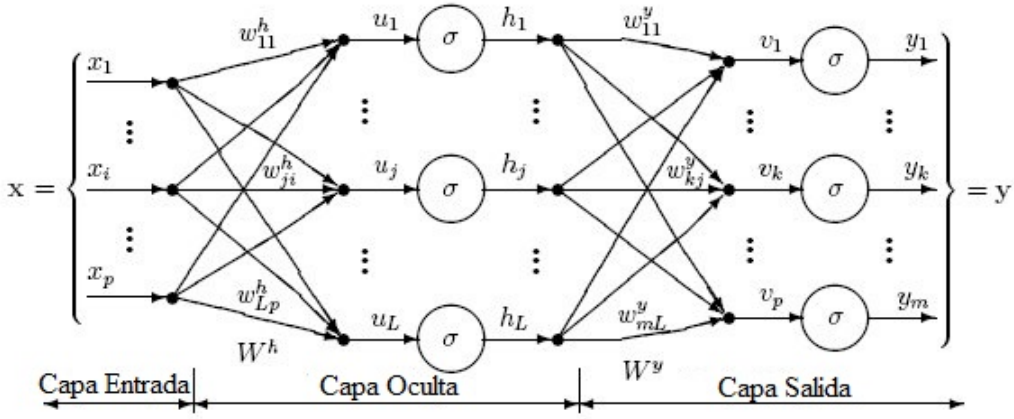


Figura 2.6: Estructura de una red neuronal con perceptrones multicapa

Durante el proceso de clasificación, la señal emitida por las unidades de entrada se propaga a través de toda la red con el fin de determinar el valor de activación de todas las unidades de salida. Para esto, cada unidad de entrada envía su valor de activación a cada unidad intermedia o escondida, las cuales calculan sus propios valores de activación, para luego ser enviados a las unidades de salida. Cada valor de activación es calculado por una *función de activación*, la cual, cómo se había descrito anteriormente, suma todas las contribuciones que recibe, donde la contribución de una unidad se define como el ponderador de la conexión entre la unidad emisora y la unidad receptora multiplicada por el valor de activición de la unidad emisora. Es justamente de estos ponderadores que depende el comportamiento de la red neuronal. Inicialmente, los ponderadores son fijados con valores *random*, para que luego

las instancias sean expuestas a la red repetidamente. Los valores de entrada correspondientes a los atributos de cada instancia son colocados en las unidades de entrada, y el output de la red es comparado con el valor de salida buscado para dicha instancia. Luego, todos los ponderadores de los inputs son ajustados de tal manera de que el output de la red se acerque al valor deseado.

Existen varios algoritmos a través de los cuales se puede realizar el ajuste iterativo de los ponderadores que permite el entrenamiento de una red neuronal [66]. Sin embargo, el más conocido y utilizado es el *Back Propagation* (BP) o *algoritmo de propagación hacia atrás*, el cuál incluye los siguientes pasos [56]:

1. Presentar un conjunto de entranamiento a la red neuronal.

2. Comparar el output actual de la red (s) con el output o *target* (b) buscado en el conjunto de entrenamiento. Calcular el error en cada neurona de salida.

3. Para cada neurona, calcular cuál debería haber sido el output y cuál es el ajuste necesario para hacerlo coincidir con el valor de salida esperado. Este ajuste corresponde al error local (δ).

4. Ajustar los ponderadores de cada neurona para minimizar el error local.

5. Asignar una "culpa" por el error local a las neuronas de la capa previa, entregando mayor responsabilidad a las neuronas conectadas con ponderadores más grandes.

6. Repetir los pasos anteriores sobre las neuronas de la capa previa, utilizando la "culpa" de cada una como su propio error.

A continuación se presentan las ecuaciones que describen la operación del modelo perceptrón de una neurona descrita en los pasos anteriores:

$$\text{Activación: } x = \sum_{\mathrm{i}=1}^{M} w_{\mathrm{i}} a_{\mathrm{i}} - \theta$$

$$\text{Señal de salida o output: } s = f(x)$$

$$\text{Error local: } \delta = b - s$$

$$\text{Ajuste de ponderadores: } \Delta w_{\mathrm{i}} = \eta \delta a_{\mathrm{i}}$$

donde η representa el parámetro que mide el ratio de aprendizaje. Un valor más grande permite al algoritmo moverse más rápido hacia la configuración de ponderadores que alcance el valor buscado, pero también incrementa la probablidad de nunca alcanzar el target.

El algotirmo de propagación hacia atrás debe realizar varias modificaciones de los ponderadores antes de alcanzar una configuración adecuada. Para n instancias de entrenamiento y W ponderadores, cada repetición o *época* en el proceso de aprendizaje toma $a(nW)$ tiempo, pero en el peor de los casos, el número de épocas puede ser exponencial al número de inputs. Es por esto que existen ciertas reglas de detención del proceso de entrenamiento, de las cuáles las cuatro más comunes son: i) Detenerse después de un número específico de épocas, ii) Detenerse cuando el error alcanza un umbral, iii) Detenerse cuando el error no presenta mejoras después de varias épocas, iv) Detenerse cuando el error del conjunto de validación está por sobre el error del conjunto de entrenamiento en alguna medida previamente determinada [56].

Múltiples aplicaciones en variadas industrias se han basado en la utilización de redes neuronales con perceptrón multicapa para construir modelos de clasificación, como por ejemplo modelos de predicción de quiebra de empresas [90] en la industria comercial; predicción del resultado de biopsias de próstata [30] o clasificación de señales de electroencéfalograma [42] en el campo de la salud; clasificación de imágenes [35], en particular de rostros [59], en el área de las ciencias y tecnología; construcción de modelos de *scoring* crediticio [**?**, 86] en la industria financiera; predicción de *churn* de clientes [47, 48] en la industria de las telecomunicaciones; y así un sin número de usos y problemas que pueden ser resueltos gracias a la utilización de esta técnica de clasificación.

2.2.3. Support Vector Machines

Support Vector Machines es un método de aprendizaje supervisado que resuelve problemas de clasificación basándose en la noción de *margen*, concepto que representa cada lado de un hiperplano que busca separar dos clases de datos (Figura 2.7). Se ha probado que maximizando este margen, y en consecuencia creando la distancia más grande posible entre el hiperplano separador y las instancias a cada lado de éste, es posible reducir el límite superior del error esperado [56].

Esta técnica de clasificación ofrece soluciones tanto para problemas donde las clases son linealmente separables como para aquellos donde no lo son, las que serán presentadas a continuación.

Caso linealmente separable

Si los datos de entrenamiento pueden ser separados de manera lineal en dos clases con las etiquetas $y_\text{i} = -1, 1$, suponemos que existe un hiperplano que es capaz de separar las instancias positivas de las negativas ("hiperplano separador"). Los puntos x que caen sobre el hiperplano satisfacen la ecuación $w \cdot x + b = 0$, donde w es normal al hiperplano, $\frac{|b|}{\|w\|}$ es la distancia perpendicular desde el hiperplano al origen, y $\|w\|$ es la norma Euclidiana de w. Sea d_+ (d_-) la distancia más corta desde el hiperplano separador a la instancia positiva (negativa) más cercana, se define el margen de éste como $\text{d}_+ + \text{d}_-$.

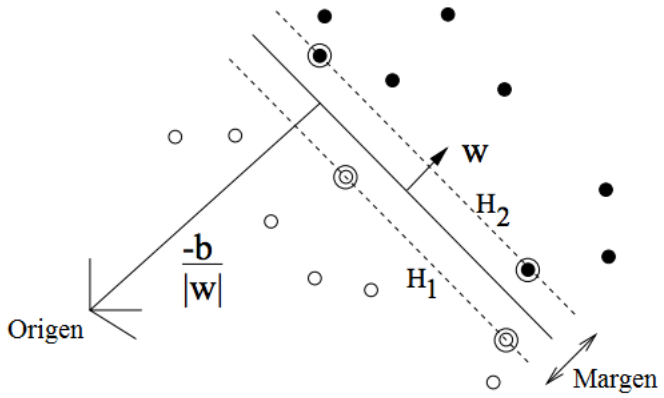


Figura 2.7: Hiperplano separador para el caso linealmente separable. Los vectores de soporte están indicados por círculos.

Para este caso linealmente separable, el algoritmo busca el hiperplano separador con el margen más amplio, lo que puede formularse como sigue.

Si se cuenta con data linealmente separable, luego un par *(w,b)* existe tal que:

$$x_{\mathrm{i}} \cdot w + b \geq +1 \text{ para } y_{\mathrm{i}} = +1$$

$$x_{\mathrm{i}} \cdot w + b \leq -1 \text{ para } y_{\mathrm{i}} = -1$$

con la regla de decisión dada por $f_{w,x}(x) = signo(x_{\mathrm{i}} \cdot w + b)$, donde w es llamado el vector de pesos y b el sesgo (-b es llamado el umbral o *threshold*).

Es fácil mostrar que cuando es posible separar linealmente dos clases, un hiperplano separador óptimo es encontrado minimizando el cuadrado de la norma de éste [13]. La minimización puede establecerse como un problema de programación cuadrática(QP):

$$min\ \tfrac{1}{2}\|w\|^2$$

$$s.a\ y_{\mathrm{i}}(x_{\mathrm{i}} \cdot w + b) \geq 1, \mathrm{i} = 1, ..., l.$$

En este caso de data linealmente separable, una vez que el hiperplano separador es encontrado, los puntos que permanecen en el margen son conocidos como *support vector points* o puntos de vectores de soporte, y la solución es representada por una combinación lineal que sólo incluye a estos puntos, ignorando el resto de ellos.

Aunque el margen máximo permite seleccionar entre muchos hiperplanos candidatos, para muchos conjuntos de datos el algoritmo SVM no es capaz de encontrar un hiperplano separador ya que los datos contienen instancias mal clasificadas. Este problema puede abordarse utilizando un *margen suave* que acepte algunas clasificaciones erróneas de las instancias de entrenamiento [56]. Esto puede llevarse a cabo introduciendo *variables de holgura* positivas ξ_i en las restricciones, con lo que se tiene:

$$x_i \cdot w + b \geq +1 - \xi \text{ para } y_i = +1$$

$$x_i \cdot w + b \leq -1 + \xi \text{ para } y_i = -1$$

$$\xi \geq 0$$

Así, para que un error ocurra el x_i correspondiente debe exceder la unidad, por lo que $\sum_i \xi_i$ es un límite superior del número de errores en el entrenamiento.

Sin embargo, la gran mayoría de los problemas del mundo real involucran data no separable para la cual no existe un hiperplano que separe exitosamente las instancias positivas de las negativas, por lo que hay que complejizar un poco el modelo para encontrar respuestas a este desafío.

Caso no linealmente separable

Una solución para el problema no separable es mapear los datos a un espacio de dimensiones superiores y definir un hiperplano separador allí. Este nuevo espacio es llamado *espacio de atributos transformados*, como opuesto al espacio de entrada ocupado por las instancias de entrenamiento.

Escogiendo apropiadamente un espacio de atributos transformados con suficiente dimensionalidad, cualquier conjunto de entrenamiento se puede hacer separable, ya que una separación lineal en este nuevo espacio corresponde a una separación no lineal en el espacio original. Para esto existe una función especial llamada *kernel* que permite al algoritmo trabajar directamente en el espacio de atributos sin necesidad de realizar el mapeo presentado anteriormente. Una vez que el hiperplano ha sido creado, la función kernel es utilizada para mapear nuevos puntos para clasificar dentro del espacio de atributos.

Esta técnica de clasificación se ha transformado en una de las más utilizadas en la actualidad, por lo que se pueden encontrar múltiples aplicaciones y problemas donde ha sido parte de la solución. Entre ellas podemos encontrar clasificación de textos [51], reconocimiento de expresiones faciales [64], análisis de genes [38], identificación de correos electrónicos *spam* [26], validación de tejidos cancerígenos [33], clasificación de suelos por medio de imágenes satelitales [49], entre muchísimas otras en lo más diversos campos de investigación.

2.3. Targeting de clientes

Ha sido probado que las estrategias de negocios que se concentran en encontrar y mantener buenos clientes tienen mayor probabilidad de generar alto valor para las compañías [84]. En este sentido, los esfuerzos por crear valor deben estar orientados en el *targeting* o selección y en la gestión de aquellos clientes que son más rentables o que tienen una rentabilidad potencial alta para la compañía. Lo anterior dice relación con las estrategias que puede aplicar el marketing directo, cuyos modelos se enfocan en identificar a los clientes que podrían responder a una solicitud específica -realizada por medio de una carta, correo electrónico o llamada telefónica- basándonse en la probabilidad estimada de que éstos respondan positivamente a una determinada campaña de marketing [55].

Según [87], para llevar a cabo un proceso de targeting de clientes exitoso, se deben definir las dimensiones más importantes que definen el proceso de selección. La primera dimensión a definir es el objetivo que se quiere alcanzar con este procedimiento, el que podría ser incrementar el valor actual del cliente, incrementar su valor potencial, disminuir el costo que genera cada uno, o reducir el riesgo asociado. Una vez el objetivo está establecido, la segunda dimensión apunta a determinar cuál es la información discriminante que define al cliente target y que lo diferencia del resto de los clientes, la que generalmente es derivada a partir de los atributos propios del cliente. Finalmente, y luego de que los clientes objetivo son seleccionados, se debe decidir qué hacer para alcanzar el objetivo planteado en la primera dimensión.

Como se menciona más arriba, la metodología para identificar a los clientes objetivo se basa en la extracción y análisis de la información discriminante que se logra obtener de cada uno de ellos, la cual tiene su origen en datos demográficos, psicográficos, de uso, de comportamiento, de valor y/o de necesidades. De acuerdo a [55], es posible extraer esta información discriminante a través de diversas herramientas de minería de datos, sin embargo para este caso particular, es necesaria la utilización de técnicas de clasificación que en vez de asignar a cada cliente una etiqueta de una de las clases posibles, calcula un *score* o puntaje para cada uno, el que representa la probabilidad que cada cliente tiene de pertenecer a cada una de estas clases. De esta manera, a partir de estos puntajes es posible seleccionar a los clientes que presentan las probabilidades más altas de responder positivamente al ofrecimiento realizado por la compañía (clase positiva), es decir, los clientes target.

El targeting de clientes ha sido ampliamente utilizado en variadas aplicaciones como la predicción de clientes interesados en comprar un seguro vehicular, la cual se desarrolla en base a la construcción de redes neuronales [55]; análisis de desgaste de clientes que potencialmente podrían fugarse de un banco de retail [46], el que se estudia a través del uso de modelos Bayesianos, árboles de decisión y también redes neuronales; identificación de clientes adecuados para ofrecer un cross-selling en la industria financiera [53]; diseño de un modelo de predicción que permitiera seleccionar a los clientes más propensos a comprar un automóvil Mercedes Benz [34], entre tantos otros desafíos cuyas soluciones fueron propuestas en base a este enfoque.

2.4. Análisis de Redes Sociales

Los registros de llamadas telefónicas móviles pueden ser estudiadas y analizadas por medio de diversas técnicas, entre las cuales se pueden encontrar *Online Analytical Processing* o Análisis OLAP [14], Análisis Estadístico [39] y *Social Network Analysis* o Análisis de Redes Sociales (SNA) [83], donde esta última, tal como señalan Faust y Wasserman [83], se basa en la identificación y estudio de las relaciones que vinculan a las entidades que interactuan al interior de la sociedad, donde estas entidades o actores sociales pueden representar personas, empresas, instituciones, países, etc.

De acuerdo con Martino y Spoto [63], existe un consenso general respecto de los orígenes de SNA, siendo los primeros estudios realizados durante la década de 1930 por Jacob Moreno, quien creó el concepto de *sociometría* [10], el cual estudia las relaciones interpersonales, y por Fritz Heider, quien construyó lo que hoy se conoce como la *teoría del balance* [43]. Estas ideas fueron desarrolladas por primera vez por Frank Harary y Dorwin Cartwright, quienes usando métodos de la entonces recién lanzada teoría de grafos [41] diseñaron una herramienta para el análisis de estructuras sociales [15]. Sin embargo, recién en 1954 Barnes [3] acuña el concepto de red social, quien la define como *un conjunto de algunos puntos (nodos) que se vinculan por líneas para formar redes totales de relaciones. La esfera informal de relaciones interpersonales se contempla así como una parte, una red parcial de una total.*

Para De Nooy [24], el análisis de redes sociales tiene como principal objetivo la detección e interpretación de patrones en los vínculos entre actores. Para alcanzar este objetivo, Nooy plantea que existen cuatro grandes áreas de investigación dentro de SNA:

- **Cohesión:** Investiga los casos de individuos que están relacionados y de aquellos que no lo están. Además estudia las razones del por qué en ciertos casos no existe un vínculo entre dos personas. La principal hipótesis plantea que quiénes coincidan en características sociodemográficas interactuarán con mayor frecuencia y que personas que interactúen regularmente fomentarán una actitud o identidad común.

- **Intermediación:** Estudia la redes sociales como estructuras que permiten el intercambio de información y analiza como ésta se difunde por la red como parte de un sistema social.

- **Ranking:** Analiza el prestigio y la posición de los actores en base a la importancia que tienen dentro de una red.

- **Roles:** Busca estudiar los roles de los participantes de una red, basándose en el análisis de los patrones existentes en sus relaciones.

El estudio de los roles se ha convertido en la principal herramienta utilizada para analizar el comportamiento del ser humano y para entender cómo esta conducta se ve afectada por las relaciones que establece con sus pares. Bajo esta mirada, la visión que entrega SNA ha sido aplicada en diversos campos de estudio, como el comportamiento individual de las personas [85], el comportamiento colectivo de grupos [58], flujos de tráfico social [71], el marketing relacionado con el incremento de la venta de productos y servicios [25], análisis y prevención del crimen [88], análisis y detección de patrones en la transmisión de enfermedades infecciosas [22], diseño de protocolos de redes inalámbricas [54], entre otras muchas aplicaciones.

2.4.1. Redes y Teoría de grafos

Esta sección presentará algunas nociones y conceptos básicos acerca de la representación de redes a través de grafos, basándose en la teoría de grafos y en el estudio de la estructura de redes.

De acuerdo a [28] un *grafo* es una forma de especificar relaciones entre un conjunto de ítems, la cual tiene dos componenetes fundamentales:

- **Nodos:** Pueden representar cualquier objeto de estudio. En particular, en una red social representan miembros o conjuntos de miembros.

- **Arcos:** Representa una relación entre un par de nodos. Para el caso de las redes sociales, puede representar un tipo de relación como grado de amistad, parentesco, nivel de interacción, etc.

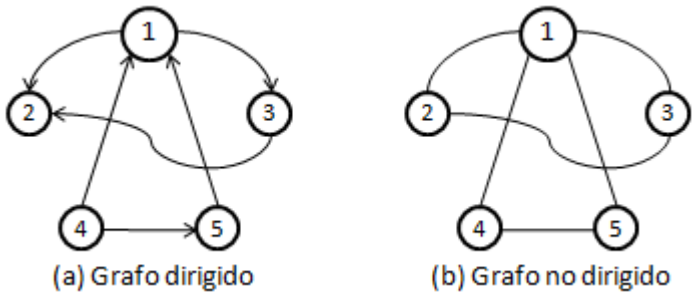


Figura 2.8: Tipos de grafos.

Un arco es denominado *arco dirigido* cuando representa una relación con dirección o jerarquía, como por ejemplo "A envía un mensaje a B" o "A es jefe de B"; mientras que es denominado *arco no dirigido* en caso de una relación bidireccional, como por ejemplo "A va a la misma clase que B". Un arco no dirigido puede construirse a partir de la combinación de dos arcos dirigidos en sentidos opuestos. Así, un grafo que tiene sólo arcos dirigidos se define como un *grafo dirigido*, y uno sólo con arcos no dirigidos se define como un *grafo no dirigido*.

Formalmente, un grafo **G** puede expresarse como **G(N,A)** donde **N** = (n_1,...,n_l) corresponde al conjunto de nodos y **A** = (a_1,...,a_k) al conjunto de arcos. Un grafo se dice no dirigido si $a_k = (n_\mathrm{i}, n_j) = (n_j, n_\mathrm{i}) \forall (n_\mathrm{i}, n_j) \in \mathbf{A}$, y dirigido si esta condición no se cumple. Otras definiciones útiles son:

- **Nodos adyacentes o vecinos:** Un nodo n_i se dice adyacente o vecino del nodo n_j si $\exists a_k \in \mathbf{A}$, tal que $a_k = (n_\mathrm{i}, n_j)$, es decir, que exista un arco entre ellos.

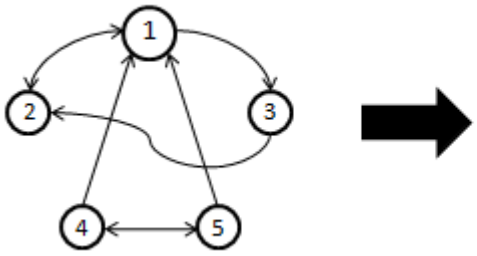
- **Camino:** Un camino es una secuencia de nodos en la cual cada par de nodos consecutivos son nodos adyancentes.

- **Tríada:** Conjunto de tres nodos en que todos son adyancentes entre ellos.

- **Subgrafo:** Se dice que **G′ = (N′,A′)** es un subgrafo de **G = (N,A)** si **N′** $\subseteq$ **N** y **A′** $\subseteq$ **A**.

- **Grafo completo:** Se dice que un grafo es completo cuando todos sus nodos son adyancentes entre ellos. Un grafo completo de n nodos tiene exactamente $\frac{n(n-1)}{2}$ arcos.

Una manera muy útil de representar un grafo es a través de la *matriz de adyacencias*, la cual contiene tantas filas y columnas como actores tenga el grafo (Figura 2.9). Formalmente, si tenemos un grafo **G(N,A)**, la matriz de adyacencias se define como una matriz A de $N \times N$ donde $A_{\mathrm{i},j} = 1$ si $(\mathrm{i}, j) \in \mathbf{A}$ y $A_{\mathrm{i},j} = 0$ en caso contrario.



	x1	x2	x3	x4	x5
x1	0	1	1	0	0
x2	1	0	0	0	0
x3	0	1	0	0	0
x4	1	0	0	0	1
x5	1	0	0	1	0

Figura 2.9: Representación de una grafo en una matriz de adyancencia.

2.4.2. Métricas utilizadas en SNA

A continuación se presentan algunas de las métricas más importantes y más utilizadas en la teoría de grafos y en el estudio de redes sociales.

- **Grado de un nodo:** Existen dos tipos de grados en un nodo n_k. El *in-degree* o *grado entrante* que corresponde al número de arcos incidentes $\mathrm{d}_k^{\mathrm{i}}$, y el *out-degree* o *grado saliente* que representa el número de arcos que salen d_k^o. Resulta importante notar que en un grafo no dirigido $\mathrm{d}_k^{\mathrm{i}} = \mathrm{d}_k^o = \mathrm{d}_k$.

- **Densidad:** La densidad de un grafo refleja qué tan entrelazada está en una red. Ésta se mide a través del número de arcos en la red sobre todos los posibles arcos existentes. En un grafo dirigido se calcula como $D = \frac{m}{n(n-1)}$, y en un grafo no dirigido como $D = \frac{2m}{n(n-1)}$.

- **Vecindario:** El vecindario de un nodo n_k se define como el conjunto de nodos que se conecta con n_k.

- **Clique:** Un clique es el máximo subgrafo completo compuesto de al menos tres nodos. Una tríada es el mínimo elemento que puede ser llamado clique.

2.5. Targeting social de clientes

Como dijimos antes, las campañas de marketing directo que se desarrollan en base a targeting de clientes apuntan a seleccionar a aquellos consumidores que son potencialmente más rentables y enfocar la campaña sólo en ellos. Sin embargo, esta estrategia presenta una limitación básica pero muy importante: trata a cada cliente como si tomara decisiones independientemente de los demás clientes [25]. En la realidad, la decisión de una persona de comprar o adquirir un producto usualmente está fuertemente influenciada por sus amigos, conocidos, familiares, compañeros de negocios, etc. Esta influencia entre el cliente y su red social es la idea principal detrás del concepto de *targeting social*, el cual depende de los atributos que puedan desprenderse de la configuración de dicha red, siendo representado a través del *valor social* que cada suscriptor posee.

Campañas de marketing basadas en las redes sociales del *boca a boca* pueden ser mucho más efectivas en términos de costos que las campañas tradicionales, ya que los mismos clientes son los que realizan gran parte del esfuerzo promocional.

En [45] se describen tres estrategias complementarias en que el targeting de clientes basado en SNA puede enfocarse para lograr sus objetivos de marketing:

- **Recomendación explícita:** Los clientes son partidarios o defensores vocales del producto o servicio, recomendándolo a sus amigos o conocidos. Un ejemplo de esto es el éxito de un libro que fue entregado de manera gratuita a 10.000 lectores *influyentes* (e.g vendedores de libros) con el fin de estimular la venta del mismo.

- **Recomendación implícita:** Incluso si los individuos no hablan sobre un producto, estos pueden recomendarlo implícitamente a través de sus acciones, especialmente a través de su propia adopción del producto. Un ejemplo de esto puede ser el esfuerzo que hacen las empresas para que atletas famosos utilicen sus marcas, influenciando a sus seguidores a utilizarlas también, efecto que tratan de replicar en pequeños grupos de personas, intentando convencer al miembro "líder" para que adopte el producto.

- **Targeting de redes:** Las firmas tratan de enfocar la campaña de marketing sobre los vecinos de la red social de los clientes que adoptaron previamente el producto, incluso si no existió ninguna recomendación por parte de él. Un ejemplo de esto fue la estrategia utilizada por Hotmail donde los correos enviados por un usuario tenían un pie de página que invitaba al receptor del mensaje a adoptar gratuitamente el producto.

Existen variadas maneras de abordar el problema de targeting social de clientes, cuyo desafío principal radica en encontrar la mejor manera de incorporar el valor social de cada cliente en el proceso de identificación del segmento objetivo. Una de ellas se enfoca en la idea de que combinando los atributos del usuario con los atributos sociales que se puedan extraer de la estructura de su red social se logran construir modelos con mayor poder de predicción en términos de la identificación de futuros adoptadores de un producto [2, 8], teoría que se prueba en una red social de mensajería instantánea, tanto para la adopción de un producto como para el interés común sobre un anuncio publicitario.

Otra vía para evaluar la importancia de la red social de cada cliente en sus decisiones corresponde a la noción de *valor de adopción social*, el cual deriva del comportamiento de los individuos que conforman su red de contactos, ya que la decisión de adoptar un producto por parte de un cliente generalmente viene influenciada por las decisiones de sus amigos, cercanos, compañeros de trabajo, etc.

Existen dos enfoques diferentes para calcular el valor de adopción social de un cliente. Por un lado está el método propuesto en [25], donde se afirma que este valor deriva de la influencia que el cliente pueda tener sobre otros clientes en el futuro, y por otro la mirada que se plantea en [8], la cual propone la existencia de una influencia sobre un individuo de parte de sus amigos que adoptaron previamente el producto en cuestión (predecesores), lo que genera una *presión social* que lo puede llevar a adquirir el mismo producto.

Ambas maneras de plantear el tema del valor social advierten que si éste no es tomado en consideración cuando se realiza el targeting de clientes se pueden llegar a resultados subóptimos en la predicción de futuros adoptadores. En este aspecto, un cliente puede no ser seleccionado para una campaña de marketing directo ya que su valor individual resulta ser menor que el costo de la misma, sin embargo, si es tomando en cuenta su valor social puede quedar por sobre el costo, revertirse la decisión y entrar en la selección, por lo que es muy relevante este concepto para el targeting de clientes.

En esta memoria se considerará tanto la inclusión de atributos sociales como el valor de adopción social en la construcción de los modelos, apuntando a investigar si es posible lograr una mejora en la selección de clientes que se planteará más adelante en el capítulo 4.

Capítulo 3

Metodología

La metodología de trabajo escogida corresponde a una adaptación del proceso jerárquico *Cross Industry Standard Process for Data Mining* (CRISP-DM) [17] para SNA, similar a las desarrolladas por Cuadra [61] y Muñoz [65] con metodologías similares como *Knowledge Discovery in Databases* (KDD). Este modelo da una mirada global al ciclo de vida de un proyecto de minería de datos, describiendo las diferentes fases del proyecto, sus respectivas tareas (tanto genéricas como específicas) y la relación que se establece entre ellas.

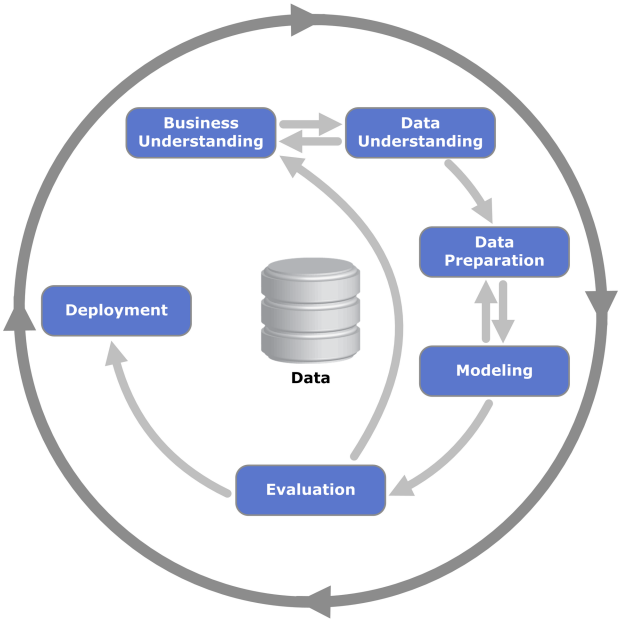


Figura 3.1: Metodología CRISP-DM.

Generalmente, el ciclo de un proyecto de minería desarrollado bajo la metodología CRISP-DM consta de seis fases, sin embargo esta adaptación para SNA que se está proponiendo requerirá de la inclusión de nuevas etapas, las que se describirán a continuación:

1. **Entendimiento de la Problemática**
   La fase inicial se enfoca en entender el problema y determinar los objetivos y requerimientos de la investigación desde una perspectiva de Inteligencia de Negocios, con el fin de transformar este conocimiento en un problema de minería de datos. En esta etapa es importante evaluar la situación inicial especificando los recursos con que se cuenta; diseñar un plan de trabajo que especifique los diferentes pasos a seguir; definir los alcances de la investigación; e indicar los resultados que se esperan obtener.

2. **Estudio y Comprensión de los Datos**
   En esta fase se da por comenzado el trabajo con la data, donde el primer paso corresponde a la recopilación inicial de las bases de datos con que se trabajará, para luego realizar una descripción de las características de éstas y evaluar si es que cumplen con los requerimientos de la investigación. Posteriormente se lleva a cabo una exploración simple de los datos para lograr familiarizarse con ellos y para determinar la calidad de los mismos.

3. **Preparación inicial de los Datos**
   Esta fase cubre todas las actividades necesarias para construir las bases de datos definitivas a partir de los datos obtenidos inicialmente. Para adaptar los *datasets* a las técnicas de modelamiento a utilizar, es imprescindible seleccionar las variables claves para la investigación y determinar la importancia de cada una de ellas dentro del modelo; limpiar las bases de datos eliminando *outliers* e imputando datos faltantes; y transformar los datos de manera adecuada, creando funciones y nuevas variables.

4. **Configuración de la Red Social**
   En esta etapa se diseña y estructura la red social de teléfonos móviles que representa las relaciones existentes entre los usuarios y la importancia de cada una de éstas. Esta red será posteriormente la fuente desde donde se extraerán los atributos sociales de cada cliente que serán incluidos para complementar las bases de datos de los modelos de targeting.

5. **Modelo de Targeting Social**
   En esta etapa se analiza el efecto que tiene la inclusión de atributos sociales en los modelos de targeting tradicional. Además se realiza una comparación de la capacidad de predicción de los modelos diseñados en base a distintas técnicas de clasificación, cuyos conjuntos de entrenamiento fueron configurados bajo diferentes estructuras temporales.

6. **Valor de Adopción Social de los Clientes**
   En base al papel que pueden jugar los amigos o cercanos de un cliente en su decisión sobre adoptar o no un producto, en esta etapa se busca identificar y calcular una componente social que permita estimar de mejor manera y desde otro enfoque la probabilidad de que un cliente finalmente adopte el producto ofrecido.

7. **Evaluación**
   Una vez obtenidos los *outputs* de los modelos, se utilizan diferentes métricas de evaluación que permiten rescatar una serie de conclusiones a partir de los resultados entregados y definir si éstos calzan con los objetivos de la investigación. Adicionalmente, se busca establecer qué técnica de clasificación es la más adecuada para realizar una investigación con este tipo de información y qué estructura del modelo de targeting es el que mejor desempeño muestra.

8. **Despliegue**
   Esta fase final tiene como objetivo resumir y organizar todo el conocimiento generado a partir de la investigación. Se busca la manera más sencilla y adecuada de describir el proceso de estudio, de presentar los resultados obtenidos y de desplegar la data final utilizada, de manera de que frente a un traspaso de información sea fácil entender cada una de las etapas desarrolladas durante la investigación.

3.1. Selección de los Datos

La gran mayoría de la información que las compañías de telecomunicaciones logran recopilar de sus clientes proviene del uso del servicio de telefonía móvil que estos contratan. Esta información puede clasificarse tanto en datos sociodemográficos como en datos de comportamiento comercial. Dentro los sociodemográficos se puede encontrar información personal del cliente, tal como la edad, género, lugar de residencia y el nivel socioeconómico; mientras que en los de comportamiento comercial, se pueden encontrar datos como la antigüedad del cliente en la compañía, cantidad de minutos utilizados, número de mensajes de texto enviados, el nivel de uso del servicio de internet y el monto de facturación.

Los datos sociodemográficos son usualmente almacenados en bases de datos al momento en que el cliente contrata el servicio y se van actualizando en contadas ocasiones, sólo cuando el cliente modifica algún dato personal. Por otro lado, la información transaccional correspondiente a los datos relacionados al tráfico de llamadas y mensajes -que son los que reflejan el comportamiento comercial del cliente- son almacenados en archivos conocidos como *Call Detail Records* o CDRs, los cuales graban detalles como los teléfonos de origen y destino de llamadas o mensajes, la duración de una llamada y el lugar desde donde ésta se realizó [73], los que posteriormente son consolidados y transformados en atributos comerciales del cliente correspondientes a un cierto periodo de tiempo.

Según [27], una red de llamados telefónicos móviles puede ser representada como una red social tradicional, lo permite que a través de diversas aplicaciones de SNA se puedan extraer una serie de *atributos sociales* que reflejan la manera en que los clientes se relacionan con sus amigos, familiares o conocidos a través de la utilización de sus telefónos móviles. Estos datos sociales permiten complementar de manera importante los datos sociodemográficos y comerciales, ya que al entregar información no explícita genera oportunidades para realizar nuevos estudios y análisis en el ámbito de las telecomunicaciones y en particular de la telefonía móvil.

Lo que esta memoria busca investigar es cómo la inclusión de estos atributos sociales en los inputs de los modelos de clasificación modifica la capacidad de predicción de éstos en relación a la etiqueta que se debe asignar a cada cliente respecto a la respuesta que éste tenga frente a la oferta de un producto por parte de la compañía.

3.2. Preprocesamiento de los Datos

De acuerdo a [40], las bases de datos utilizadas en la actualidad son muy susceptibles a presentar ruidos, datos faltantes e información inconsistente debido principalmente a su alto volumen y a las múltiples fuentes desde donde son extraídas, lo que significa un problema importante ya que generalmente datos de baja calidad generan resultados de baja calidad. A continuación se presentan diversas técnicas de preprocesamiento de datos que permiten mejorar la calidad de éstos en términos de los elementos que la definen: precisión, integridad y consistencia.

3.2.1. Limpieza de los Datos

El proceso de limpieza de los datos apunta a completar datos faltantes o *missing values*, suavizar ruidos a medida que se identifican valores fuera de rango o *outliers*, y corregir inconsistencias de la data.

- **Datos faltantes**
  Para generar un buen modelo de clasificación es clave contar con un gran volumen de datos, por lo que es importante disponer de la información de un número elevado de clientes correspondiente a todos los atributos que caracterizan a cada uno de ellos: sociodemográficos, comerciales y sociales. Para esto es necesario contar con estrategias que permitan reemplazar una proporción pequeña de datos vacíos o en blanco que no conviene descartar, ya que conllevaría la eliminación de clientes de las bases de datos de entrada de los modelos. Entre las estrategias para completar datos se pueden encontrar ignorar la tupla, completar manualmente, usar una constante, usar una medida de tendencia central -por ejemplo el promedio o la mediana-, y usar el valor más probable.

- **Datos ruidosos**
  Un *ruido* corresponde al error o varianza aleatoria presente en una variable que hace que su valor no esté dentro de los rangos comunes de la base. Con esta estrategia se busca *suavizar* la totalidad de los datos con la finalidad de eliminar por completo el ruido que pueda existir en el input de un modelo. Algunas técnicas que permiten llevar a cabo este proceso son el suavizamiento de datos en función del vecindario, regresiones y análisis de outliers.

3.2.2. Integración de los Datos

Los proyectos de minería de datos generalmente necesitan combinar muchos datos que vienen de múltiples fuentes, proceso que debe realizarse cuidadosamente para evitar redundancias e inconsistencias en la información que se utilizará como entrada de los modelos, siendo los principales retos la heterogeneidad semántica y la estructura de la base de datos.

- **Problema de identificación de entidades**
  Durante el proceso de integración puede presentarse la necesidad de combinar datos de un mismo cliente provenientes de diferentes fuentes, desafío que requiere de la identificación del mismo en todas las fuentes desde donde se extrae la información -puede ser a través de un número o código de cliente- que permita realizar un *match* de todos los atributos correspondientes a dicho cliente, de manera de generar una sola gran base de datos integrada.

- **Redundancia y análisis de correlación**
  Cuando se requiere integrar datos que poseen diferentes orígenes, la *redundancia* se presenta como un importante asunto a considerar. Un atributo puede ser redundante si puede ser derivado de otro atributo o conjunto de atributos, lo que puede llevar a obtener resultados redundantes. Algunas redundancias pueden ser detectadas a través de un análisis de correlación, el que es capaz de medir la intensidad con que un atributo implica a otro. Para datos nominales se utiliza el test χ^2 (chi-cuadrado), mientras que para variables numéricas se usa el coeficiente de correlación y la covarianza.

- **Duplicación de Tuplas**
  Adicional a la identificación de redundancias entre atributos, la duplicación de datos puede ser detectada a nivel de tuplas, pudiendo existir dos o más tuplas para un único cliente, hecho que podría generar aún más inconsistencias en los datos. Por ejemplo, una discrepancia se da cuando en una base de datos de compras, un cliente aparece con dos domicilios diferentes.

3.2.3. Reducción de los Datos

Las bases de datos que se logran construir en la industria de las telecomunicaciones suelen ser de volumenes muy grandes, por lo que el tiempo que toma realizar análisis y minería de datos tiende a ser mucho más largo de lo común, haciendo a veces impracticables o inviables estos análisis. Ciertas técnicas de *reducción de datos* pueden ser aplicadas con el objetivo de representar los datos en volumenes más pequeños, sin la necesidad de perder la integración de la data original, obteniendo resultados similares a partir de procesamientos más eficientes. Entre ellas se encuentran la reducción de dimensionalidad, reducción de numerosidad y compresión de los datos.

- **Reducción de dimensionalidad**
  Es el proceso de reducir el número de variables o atributos que se consideran como parte del input de los modelos. Dentro de los métodos existentes para reducir dimensionalidad, se encuentran la transformaciones de wavelet y el análisis de componentes principales (ACP), los cuales transforman o proyectan los datos originales sobre un espacio de dimensionalidad más pequeño.

- **Reducción de numerosidad**
  Esta estrategia busca reemplazar el volumen de datos original por representaciones alternativas más pequeñas de la data, las que pueden ser construidas de manera paramétrica como no paramétrica. En el caso de métodos paramétricos, un modelo es utilizado para estimar los datos, por lo que en vez de almacenar toda la data actual, sólo se encarga guardar los parámetros de ésta, siendo la regresión y los modelos loglineales ejemplos de este caso. Los métodos no paramétricos que permiten almacenar representaciones reducidas de los datos incluyen histogramas, *clustering*, *sampling* y *data cube aggregation*.

- **Compresión de datos**
  En este caso se aplican transformaciones con el fin de obtener una representación reducida o comprimida de los datos originales. Si la data original puede ser reconstruida a partir de la comprimida sin perder información, la reducción de datos se llama "sin pérdida". Por el contrario, si sólo se puede construir una aproximación de la data original, esta es llamada "con pérdida".

3.2.4. Transformación de los Datos

En esta etapa del preprocesamiento los datos son transformados o consolidados a formas apropiadas para lograr que el proceso de minería resultante sea más eficiente y que los patrones encontrados sean más fáciles de enteder. Entre las estrategias existentes para transformar datos se encuentran:

1. **Suavizamiento,** la cual trabaja removiendo ruidos de la data. Dentro de las técnicas se incluyen la regresión y clustering.

2. **Construcción de atributos,** donde nuevas variables son construidas y adicionadas a la base de datos de entrada de los modelos para ayudar al proceso de minería y análisis.

3. **Agregación,** donde se resumen o agregan datos con el fin de poder manejarlos de manera consolidada.

4. **Normalización,** donde los datos de los atributos son escalados de manera de que caigan dentro un rango más pequeño, como -1 a 1, o 0 a 1.

5. **Discretización,** donde los valores brutos de una variable numérica (e.g., *edad*) son reemplazados por etiquetas de intervalos (e.g., 0-10, 11-20, etc.) o etiquetas conceptuales (e.g., *joven*, *adulto*, *anciano*).

3.2.5. Balanceo de los Datos

El *problema de clases desbalanceadas* se da típicamente cuando en un modelo de clasificación hay muchas más instancias de unas clases que de otras, teniendo casos donde el ratio de clase minoritaria a clase mayoritaria puede llegar a ser 1 es a 1.000 o 1 es a 10.000. En estos casos, los clasificadores estándar tienden a estar agobiados por las clases más grandes por lo que ignoran a las más pequeñas. Particularmente, éstos tienden a producir mayor precisión en la predicción en la clase mayoritaria, pero una precisión muy baja en la clase minoritaria. Este problema se da en muchas aplicaciones como la detección de fraude o intrusión, gestión de riesgo, diagnósticos y monitoreos médicos, entre muchos otros casos [19].

Algunas soluciones para el problema de clases desbalanceadas han sido propuestas tanto a nivel de datos como a nivel algorítmico. A nivel de datos [20], estas soluciones incluyen varias formas diferentes de *re-sampling* o remuestreo, tales como *over-sampling* o sobremuestreo y *under-sampling* o submuestreo, y técnicas que modifican la probabilidad a priori de las clases mayoritaria y minoritaria en el conjunto de entrenamiento para obtener un número más balanceado de instancias en cada clase. El método de under-sampling extrae un conjunto más pequeño de las instancias mayoritarias manteniendo todas las instancias minoritarias, lo que es adecuado para aplicaciones de gran escala donde el número de muestras mayoritarias es muy grande, ya que al disminuir el tamaño del conjunto de entrenamiento se reduce el tiempo de procesamiento haciendo el problema de aprendizaje mucho más manejable. En contraste, el método de over-sampling incrementa el número de instancias minoritarias realizando un sobresampleo en ellas.

A nivel de algoritmos [19], las soluciones incluyen el ajuste de los costos de varias clases con el fin de contrarrestar el desbalanceo de clases cuando se entrena el modelo; el ajuste de la estimación probabilística en las hojas cuando se utilizan árboles de decisión; y el ajuste del umbral de decisión.

SMOTE

Synthetic Minority Over-sampling Techinique (SMOTE) es un enfoque de resampleo propuesto y diseñado en [18], el que genera ejemplos sintéticos de una manera genérica independiente de la aplicación. La clase minoritaria es sobresampleada tomando pequeños vecindarios de instancias minoritarias e introduciendo ejemplos sintéticos a lo largo del segmento de línea

que une algunos o todos los k vecinos más cercanos. Dependiendo de la magnitud del over-sampling requerido, los k vecinos más cercanos son escogidos aleatoriamente, sin embargo la implementación más típica utiliza cinco instancias para cada vecindario. Por ejemplo, si se requiere un sobresampleo de un 200 %, sólo dos de los cinco vecinos son escogidos y una instancia sintética es creada entre ellas.

Las instancias sintéticas son creadas de la siguientes forma: se toma la diferencia entre el vector de atributos de la instancia en consideración y su vecino más cercano; se multiplica esta diferencia por un número aleatorio entre 0 y 1, y ésta se añade al vector de atributos en consideración. Esto causa la selección de un punto aleatorio a lo largo del segmento de línea entre dos atributos específicos (Figura 3.2).

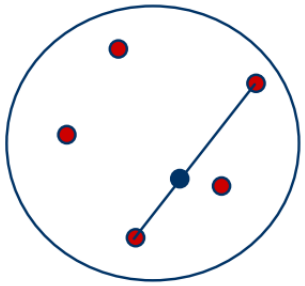


Figura 3.2: Over-sampling a través de SMOTE. La clase minoritaria es sobresampleada tomando las instancias minoritarias e introduciendo ejemplos sintéticos (círculos azules) a lo largo del segmento de líneas que uno los k vecinos más cercanos de la vecindad (círculos rojos).

3.3. Configuración de la Red

Para construir la red social de teléfonos móviles es necesario tener en consideración las interacciones que se producen entre los clientes a través de sus celulares, ya sea por medio de llamadas telefónicas o por mensajes de texto. Ya establecida esta noción, la red estará configurada de la siguiente manera: los *nodos* representarán a los clientes o usuarios y los *arcos* representarán a las interacciones o relaciones entre ellos. Adicionalmente cada arco llevará asociado un *peso* que representará la importancia de la relación establecida entre dos usuarios, el cual generalmente se calcula en base a la actividad telefónica entre los involucrados.

En la literatura existen dos estructuras que permiten generar este tipo de redes telefónicas [69]. La primera de ellas se enfoca en la construcción de una *red no recíproca* donde dos usuarios están conectados con un arco no dirigido si es que se produce al menos una llamada telefónica entre ellos, es decir, i llama a j o j llama a i. La segunda se basa en la construcción de una *red recíproca*, en la cual dos usuarios se conectan con un arco no dirigido si al menos hubo un par de llamadas recíprocas entre ellos, es decir, i llama a j y j llama a i. En esta memoria se utilizará el enfoque de *red no recíproca*.

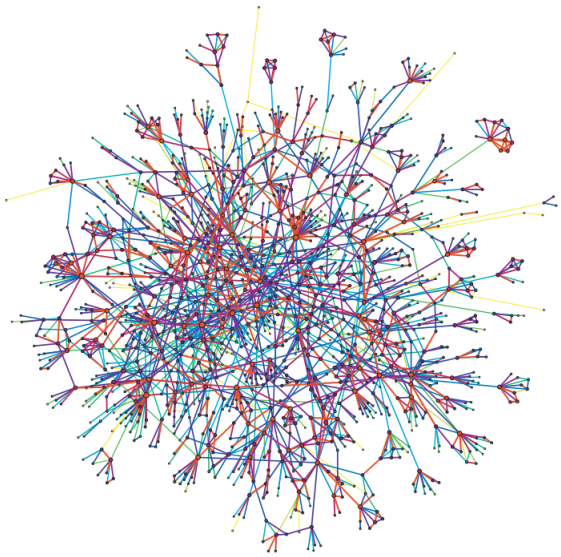


Figura 3.3: Estructura de una red social de teléfonos móviles

El hecho de que una red de llamadas telefónicas pueda ser representada por una red social ha sido estudiado en [27, 69], en la cual un teléfono móvil representa una entidad social y el tráfico de llamadas el vínculo entre usuarios, permitiendo que a través de el uso de SNA se pueda analizar la data de tráfico de llamadas. Bajo esta mirada, Baruah y Angelov [5] han logrado identificar miembros clave cuantificando la significancia de cada cliente en la red y han estudiado como ésta evoluciona en el tiempo. Así mismo Xu et al.[cita] se han enfocado en la búsqueda de comunidades en base a la influencia entre miembros de la red. Por otro lado, Hidalgo y Rodriguez- Sickert [44] han investigado empíricamente la dinámica de una red de teléfonos móviles a través del estudio de la persistencia de los vínculos entre las personas, complementando el trabajo que Onnela y Saramäki [69] han realizado al analizar la estructura de la red y la fuerza de los lazos existentes entre los usuarios de teléfonos móviles.

3.4. Modelo de Targeting de clientes

Para llevar a cabo el proceso de targeting de clientes (Figura 3.4), primero se deben escoger al menos dos técnicas de clasificación -acorde al conjunto de datos disponible- de manera de hacer un comparación entre los resultados finales obtenidos por cada uno. Una vez se decide qué técnicas se usarán, se debe determinar qué ventanas de tiempo se desean evaluar, estableciendo el mes de datos que será utilizado como base de entrenamiento y sobre qué mes se desea realizar la predicción. Se recomienda que se prueben distintas ventanas de tiempo para entender cómo influye el espacio de tiempo existente entre entrenamiento y prueba en la precisión del modelo.

Luego de definidas la técnica a utilizar y la ventana de tiempo a evaluar, se ejecuta el algoritmo de clasificación, el cual extrae información predictiva desde el conjunto de entrenamiento y aprende patrones a partir de los datos. Ya aprendidos ciertos patrones, el modelo

entrenado se evalúa sobre la base de prueba definida previamente. Esta evaluación entrega una predicción para cada una de las instancias, la que puede ser representada a través de una etiqueta que refleja a qué clase pertenece o unos *scores* que corresponden a la probabilidad de pertenencia a cada una de las clases. Si se toma el camino de las etiquetas categóricas, se evalúa directamente la precisión del modelo a través de lás métricas que se especificarán en la sección 3.7.

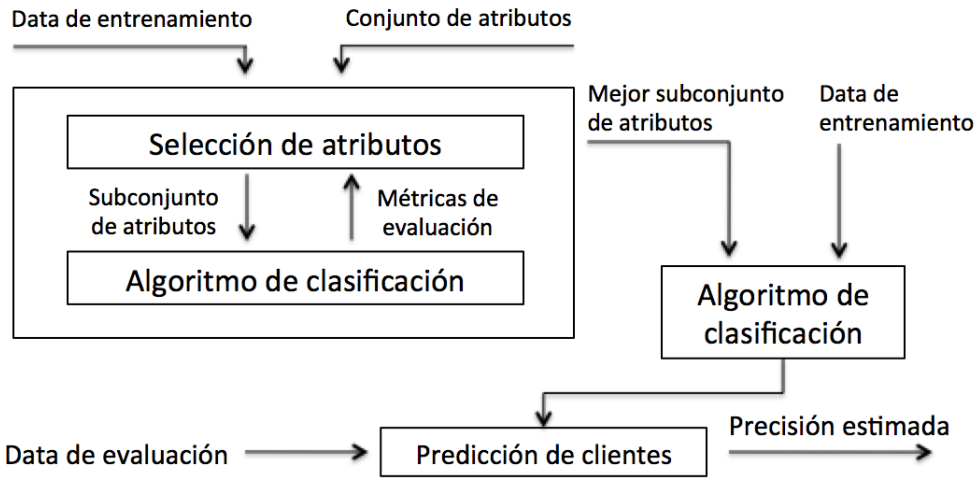


Figura 3.4: Estructura del modelo de Targeting de clientes.

Al contrario, si se opta por el enfoque de los scores [55], se debe construir un *ranking* en orden decreciente en función de la probabilidad de pertenecer a la clase de interés (e.g. mayor probabilidad de *churn*, de adopción de un producto, etc.), es decir, de las instancias con mayor probabilidad de pertenencia a las con menor probabilidad. Con el ranking listo, se procede a seleccionar el i % superior de las instancias rankeadas, sobre el cual se realiza la evaluación a través de las mismas métricas anteriores (Sección 3.7), y por medio de la construcción de una curva que resume la precisión acumulada en función del porcentaje superior de clientes seleccionados para la evaluación.

Con el fin de obtener una mejor predicción para todas las instancias, se utiliza un procedimiento de validación muy riguroso denominado *k-fold cross validation* o *validación cruzada k-veces*, en el cual el conjunto de entrenamiento se divide en k grupos no traslapados. El modelo se entrena usando los primeros k - 1 grupos como data de entrenamiento, para luego probarlo sobre el grupo *k-ésimo*. El proceso se repite hasta que todos los grupos son utilizados una vez como conjunto de prueba. Finalmente se toma el promedio de las mediciones de precisión de los k grupos.

3.5. Valor de adpoción social de un cliente

En la sección 3.4 se planteó una metodología para seleccionar a los clientes con mayores posibilidades de adoptar un producto ofrecido bajo un enfoque que se centra de manera individual en cada cliente sin poner demasiado énfasis en lo que ocurre a su alrededor con los amigos, familiares y cercanos que lo rodean. Es por esto que se ha querido plantear una nueva manera de calcular la propensión de un cliente a aceptar una oferta por un producto que se basa en la metodología anterior pero que esta vez sí considera de manera importante el comportamiento de las personas del círculo cercano del cliente, el que se refleja a través de un valor o score de adopción social, enfoque explicado en la sección 2.5.

El objetivo principal de este enfoque es lograr entender cómo el hecho de que los amigos de un cliente adopten un producto puede hacer que dicho cliente se decida adiquirir el mismo producto a lo largo de un periodo de tiempo de 8 semanas, lo que representa la presión social que círculo cercano logra ejercer sobre el individuo en cuestión. Esta presión podría elevar la propensión del cliente a aceptar la oferta por el producto, por lo que pasaría a ser uno de los seleccionados para entrar en la campaña de éste, a diferencia de lo que pasaría si el sujeto fuera estudiado de manera individual sin considerar estos efectos sociales.

La idea consiste en calcular una componente social del score con que se rankea a los clientes y añadirlo al puntaje determinado anteriormente. Esto permitirá que el ranking se reordene a partir de estos nuevos scores y así nuevos clientes podrán ser seleccionados para entrar en la campaña de marketing directo de un producto.

Para concretar este nuevo modelo propuesto, se debe en primera instancia contabilizar el número de amigos que cada cliente posee al interior de la red social de teléfonos móviles, los que se caracterizan por estar a un arco de distancia y tener una relación con un peso mayor a 1. Luego se debe calcular la cantidad de amigos que adoptaron el mismo producto en campaña durante un periodo compuesto por los meses previos al de estudio, con lo que se logrará calcular una proporción de amigos que adoptan. Una vez se conoce este valor, se procede a normalizar todos los scores sociales y se suman a los scores calculados previamente de manera individual, lo que obviamente alterará el orden original permitiendo construir un nuevo ranking de clientes.

3.6. Evaluación

Una vez finalizado el proceso de clasificación, es necesario medir la calidad de los resultados obtenidos. Según [4], para esto se debe utilizar el concepto de *matriz de confusión*, la cual tiene por objetivo comparar las clasificaciones reales con las predichas por el modelo. Si una instancia positiva es clasificada como positiva se trata de un *positivo verdadero* o *true positive*(TP); si es clasificada como negativa se trata de un *negativo falso* o *false negative*(FN).

Si una instancia negativa es clasificada como negativa se cuenta como *falso verdadero* o *true negative*(TN); si es clasificada como positiva se cuenta como *positivo falso* o *false positive*(FP) [29]. A partir de estas definiciones y del ejemplo de la Tabla 3.1, se pueden definir una serie de medidas que serán presentadas a continuación:

		Clases reales	
		0	1
Clases predichas	0	TN	FN
	1	FP	TP
	Totales	N = TN + FP	P = FN + TP

Tabla 3.1: Matriz de confusión genérica

$$Tasa\ positivos\ verdaderos = \frac{TP}{P}$$

$$Tasa\ positivos\ falsos = \frac{FP}{P}$$

A partir de medidas anteriores, se pueden construir otras más complejas que reflejan de manera más realista los resultados obtenidos. Entre ellas se pueden destacar [4]:

- **Recall:** Medición que refleja la proporción de casos positivos reales que el modelo consigue predecir correctamente como positivos.

$$Recall = \frac{TP}{P}$$

- **Precision:** Medición que refleja la proporción de casos que el modelo predice como positivos que realmente corresponden a casos positivos.

$$Precision = \frac{TP}{TP+FP}$$

- **Accuracy:** Medición referida al nivel de certeza total del modelo dentro del universo en el cual está inserto el problema.

$$Accuracy = \frac{TP+TN}{P+N}$$

- **F-Measure:** Esta medida es una media geométrica entre dos cocientes relativos a la medidas *precision* y *recall*. La principal ventaja de este indicador radica en que la combinación de estas dos medidas permite comparar resultados de manera más fácil que haciéndolo con cada una por separado.

$$F\text{-}Measure = \frac{2}{\frac{1}{recall}+\frac{1}{precision}}$$

- **Curva ROC:** Esta curva muestra la habilidad del clasificador para posicionar las instancias verdaderas respecto a las falsas, es decir, la curva ROC mide la relación de la tasa de positivos verdaderos (predicciones acertadas) versus la tasa de positivos falsos (predicciones erradas). Si bien esta curva no tiene una fórmula asociada, sí tiene una métrica: el *área bajo la curva ROC* o *area under the ROC curve* (AUC). Ésta área bajo la curva es equivalente a la probabilidad que el clasificador posicione una instancia aleatoria positiva más alto que una instancia aleatoria negativa.

- **Lift:** Es una medida que compara la precisión del modelo con respecto a la proporción total de instancias positivas en la base de testeo, la cual se representa por medio de la siguiente ecuación:

$$Lift = \frac{precision}{\frac{P}{P+N}}$$

En este estudio se utilizarán las métricas *recall*, *precision* y *F-measure* para medir el rendimiento de los distintos modelos en los diferentes experimentos que se llevarán a cabo. Los dos primeras serán las métricas principales de la evaluación ya que se enfocan únicamente en los casos positivos, por ende demuestran rápidamente y de manera simple la capacidad de los modelos de identificar a los clientes que efectivamente adoptaron un producto ofrecido en una de las campañas. Y la medida F será una medida secundaria pero de todas maneras importante, dado que al combinar las dos métricas anteriores en una sola permite capturar información sobre los errores cometidos por el modelo, reflejando la capacidad predictiva del modelo en su totalidad a través de un sólo número.

3.7. Despliegue

Finalmente, una vez obtenidos los resultados, todas las especificaciones respecto a la metodología utilizada, los algoritmos diseñados, los modelos construidos y los resultados obtenidos será entregada a los analistas en una serie de documentos de traspaso de información que permitirán que estos puedan replicar el trabajo llevado a cabo de manera de ejecutar los modelos en sistemas propios. En particular, los resultados serán entregados en tablas que corresponden al ranking que detalla el identificador del cliente en cuestión y el score respectivo asociado a la probabilidad de adopción del producto bajo estudio.

Capítulo 4

Aplicación en datos reales

En este capítulo será presentada una aplicación sobre datos reales de una compañía de telecomunicaciones chilena. La estructura de la sección se basará en la metodología descrita en el capítulo 3. En primer lugar se realizará una breve explicación del diseño de los diferentes experimentos que se llevarán a cabo. Más adelante, se expondrán las características de la data utilizada y se detallarán todos los pasos del preprocesamiento de los datos. Luego se describirá la manera cómo se construye la red social de teléfonos móviles y qué tipo de información adicional se puede extraer a partir de ella. Una vez que la data ya está lista para ser utilizada, se explicará la manera en que fueron construidos los modelos de clasificación utilizados para la selección de los clientes target, para luego detallar la forma en que se puede incluir en los modelos el valor de adopción social de los clientes. Finalmente, se presentan los resultados obtenidos a partir de todos los modelos construidos en las secciones previas.

4.1. Diseño de los Experimentos

Para lograr un diseño integral de los experimentos a realizar, el primer paso que se debe llevar a cabo es evaluar qué datos se encuentran disponibles y definir cuáles serán utilizados durante la investigación. En este caso, se cuenta con dos fuentes de información relativa a los clientes, una que contiene datos correspondientes al comportamiento comercial del suscriptor y otra que posee datos que dan cuenta de la relación social que mantiene el usuario con sus contactos más cercanos.

Posteriormente se deben determinar las técnicas de preprocesamiento de datos que será necesario incorporar para preparar las bases de datos, las que pueden considerar la limpieza, transformación, integración, reducción o balanceo de éstas. Lo anterior permitirá trabajar con bases más ordenadas y mejor estructuradas, lo que es una ventaja al momento de ejecutar los modelos ya que la información que se logre extraer a partir de ellos podría ser más fidedigna.

Una vez que se cuenta con conjuntos de datos bajo los estándares de calidad requeridos, se continúa con la combinación de las bases que contienen ambos tipos de datos, esto con el objetivo de trabajar con una sola gran base y así permitir que la información sea extraída a partir de patrones conjuntos.

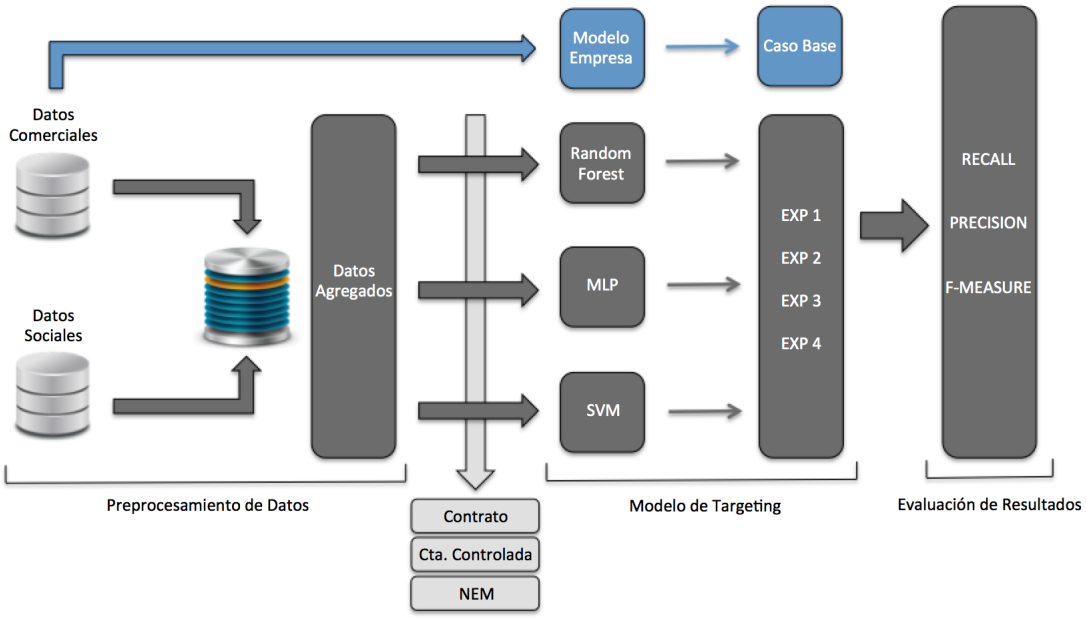


Figura 4.1: Diagrama que resume el diseño de los experimentos

Luego de que las bases de datos que se ocuparán como input de los algoritmos de clasificación están listas para ser utilizadas, los modelos de targeting se ejecutan sobre tres productos que la compañía de telecomunicaciones ofrece: up-selling de contrato, up-selling de cuenta controlada y cross-selling de navegación en el móvil, basándose en tres técnicas de clasificación con diferentes enfoques: Random Forest, Multilayer Perceptron y Support Vector Machines. La ejecución de estos modelos estará enmarcada en una serie de experimentos diseñados para comparar los resultados de las diferentes configuraciones de los modelos y encontrar aquel que tenga el mayor poder de predicción de la adopción de productos.

Expermientos	Mes entrenamiento	Mes a predecir
Experimento 1	Julio	Agosto
Experimento 2	Agosto	Septiembre
Experimento 3	Julio	Septiembre
Experimento 4	Julio + Agosto	Septiembre

Tabla 4.1: Experimentos que se realizan para evaluar poder de predicción de los modelos

Lo experimentos que se llevarán a cabo estan diseñados de manera que entre ellos vaya variando la manera de estructurar el conjunto de entrenamiento que será utilizado como input para los algoritmos de clasificación que se usan como base de los modelos de targeting. A continuación se detallará cada uno de los cuatro experimentos que se llevarán a cabo (Tabla 4.1).

Experimentos 1 y 2

Estos primeros experimentos contemplan la configuración más simple de los modelos de targeting que se usarán en este trabajo de investigación, la cual busca predecir la adopción de productos en el mes t a partir de un conjunto de entrenamiento del mes $t-1$. Así, el objetivo de estos experimentos será predecir lo ocurrido en los meses de Agosto (Exp. 1) y Septiembre (Exp. 2) utilizando como conjunto de entrenamiento bases de datos del mes inmediatamente anterior, es decir, Julio y Agosto respectivamente.

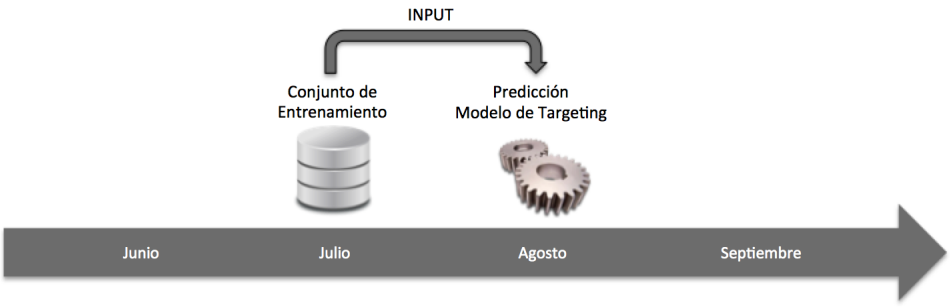


Figura 4.2: Experimento 1. Predicción mes de Agosto con ventana de tiempo de 1 mes.

Experimento 3

El tercer experimento consiste en replicar los anteriores pero generando una ventana de tiempo de dos meses entre el mes correspondiente a los datos del conjunto de entrenamiento y el mes sobre el cual se quiere predecir. En otras palabras, se busca predecir la adopción en el mes t en base a un conjunto de entranamiento generado en $t-2$. De esta manera, en la realidad se estaría prediciendo el mes de Septiembre utilizando los datos de entrenamiento del mes de Julio.

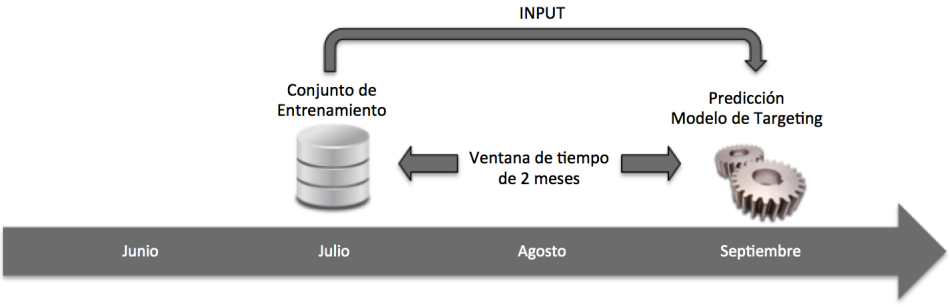


Figura 4.3: Experimento 3. Predicción mes de Septiembre con ventana de tiempo de 2 meses.

Experimento 4

Mediante este último experimento se desea evaluar el efecto que tiene sobre el poder de predicción el hecho de agregar en el conjunto de entrenamiento datos de dos meses consecutivos, es decir, predecir la adopción en el mes t teniendo como base de entrenamiento los datos agregados de los meses $t-2$ y $t-1$. Por lo tanto, siguiendo lo recién descrito, se apunta a predecir la adopción del mes de Septiembre con un conjunto de entrenamiento construido a partir de datos de Julio y Agosto.

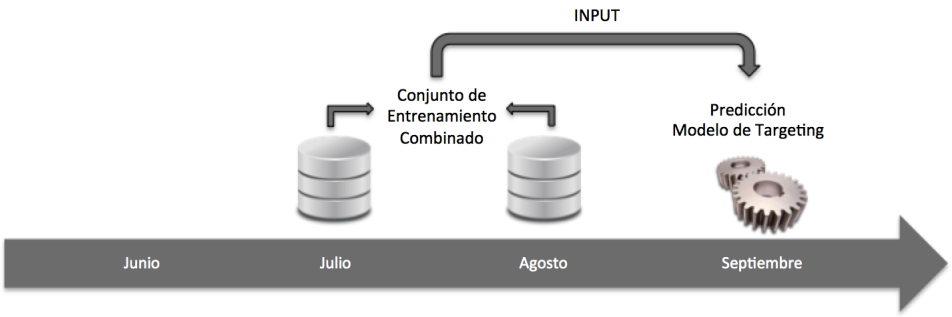


Figura 4.4: Experimento 4. Predicción mes de Septiembre con entrenamiento combinado.

Llevados a cabo los cuatro experimentos descritos anteriormente, resulta interesante buscar una manera de no sólo basar los modelos de targeting en el valor individual del cliente, si no que también en el valor de adopción social que estos poseen dada la estructura y dinamismo de la red social de teléfonos móviles con la que interactúa a diario. Para esto se debe crear una medida del valor de adopción social del suscriptor, a partir del cual se modificará el ranking de clientes y por ende se podrían indentificar nuevos posibles adoptadores de producto que antes no era posible captar.

Finalmente, para evaluar los resultados obtenidos se utilizarán una serie de medidas que buscan reflejar la capacidad de predicción que tiene los diferentes modelos usados en cada uno de los experimentos que fueron diseñados. El principal objetivo es identificar qué configuración es la que permite encontrar la mayor cantidad de clientes propensos a adoptar un producto ofrecido por la compañía de telecomunicaciones, de manera de captar a dichos usuarios utilizando un volumen menor de recursos económicos.

4.2. Datos de teléfonos móviles

Los datos fueron obtenidos a través de una compañía líder en el negocio de la telefonía móvil en Chile con una base de clientes cercana a los 9 millones de suscriptores. Los datos recopilados abarcan los meses de Julio, Agosto y Septiembre del año 2013, y corresponden sólo a aquellos clientes que tienen algún tipo de plan de telefonía móvil contratado con la compañía, alcanzando un número cercano a los 2,7 millones de clientes mensuales. Se cuenta

con una base de datos bruta[1] de cada mes, construidas a partir de 177 atributos tanto socio-demográficos como de comportamiento comercial de cada cliente.

Además, se cuenta con la información de tres campañas de marketing realizadas por la compañía durante los tres meses indicados, donde cada una buscaba ofrecer a un grupo de clientes un producto diferente: up-selling de contrato[2], es decir una mejora en el contrato actual que posee; up-selling de contrato con cuenta controlada[3]; y cross-selling que apuntaba a la venta de una bolsa de internet para navegar en el móvil (NEM).

Estas campañas son diseñadas en base al uso del canal *call center* para llegar al cliente objetivo, teniendo dos posibles estados luego de un intento de llamada telefónica: cliente *contactado* o cliente *no contactado*. El cliente contactado es aquel con quién se pudo establecer comunicación y se logró realizar el ofrecimiento del producto que forma parte de la campaña, mientras que el no contactado es aquel al que no se logró hacer la oferta, ya sea porque no contestó la llamada o porque sí lo hizo pero la llamada falló, resultó no ser el titular de la suscripción, o solicitó que lo contactaran nuevamente más tarde. El cliente contactado puede aceptar o rechazar la oferta. Si la acepta y contrata el producto se deja en periodo de latencia durante tres meses antes de que pueda formar parte de otra campaña. Si la rechaza, una vez más se da un espacio de tiempo de tres meses antes de que sea enviado a la misma campaña o a una de otro producto. Por último, si el cliente no logra ser contactado, éste se reinserta al proceso de comunicación siguiente correspondiente a la misma campaña.

En esta ocasión se tuvo acceso a una base de datos bruta que especificaba los clientes que fueron seleccionados para cada campaña de marketing, los clientes que respondieron ante una llamada telefónica ofreciendo el producto (contactados), y cuál fue la respuesta de quiénes tomaron la llamada (positivos)[4] (Tablas 4.2, 4.3, 4.4).

Producto en campaña	En campaña	Contactados	Positivos	% Positivos
Contrato	118.659	36.310	4.296	3,6 %
Cuenta controlada	103.977	38.802	3.698	3,6 %
NEM	54.817	14.909	1.150	2,1 %

Tabla 4.2: Estadísticas campañas mes de Julio

[1]Base de datos extraída directamente de los sistemas de la compañía, por lo que necesitan ser procesadas antes de ser utilizadas

[2]Plan de telefonía móvil por el que se paga un cargo fijo, pero que si se sobrepasa dicho monto, no interrumpe el servicio, si no que realiza cobros extra.

[3]Plan de telefonía móvil por el que se paga un cargo fijo que no permite sobrepasar dicho monto interrumpiendo el servicio.

[4]El porcentaje de positivos es calculado como el número de respuestas positivas por sobre el total de clientes que entraron en campaña.

Producto en campaña	En campaña	Contactados	Positivos	% Positivos
Contrato	174.738	16.692	4.059	2,3 %
Cuenta controlada	104.001	11.368	7.353	7,1 %
NEM	48.824	3.500	1.284	2,6 %

Tabla 4.3: Estadísticas campañas mes de Agosto

Producto en campaña	En campaña	Contactados	Positivos	% Positivos
Contrato	181.068	28.402	5.954	3,3 %
Cuenta controlada	112.703	29.260	7.104	6,3 %
NEM	52.014	9.049	1.501	2,9 %

Tabla 4.4: Estadísticas campañas mes de Septiembre

Adicionalmente, se poseen dos bases de datos que almacenan la información social de los clientes. Una de ellas corresponde al detalle de las relaciones que se construyen al interior de la red social de teléfonos móviles, representadas a través de un *grafo social* que posee alrededor de 5 millones de clientes y más de 125 millones de enlaces, y que especifica el nivel de importancia de dicha relación. A partir de ella se extraen 17 atributos sociales que caracterizan a cada cliente, los que son almacenados en la segunda base de datos, y que dicen relación con el número de contactos, tanto fuertes como débiles, que tiene un cliente y con las comunidades a las que pertenece.

4.3. Preprocesamiento de los datos

Como se mencionó en la sección 3.3, para contar con los datos adecuados que se ajusten a los modelos a utilizar, es necesario realizar un preprocesamiento de los datos que incorpore un análisis de la calidad de la data, en relación a la existencia de datos faltantes y/o ruidosos, a la integración y reducción de los datos, y a la posterior transformación de los mismos. Todo el detalle respecto de las variables sociodemográficas y de comportamiento comercial se podrá encontrar más adelante.

4.3.1. Limpieza de los datos

En las bases de datos que contienen la información sociodemográfica y de comportamiento comercial de los clientes se encontraron una serie de atributos que no cumplían con las condiciones mínimas de calidad para ser incluídos en los modelos por lo que fueron eliminadas. En particular se realizaron las siguientes acciones, las que redujeron el número de variables de 177 a 92.

- Se eliminan 35 atributos que no cuentan con valores en sus filas, es decir, que tienen un 100 % de datos faltantes.
- Se elimina un grupo de 12 variables que presenta un proporción de datos faltantes superior al 55 %.
- 8 variables se eliminan ya que no corresponden a data que fuera de utilidad para el modelo. En particular, 5 de ellas hacen alusión a datos personales del cliente como su código de cliente o su RUN[5], y las restantes 3 son iguales para todos los clientes, por lo que no aportan mayor información.
- Se suprimen 5 variables por información redundante ya presente en otros atributos. Específicamente se eliminan variables correspondientes a fechas de ciertos hitos, ya que son mejor representadas por otros atributos que especifican antigüedad a partir de dicha fecha.
- Se elimina un conjunto de 6 variables que dicen relación con el lugar de residencia del cliente, ya que por un lado, 4 de ellas son igualmente representadas por un atributo que especifica la región donde vive el suscriptor, y por otro lado, 2 de ellas relativas al lugar de trabajo no se consideran ya que no se tiene certeza de que sean correctas.
- 19 atributos de valores binarios son borrados de la base de datos ya que más del 98 % de las observaciones poseen valor 0, por lo que se descarta que puedan tratarse de atributos relevantes para el estudio.

Atributo	% Datos faltantes
genero	28,70 %
pac	28,40 %
valor	19,60 %
dias_mora_sum_6	17,70 %
dias_mora_min_6	17,70 %
dias_mora_max_6	17,70 %
gse	17,50 %
equipo_articulo	1,80 %
equipo_antiguedad	1,70 %
factura	1,30 %
nota_de_credito	1,30 %

Tabla 4.5: Atributos con una menor proporción de datos faltantes

Por otro lado, se identificaron algunas variables que presentan una proporción menor de datos faltantes (Tabla 4.5), los que pueden ser completados a través de las diferentes estrategias nombradas en la sección 3.3. En particular, en esta ocasión se decidió reemplazar por la media, la moda o manteniendo las proporciones de los datos, tal como se menciona en [16], además buscando introducir las menores variaciones posibles a los datos. Los datos faltantes de las variables binarias *genero*, *pac* y *valor* fueron completados aleatoriamente siempre manteniendo la proporción de cada clase, mientras que los faltantes de las variables categóricas *gse* y *equipo_articulo* fueron reemplazados por la moda correspondiente a cada atributo. A

[5]Rol Único Nacional, Número de identificación de los ciudadanos chilenos.

diferencia de las anteriores, los datos faltantes de las demás variables fueron completados por la media, ya que se trataba de atributos numéricos.

Con respecto a los datos ruidosos, se descubrieron sólo casos aislados en que las observaciones presentaban valores fuera de rango, por lo que se decidió eliminarlas de las bases de datos. Las variables que más presentaron valores ruidosos o fuera de rango fueron *edad* y *factura*.

4.3.2. Transformación de los datos

Como parte del proceso de preparación de los datos, las variables categóricas presentes en la base de datos debieron ser transformadas a variables binarias. Para algunas como *cod_categoria*, *genero*, *marca_plan*, *cod_situacion*, y *valor* que presentaban 2 categorías sólo fue necesario reemplazar sus valores por ceros y unos. En el caso de los atributos *region*, *gse* y *viral_mkt* que contaban con 15, 5 y 4 categorías respectivamente, se construyen nuevas variables binarias que despliegan la información presente en ellos. Para *region* se optó por reducir las 15 regiones a 4 zonas correspondientes a las zonas geográficas *region_norte*, *region_centro*, *region_sur* y *region_metropolitana*. Para *gse* se diseñaron 5 variables binarias para las categorías correspondientes a los 5 grupos socioeconómicos, *gse_ABC1*, *gse_C2*, *gse_C3*, *gse_D* y *gse_E*, mientras que para *viral_mkt* se construyeron 4 variables binarias para las 4 categorías, *viral_0*, *viral_1*, *viral_2* y *viral_3*[6]. Luego de finalizado el preprocesamiento de los datos, las bases de los meses de Julio, Agosto y Septiembre terminan con 102 atributos entre sociodemográficos y de comportamiento comercial. La Tabla 4.6 presenta el detalle de las variables.

[6]Estas variables binarias representan la influencia que posee un cliente en su red social, donde *viral_3* corresponde a aquel más influyente y *viral_0* al menos influyente

ID Atributo	Descripción de Atributos
1 - 3	Número de teléfono, código de abonado y número de cliente
4	Región del país donde el cliente declara vivir
5	Segmento comercial al que pertenece el cliente
6 - 7	Características del plan contratado
8	Proporción de clientes que tiene plan multimedia
9	Edad del cliente
10 - 12	Antigüedad de los servicios contratados por el cliente
13	Tipo de equipo móvil que posee el cliente
14 - 18	Importe pagado por el cliente por diferentes servicios
19 - 21	Tráfico de navegación: cantidad de datos, minutos conectado y número de conexiones a internet
22 - 26	SMS enviados y recibidos de diferentes operadores
27 - 31	Minutos de salida hacia teléfonos de otros operadores
32 - 36	Llamadas de salida hacia teléfonos de otros operadores
37 - 42	Contactos de salida correspondientes a teléfonos de otros operadores
43 - 48	Contactos de entrada correspondientes a teléfonos de otros operadores
49 - 53	Contactos totales correspondientes a teléfonos de otros operadores
54 - 58	Minutos de entrada hacia teléfonos de otros operadores
59 - 63	Llamadas de entrada hacia teléfonos de otros operadores
64 - 65	Ingresos obtenidos por conexiones con diferentes operadores
66 - 69	Características de mora del cliente
70 - 76	SMS de entrada y salida onnet y offnet
77	Días de navegación en el móvil
78 - 81	Nivel de importancia del clientes en la red
82 - 86	Grupo socioeconómico al que pertenece el cliente
87 - 88	Cambios de plan y equipo que realiza el cliente
89 - 95	Pagos realizados por el cliente
96 - 98	Perfil de valor del cliente
99 - 100	Perfil de navegación en el móvil del cliente
101	Ganancia promedio obtenida por ese cliente (ARPU)
102	Proporción de clientes que tiene el servicio 3G

Tabla 4.6: Atributos sociodemográficos y de comportamiento comercial del cliente.

4.3.3. Integración y Reducción de los datos

Con el fin de construir -para cada uno de los tres productos bajo investigación- la base de datos definitiva que será utilizada en el entrenamiento de los modelos de clasificación, es necesario integrar tres bases de datos: por un lado la base que contiene las variables sociodemográficas y de comportamiento comercial de los clientes, por otro la base que detalla la respuesta de éstos a la oferta realizada por la compañía, y por último la base que incluye los atributos sociales de cada uno, cuya extracción y obtención será explicada en la sección 4.3.

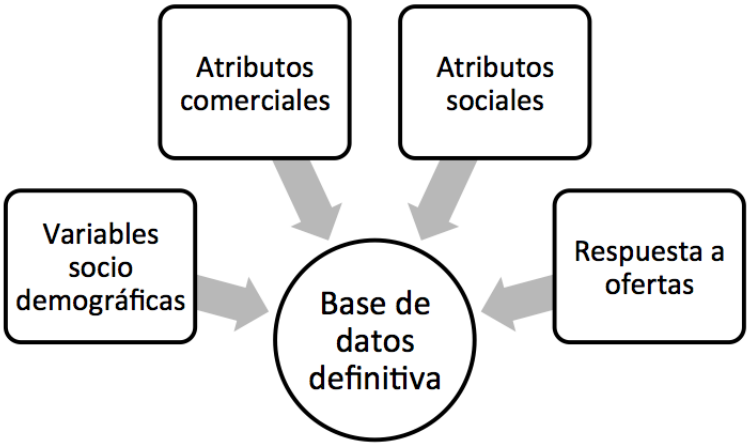


Figura 4.5: Integración de las bases de datos

El primer paso consiste en agregar las bases de datos que poseen atributos de los clientes, es decir, reunir las variables sociodemográficas, comerciales y sociales para formar una base de 119 variables. Posterior a esto, se une la base recién construida con la variable de respuesta de cada cliente a la oferta, variable binaria que en el modelo de clasificación servirá como el *target* de entrenamiento sobre el cual el modelo debe aprender para encontrar el patrón que permitirá identificar a los clientes propensos a contestar positivamente. Para lograr esto, se utiliza como *llave* el número de teléfono del cliente, de manera de unir las bases de datos a través de la coincidencia de los números de teléfono que identifican a cada cliente en cada una de las bases.

Adicionalmente, con el objetivo de evitar redundancias existentes en la base de datos integrada, se propone la realización de una *selección de atributos* basada en el análisis de correlación entre las variables. En una primera instancia esto se realiza sólo para los datos del mes de Julio, ya que se desea evaluar si se logran o no mejores resultados, para así poder replicarlo para el resto de los meses. A partir de lo anterior, se construyen dos nuevas configuraciones de la data: una donde se identifican las variables que poseen un coeficiente de correlación superior a 0.9, y otra en la cual se detectan los atributos que presentan un coeficiente mayor a 0.7. En ambos casos, de una pareja de variables correlacionadas, se mantiene aquella que representa un total, y se elimina la que representa un desglose de la anterior (por ejemplo *minutos_totales* se mantendría y *minutos_compañia1* se eliminaría).

En la Tabla 4.7 se detalla el número de variables que posterior a la selección se utilizan para entrenar el modelo de clasificación.

Producto en campaña	Corr > 0.9	Corr > 0.7
Contrato	106	68
Cuenta controlada	110	68
NEM	103	69

Tabla 4.7: Selección de atributos para correlación superior a 0.9 y 0.7. Mes de Julio

4.3.4. Balanceo de los datos

Dado que en este problema el número de clientes que aceptan la oferta por un producto es muy reducido en comparación con la cantidad de clientes que no lo hace, se produce un claro desbalanceo de las clases (Tablas 4.8, 4.9, 4.10). Como se comentó en la sección 3.2.5, para solucionar este problema se pueden considerar las técnicas de over-sampling y/o under-sampling. En este caso, para poder realizar una comparación de los resultados, ambas técnicas serán utilizadas, donde para el sobresampleo se usará la técnica SMOTE, y para el subsampleo se reducirá aleatoriamente el número de instancias de la clase negativa.

Producto en campaña	Clase Positiva	Clase Negativa
Contrato	3,6 %	96,4 %
Cuenta controlada	3,6 %	96,4 %
NEM	2,1 %	97,9 %

Tabla 4.8: Problema de clases desbalancedadas mes de Julio

Producto en campaña	Clase Positiva	Clase Negativa
Contrato	1,5 %	98,5 %
Cuenta controlada	7,1 %	92,9 %
NEM	2,6 %	97,4 %

Tabla 4.9: Problema de clases desbalancedadas mes de Agosto

Producto en campaña	Clase Positiva	Clase Negativa
Contrato	3,3 %	96,7 %
Cuenta controlada	6,3 %	93,7 %
NEM	2,9 %	97,1 %

Tabla 4.10: Problema de clases desbalancedadas mes de Septiembre

Con las dos técnicas escogidas se busca equilibrar los datos de tal manera que las clases, en un primer caso, queden en proporciones de 50 % para cada una, y en un segundo caso, queden 30 % para la clase positiva y 70 % para la clase negativa. Al igual que antes, de manera preliminar esto se realizó sólo para el mes de Julio con el fin de constatar si se evidenciaban diferencias entre ambas configuraciones del conjunto de datos de entrenamiento, para que en caso de observar mejores resultados en una de las dos, se replicara el procedimiento para los otros meses bajo estudio.

Producto en campaña	SMOTE (over/under)	Clase Positiva	Clase Negativa
Contrato	100 %/200 %	6.842	6.842
Cuenta controlada	100 %/200 %	3.374	3.374
NEM	300 %/150 %	1.017	904

Tabla 4.11: Balanceo de clases SMOTE caso 50 % ambas clases

Producto en campaña	SMOTE (over/under)	Clase Positiva	Clase Negativa
Contrato	100 %/400 %	6.842	13.684
Cuenta controlada	100 %/400 %	3.374	6.748
NEM	300 %/300 %	904	2.034

Tabla 4.12: Balanceo de clases SMOTE caso 30 %-70 %

Con la técnica SMOTE fue necesario realizar un over-sampling de un 100 % para Contrato y Cuenta Controlada, y de un 300 % para NEM, además de un under-sampling que en el caso 50 %-50 % fue de un 200 % para los dos primeros productos y de un 150 % para el último, y que en el caso de 30 %-70 % fue de un 400 % para los dos primeros y de un 300 % para el restante (resumen en Tablas 4.11, 4.12). El subsampleo fue necesario para no crear un conjunto de entrenamiento demasiado grande que estuviese construido a partir de prácticamente sólo instancias sintéticas.

4.4. Extracción de atributos sociales

Tal como se planteó en la sección 3.4, una serie de *atributos sociales* pueden ser extraídos de una red social construida a partir de las interacciones generadas entre teléfonos móviles. En esta sección se detallará la manera en que se construye esta red social y la forma en que a través de esta estructura es posible rescatar nueva información que puede ser relevante para el modelo de targeting.

4.4.1. Construcción de la red

En la sección 3.2 se indicó que los datos de las transacciones telefónicas son almacenadas en bases de datos llamadas CDRs, las cuales guardan una serie de detalles de las llamadas telefónicas (tipo 1) o mensajes de texto (tipo 2), tal como muestra la Tabla 4.13.

Emisor	Receptor	Fecha	Hora	Duración(seg)	Tipo
111	112	20131012	112435	12	1
112	111	20131012	011436	190	1
113	114	20131109	172158	0	2
114	116	20131114	215746	68	1
...	...	...	...	...	...

Tabla 4.13: Detalles de transacciones al interior de un CDR

Como primer paso en la construcción de la red, se debe agregar la información de todas las transacciones que se hayan generado en el periodo de estudio, resumiendo todas las interacciones que se generan de emisor a receptor. En esta etapa aún existen dos filas que describen la relación entre dos usuarios, donde cada una representa esta relación en una de las dos direcciones posibles, por lo que existen dos *links unidireccionales* opuestos.

Una vez se tiene agregada la información, es escencial que las interacciones entre emisor y receptor queden descritas en una sola fila, por lo que los datos se reagrupan generando un único *link bidireccional* entre dos usuarios. La Tabla 4.14 muestra cómo debe quedar la base de datos previo a que sea transformada en un grafo. Cabe mencionar que los tratamientos y transformaciones a las que son sometidas estas bases de datos replican los procedimientos realizados por la compañía de telecomunicaciones que puso a disposición la información, esto con el fin de generar grafos comparables donde la diferencia de los resultados de los experimentos se den sólo por la aplicación de los modelos de targeting de clientes y no por diferencias en la configuración de éstos.

Tel1	Tel2	Llam12	Llam21	SMS12	SMS21	Dur.llam12	Dur.llam21
111	112	1	1	0	0	12	190
113	114	0	0	1	0	0	0
114	116	1	0	0	0	68	0
115	116	2	1	0	1	79	23
...	...	...	...	...	...	...	...

Tabla 4.14: Detalles de transacciones agregadas con link bidireccional

Para lograr la extracción de los atributos desde la red social es necesario construir un grafo que contenga un *peso* y una *direccionalidad* para cada link, los que serán calculados a partir de los datos almacenados en la base resumida correspondientes a un mes completo. El peso representará la importancia de la relación entre dos usuarios y se calculará en base al total de llamadas y mensajes de texto del mes, pudiendo presentar valores mayores a 1; mientras que la direccionalidad reflejará qué usuario es más responsable del peso de la relación, la que será calculada en base a la proporción que representan las llamadas y mensajes de emisor a receptor del total de interacciones que se generan entre ellos a lo largo del mes, pudiendo adoptar valores acotados entre 0 y 1. En la Tabla 4.15 puede apreciarse cómo queda construida la base de datos que representa al grafo correspodiente a la red social de teléfonos móviles.

$$Peso = \frac{\sum_{\mathrm{i}=1}^{dias\_mes} llam_{\mathrm{i}} + \sum_{\mathrm{i}=1}^{dias\_mes} sms_{\mathrm{i}}}{4}$$

$$Dir = \frac{\sum_{\mathrm{i}=1}^{dias\_mes} llam12_{\mathrm{i}} + \sum_{\mathrm{i}=1}^{dias\_mes} sms12_{\mathrm{i}}}{\sum_{\mathrm{i}=1}^{dias\_mes} llam_{\mathrm{i}} + \sum_{\mathrm{i}=1}^{dias\_mes} sms_{\mathrm{i}}}$$

Teléfono 1	Teléfono 2	Peso	Direc
111	112	0,47	0,90
111	115	0,16	0,38
113	114	1,26	0,24
113	116	1,87	0,65
115	116	0,82	0,73
...	...	...	...

Tabla 4.15: Ejemplo de un grafo social que describe las relaciones entre teléfonos móviles

Buscando replicar la configuración adoptada por la compañía de telecomunicaciones, las conexiones relevantes para el grafo utilizado en este estudio serán aquellas que posean un peso mayor o igual a 1, es decir las relaciones entre clientes que involucren 4 o más interacciones por mes, ya sea a través de llamados telefónicos, mensajes de texto o mensajes multimedia. Dado esto, se decide filtrar el grafo y descartar todos los enlaces que tengan un peso menor a 1, quedando estructurado a partir de aproximadamente 5.2 millones de nodos y 13.5 millones de arcos, reflejando una reducción importante de su estructura, lo que permitirá un manejo más sencillo del mismo. Cabe destacar que existen otras maneras más elaboradas de generar umbrales para filtrar grafos, tal como se describe en [82].

En relación a lo anterior, la figura 4.6 muestra el comportamiento del grafo en cuanto al número de clientes que integran el grafo social y al número de interacciones mensuales que se da entre ellos considerando el umbral de 4 interacciones establecido. El gráfico muestra que a medida que se incrementa el número de interacciones que posee un arco, el número de clientes que están conectados por dicho enlace va disminuyendo velozmente, comportamiento que representa la distribución *power-law*, hecho muy importante ya que es por esta razón que pueden calcularse las diferentes métricas de la red social asociadas a comunidades.

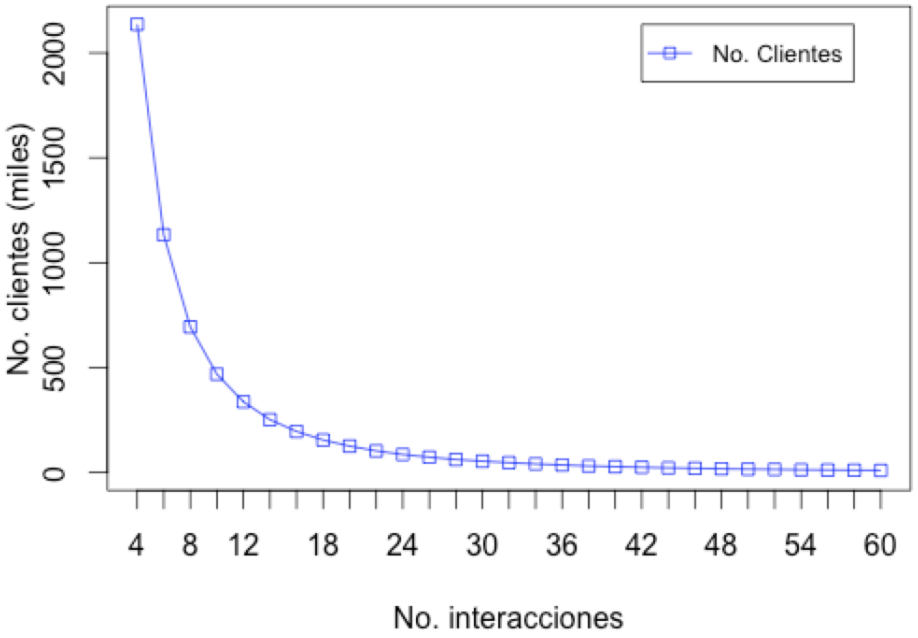


Figura 4.6: Distribución powerlaw del grafo social de llamadas telefónicas

El grafo que finalmente logra construirse (Tabla 4.15) puede representarse en una red social de la manera en que la Figura 4.7 lo presenta.

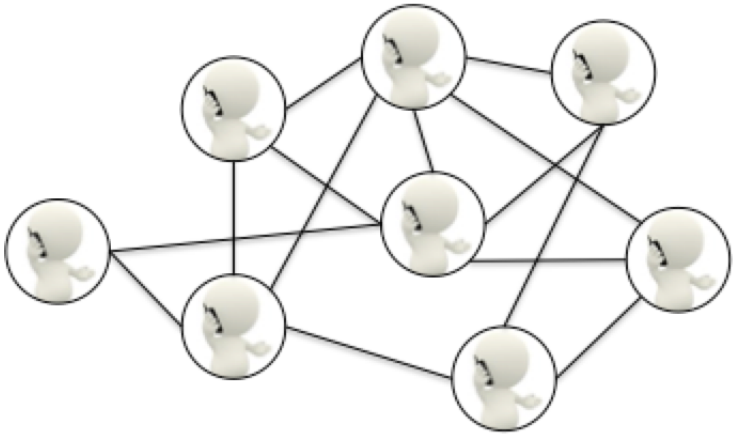


Figura 4.7: Grafo social telefónico

4.4.2. Extracción atributos

Los atributos sociales de cada cliente se extraen a partir de la red social de teléfonos móviles descrita en la sección 4.3.1 de la cual son parte. Las variables que dicen relación con el número de contactos de un cliente se obtienen simplemente a través de la contabilización de los enlaces que cada uno posee, separando entre contactos fuertes y débiles dependiendo del peso de dicho enlace, donde los enlaces fuertes serán aquellos con un peso mayor a uno y los débiles los que posean un peso menor. En este estudio, los atributos referidos a las comunidades a las que el cliente puede pertenecer se extraen a través del descubrimiento de cliques, los que corresponden al máximo subgrafo compuesto de al menos tres nodos en el cual todos sus nodos están conectados entre sí [61]. Otras maneras de extraer información a partir de comunidades pueden ser técnicas como Particionamiento de Enlaces, Expansión y Optimizción Local, Detección Fuzzy y Algoritmos Dinámicos, acerca de las cuales se pueden encontrar mayores detalles en [82].

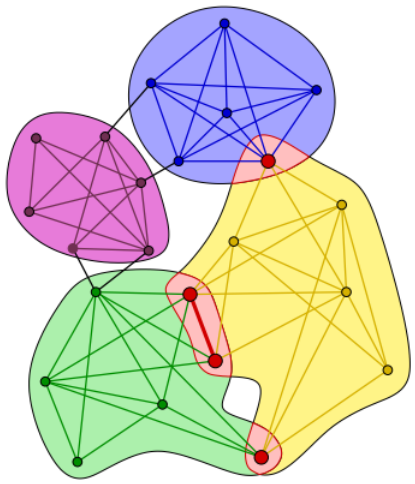


Figura 4.8: Comunidades al interior de una red social

El método más conocido para encontrar cliques es el propuesto por [70], denominado *percolación de cliques*, el cual se basa en la probabilidad de que los arcos internos de una comunidad formen cliques. Por un lado, es muy probable que se formen cliques si es que las comunidades son densas en arcos internamente, y por otro, es probable que no se formen cliques entre comunidades si es que no existen muchos arcos entre ellas. Se utiliza el término *k-clique* para referirse a un subgrafo de k elementos que forma un clique, para luego determinar que dos k-clique son adyacentes si comparten *k-1* nodos. Así, se determina que una cadena de k-cliques corresponde a la unión de dichos elementos adyacentes, para luego definir una comunidad de k-cliques como el subgrafo con el mayor número de éstos conectados.

Luego, para encontrar estos k-cliques al interior de la red es necesario diseñar una heurística, la que permitirá ir construyendo las comunidades a partir de las definiciones anteriores.

Atributo	Descripción de Atributos
contact_ count	Número de contactos totales
strong_ contact_ count	Número de contactos fuertes
strong_ contacts_ weights	Suma de los pesos de contactos fuertes
weak_ contact_ count	Número de contactos débiles
weak_ contact_ weights	Suma de los pesos de contactos débiles
competitor_ strong_ contacts	Número de contactos fuertes con teléfonos de la competencia
competitor_ weak_ contacts	Número de contactos débiles con teléfonos de la competencia
community_ count	Número de comunidades a las que pertenece
community_ count_ as_ strong	Número de comunidades a las que pertenece como miembro regular
community_ count_ as_ weak	Número de comunidades a las que pertenece como miembro asociado
community_ size_ mean	Tamaño promedio de las comunidades a las que pertenece
community_ size_ regulars_ mean	Número medio de miembros regulares de las cumunidades a las que pertenece
community_ size_ orphan_ mean	Número medio de miembros asociados de las cumunidades a las que pertenece
reach_ two_ step	Alcance en dos saltos
reach_ two_ step_ competitors	Alcance en dos saltos a nodos de la competencia
clust_ coef	Coeficiente de cluster

Tabla 4.16: Atributos sociales extraídos de la red social de teléfonos móviles

4.5. Targeting social de clientes

Con el diseño de los experimentos listo y las bases de datos completas, se procede a aplicar las diferentes técnicas de clasificación descritas previamente en la sección 2.2. Las técnicas Random Forest, Multilayer Perceptron y Support Vector Machines son ejecutadas sobre las bases de entrenamiento construidas para los tres productos que son parte de la investigación (Contrato, Cuenta Controlada y Navegación en el Móvil).

Las tres técnicas toman como input el conjunto de datos que agrupa los atributos socio-demográficos, de comportamiento comercial y sociales, más la variable dependiente binaria o target que se desea predecir -en este caso la pertenencia de los clientes a la clase de quienes adoptan un producto o a la clase de quienes no lo hacen- el cuál posee en su interior una serie de patrones que permiten realizar la clasificación de los clientes. Para los tres productos, las técnicas de clasificación utilizadas generan una lista con dos scores para cada cliente que está en la base de testeo o de prueba. Estos scores representan la probabilidad que cada cliente tiene de pertenecer a una de las dos clases, por lo que cada puntaje pertenece al intervalo [0,1] y la suma de ellos da como resultado exactamente 1, donde la clase negativa corresponde a aquellos clientes que aceptaron la oferta y terminaron adoptando el producto ofrecido, mientras que la clase positiva corresponde a los clientes de quienes no se obtuvo una respuesta afirmativa respecto de la oferta.

Luego de finalizado el proceso de obtención de los puntajes, se debe construir un ranking de clientes en función del score que cada uno tiene para la clase positiva. En este caso particular, el ranking se estructura como una lista de clientes ordenada de mayor a menor probabilidad de pertenecer a la clase positiva (Tabla 4.17).

No. Cliente	Score Clase Pos.
102	0.861
101	0.529
103	0.380
...	...

Tabla 4.17: Ejemplo ranking clientes

Una vez terminada la construcción del ranking, se realizan cortes en la parte superior de éste (scores más altos) considerando diferentes proporciones de clientes de manera de analizar cómo va evolucionando la capacidad de predicción de los modelos de clasificación a medida que la cantidad de clientes que cae dentro del tramo es mayor. Específicamente, aquí se generan cortes en el 10 %, 20 %, 30 % y 40 % superior del ranking, con el fin de ir midiendo cuántos clientes que adoptan el producto en la realidad logran ser identificados por los modelos de manera acumulada. Finalmente, para cada producto se obtienen seis rankings diferentes dependiendo de la técnica de clasificación y la manera en que se balancean las clases en la base de entrenamiento, los que permiten comparar y analizar qué combinación es la que mejor permite predecir la adopción de productos.

A modo de prueba, se lleva a cabo un proceso de clasificación convencional (enfoque de clases y no de scores) sobre el conjunto de datos de Julio para evaluar si tanto la selección de atributos como el balanceo de las clases permiten obtener mejores resultados frente al escenario en que estas estrategias no se hubieran utilizado. Para esto, la base de datos del mes de Julio se divide en un conjunto de entrenamiento y un conjunto de prueba en proporciones de 1/3 y 2/3 respectivamente, tal como se propone en [56], con el fin de realizar la primera predicción sobre datos correspondientes al mismo mes de los datos de entrenamiento, lo que se replica para las tres campañas en estudio. En base a los resultados obtenidos para el mes de Julio, se decidió cuáles de las alternativas de modelamiento para predecir los meses de Agosto y Septiembre se utilizarían, a partir de las cuales se diseñaron los cuatro experimentos descritos en la Sección 4.1.

4.6. Valor de adopción social de los clientes

Es conocido que la metodología de selección de clientes basado en atributos sociales sólo se enfoca en el valor individual o intrínseco que posee un cliente, por lo que sería interesante poder replicar este proceso de targeting alterando el ranking de clientes dado el valor de adopción social que cada uno posee gracias al comportamiento relacionado a la adopción de productos de la red de usuarios de teléfonos móviles que lo rodea.

Tal como describe la sección 3.5, el primer paso para lograr calcular el valor de adopción social de un cliente es contabilizar el número de amigos que éste tiene en el mes de Septiembre, para lo cual sólo basta con contar los usuarios con que el cliente posee un enlace y cuya relación es representada con un peso mayor a la unidad. Una vez conocido el número de amigos que posee un cliente, es vital para esta parte del estudio saber qué proporción de esos amigos han adoptado el producto previamente, lo que se logra analizando el grafo y contabilizando el número de amigos que adquieren el producto dentro de un periodo de dos meses inmediatamente anteriores al de estudio, los que para este caso corresponden a Julio y Agosto. Con estos datos se consigue establecer el *score de adopción social*, el que no es más que el número de amigos que adoptan en dicho periodo dividido el número total de amigos del cliente en cuestión.

No. Cliente	No. Amigos	No. Amigos adoptan	Score Social
102	3	1	0.33
101	1	0	0
103	8	3	0.38
...	...	...	...

Tabla 4.18: Cálculo del valor de adopción social de un cliente

Con el valor de adopción social ya calculado, éste se añade al score individual calculado para el mes de Septiembre en la sección previa con el conjunto de entrenamiento combinado entre Julio y Agosto (que es la que mejores resultados entrega), simplemente sumando el score social y el individual, y normalizando para dejar los valores entre 0 y 1. Este procedimieto genera un nuevo ranking de clientes, para el cuál se utiliza la misma metodología de evaluación que antes, basándose en la capacidad de indentificar nuevos clientes de la clase que adopta el producto y en la proporción de aciertos acumulados.

4.7. Resultados del targeting de clientes

Considerando las tres técnicas de clasificación y las dos estrategias de balanceo de datos se logran construir seis rankings para cada producto en cada uno de los meses en estudio, haciéndose necesario descubrir cuál de ellos es el que representa al modelo con el mayor poder de predicción. Para esto se utilizan las métricas de evaluación *recall*, *precision* y *f-measure* las cuales fueron descritas en la sección 3.6, y que son calculadas para los grupos de clientes correspondientes a los cuatro cortes realizados en el ranking (de 10 % a 40 % del total).

Producto en Campaña	Total Clientes	10 %	20 %	30 %	40 %
Contrato	325.751	32.575	65.150	97.725	130.300
Cuenta Controlada	197.534	19.753	39.507	59.260	79.014
NEM	887.619	88.762	177.524	266.286	355.102

Tabla 4.19: Cortes sobre ranking de clientes mes de Agosto

Producto en Campaña	Total Clientes	10 %	20 %	30 %	40 %
Contrato	487.403	48.740	97.481	146.221	194.961
Cuenta Controlada	199.290	19.929	39.858	59.787	79.716
NEM	855.525	85.553	171.105	256.658	342.210

Tabla 4.20: Cortes sobre ranking de clientes mes de Septiembre

Como se mencionó antes, el primer paso es realizar algunas pruebas de entrenamiento y testeo tradicionales sobre el mes de Julio en base a configuraciones de datos que consideran selección de atributos en función de la correlación existente entre ellos (mayor a 0,7 y a 0,9), y balanceo de clases en proporciones de 50 %-50 % y 30 %-70 %. Los resultados expuestos en las Tablas 4.21, 4.22 y 4.23 muestran que la estrategia consistente en seleccionar un conjunto de atributos eliminando aquellas variables que presentan alta correlación no genera mejoras en los resultados, por lo que se opta por mantener todas las variables para el modelamiento correspondiente a los otros meses. También es posible advertir que los mejores resultados se obtienen cuando se trabaja con un balanceo que equilibre ambas clases en un 50 %. En base a estos resultados, se decide continuar con los experimentos sin utilizar selección de atributos y realizando un balanceo de manera de equiparar el peso de ambas clases (tanto en over-sampling como en under-sampling).

Configuración de los datos	Recall DT	Recall NN	Recall SVM
Under(50/50)	63.6 %	57.8 %	34.1 %
Under(50/50)_FS_0.9	63.3 %	57.3 %	32.4 %
Under(50/50)_FS_0.7	62.1 %	50.6 %	29.5 %
SMOTE(50/50)	18.9 %	39.9 %	26.2 %
SMOTE(50/50)_FS_0.9	17.8 %	37.1 %	24.7 %
SMOTE(50/50)_FS_0.7	16.4 %	36.6 %	23.8 %
SMOTE(30/70)_FS_0.9	0.69 %	24.5 %	0 %
SMOTE(30/70)_FS_0.7	0 %	14.1 %	0 %

Tabla 4.21: Resultados de clasificación tradicional realizada sobre el mes de Julio para Contrato

Configuración de los datos	Recall DT	Recall NN	Recall SVM
Under(50/50)	55.9 %	58.1 %	34.9 %
Under(50/50)_FS_0.9	54.7 %	47.4 %	31.7 %
Under(50/50)_FS_0.7	53.9 %	46.8 %	31.3 %
SMOTE(50/50)	10.6 %	36.7 %	26.1 %
SMOTE(50/50)_FS_0.9	9.1 %	37.1 %	26.0 %
SMOTE(50/50)_FS_0.7	8.8 %	36.2 %	26.9 %
SMOTE(30/70)_FS_0.9	0.35 %	25.5 %	0 %
SMOTE(30/70)_FS_0.7	0.18 %	24.1 %	0 %

Tabla 4.22: Resultados de clasificación tradicional realizada sobre el mes de Julio para Cuenta Controlada

Configuración de los datos	Recall DT	Recall NN	Recall SVM
Under(50/50)	61.4 %	58.6 %	39.1 %
Under(50/50)_FS_0.9	58.9 %	51.5 %	33.6 %
Under(50/50)_FS_0.7	54.3 %	48.8 %	32.4 %
SMOTE(50/50)	0 %	28.6 %	15.8 %
SMOTE(50/50)_FS_0.9	0 %	28.2 %	0 %
SMOTE(50/50)_FS_0.7	0 %	28.2 %	0 %
SMOTE(30/70)_FS_0.9	0 %	0 %	0 %
SMOTE(30/70)_FS_0.7	0 %	0 %	0 %

Tabla 4.23: Resultados de clasificación tradicional realizada sobre el mes de Julio para Navegación en el Móvil

Habiendo establecido la configuración de los datos que se utilizará, se continúa con los **experimentos 1 y 2**, cuyos primeros resultados se muestran en las Tablas 4.24, 4.25, 4.26. A través de gráficos de barras se compara la proporción de clientes que adopta que es identificada únicamente por el modelo base[7] (porción celeste), la que es encontrada únicamente por los modelos sociales (porción azul oscuro) y la que es detectada por ambos modelos (porción azul claro), proceso que se repite para las tres técnicas de clasificación (detalle en Anexos).

[7]El modelo base corresponde al modelo de comparación que sólo incorpora las variables sociodemográficas y de comportamiento comercial, sin incluir los atributos sociales.

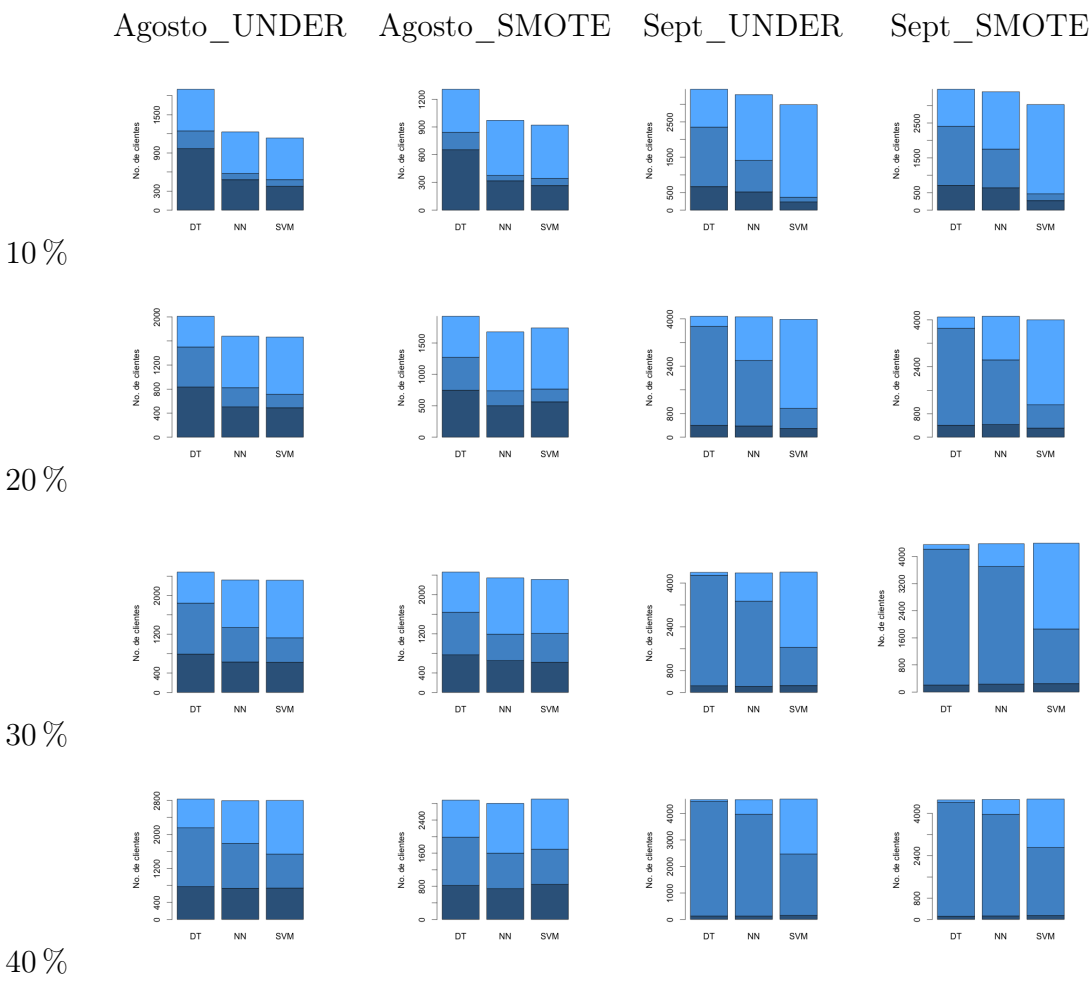


Tabla 4.24: Resultados experimentos 1 y 2 para el producto Contrato

Los resultados indican que para el producto Contrato (Tabla 4.24) el modelo propuesto a partir de la técnica Random Forest(DT) presenta en prácticamente todos los cortes del ranking un nivel de precisión mayor que el modelo base de comparación. Esto se refleja con mayor claridad en la predicción realizada para el mes de Agosto ya que el número total de clientes de la clase positiva que son identificados es mayor y que la cantidad de clientes "nuevos" que son encontrados por el método propuesto y no por el base también es mayor (independiente del método de balanceo de clases utilizado). Sin embargo, con Multilayer Perceptron(NN) y Support Vector Machines(SVM) el rendimiento de los clasificadores muestra un nivel similar o inferior al método sin atributos sociales, lo que se deduce de la menor cantidad de clientes identificados y del menor número de clientes nuevos que logran detectar.

Por otro lado, para el producto Cuenta Controlada (Tabla 4.25) es posible notar que tanto para el mes de Agosto como para el de Septiembre, la capacidad de predicción del modelo social es muy similar para las tres técnicas de clasificación utilizadas. Los resultados muestran que independiente del método de balanceo de clases, en el mes de Agosto el modelo social logra identificar un número mayor de clientes nuevos en comparación con el

modelo base, sobre todo cuando se analizan los cortes de 10 % y 20 %. Sin embargo, para el mes de Septiembre, si bien a medida que se va incrementando la proporción del corte del ranking el modelo propuesto detecta una mayor cantidad de clientes, el número de clientes nuevos encontrados es cada vez menor. Entre las técnicas utilizadas, las que muestran mejor desempeño con balanceo under-sampling son Multilayer Perceptron para el mes de Agosto y Random Forest para Septiembre, mientras que para over-sampling son SVM y Random Forest respectivamente.

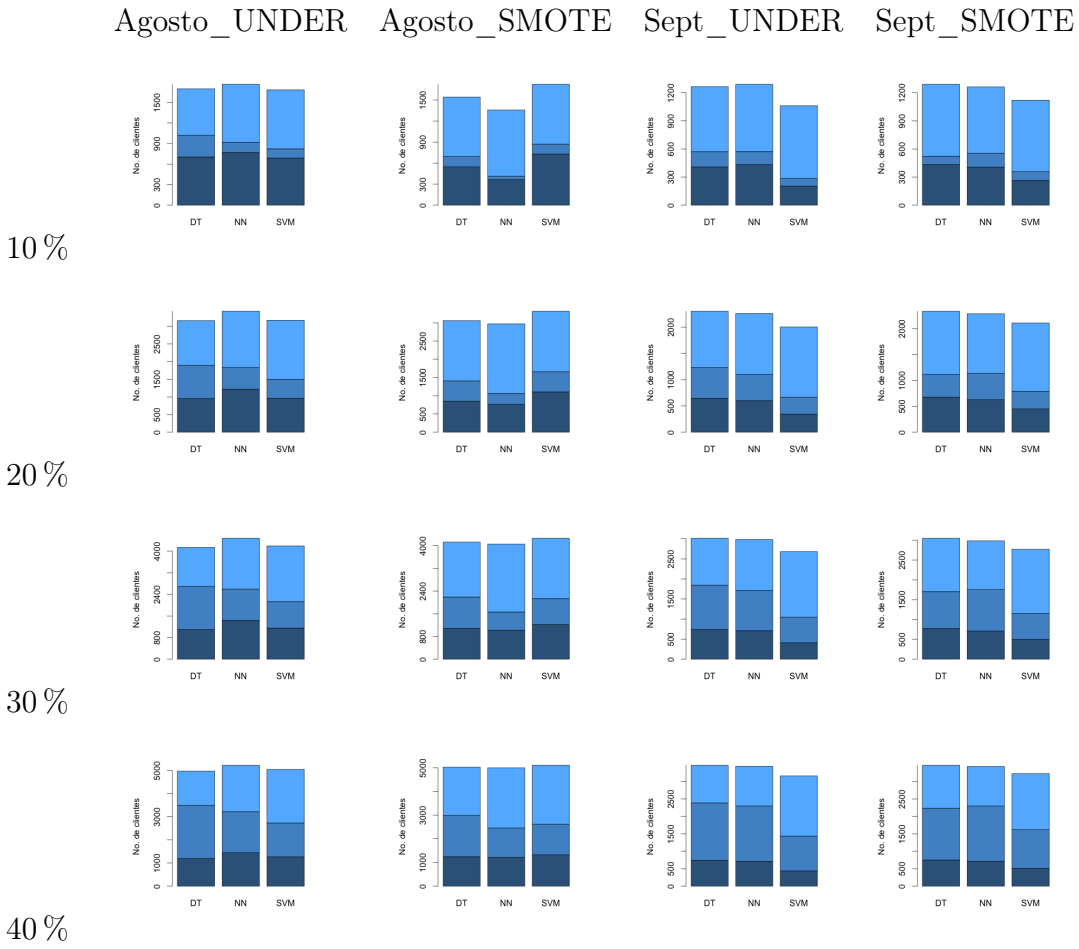


Tabla 4.25: Resultados experimentos 1 y 2 para el producto Cuenta Controlada

Por último, en el caso del producto Navegación en el Móvil (Tabla 4.26), la gráfica muestra que los resultados obtenidos para el mes de Agosto en general están por debajo de los resultados mostrados por el método base, tanto para el caso balanceado con under-samplig como para el realizado con over-sampling, a excepción del corte de 10 % donde el modelo propuesto logra detectar un número importante de clientes nuevos, fracción que va disminuyendo a medida que se incrementa la proporción del corte. Contrario a esto, para la predicción del mes de Septiembre el modelo basado en la técnica Random Forest logró resultados muy positivos, encontrando muchos más clientes que el método base e identificando una gran cantidad de clientes nuevos no detectados previamente. Estos resultados se repiten para los dos métodos de balanceo de clases.

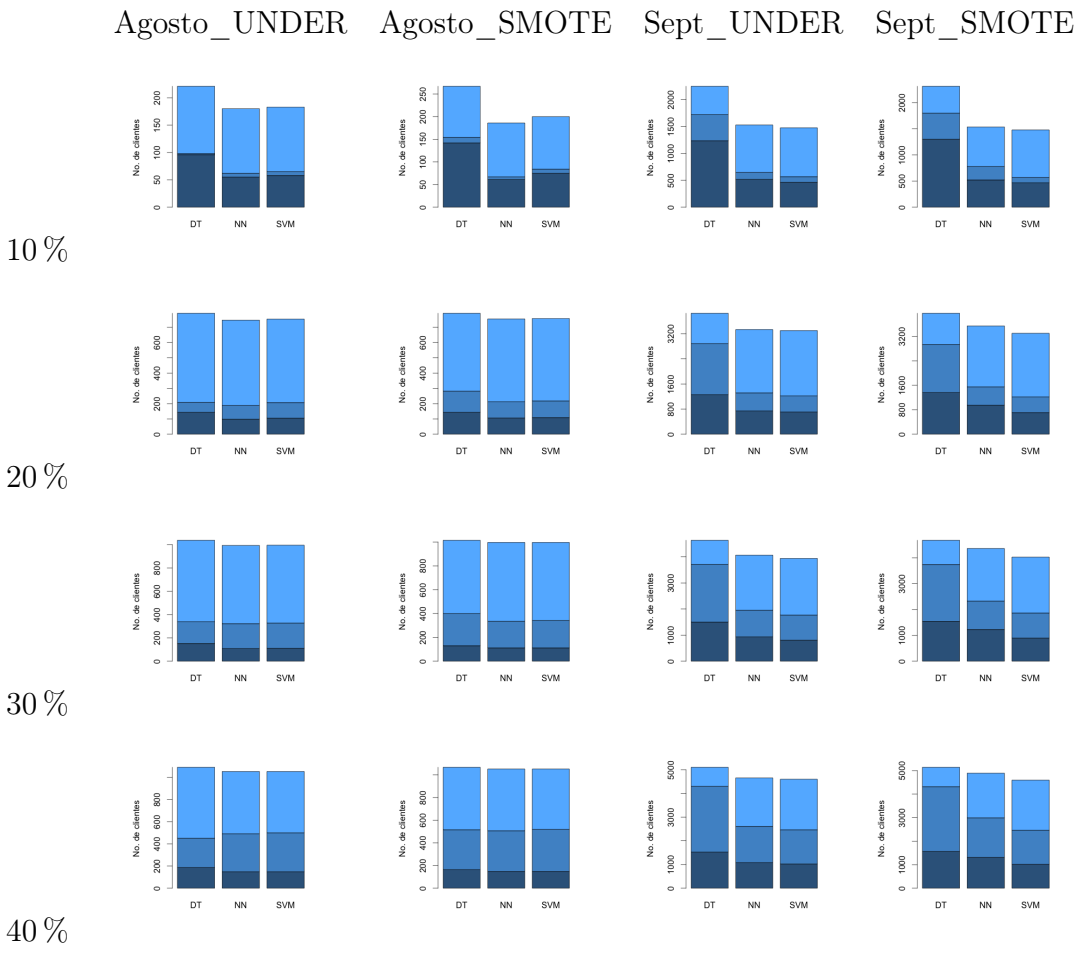


Tabla 4.26: Resultados experimentos 1 y 2 para el producto Navegación en el Móvil

Otra manera de evaluar la capacidad de predicción de los modelos propuestos es a través de la métrica *recall*, con la cual es posible ir analizando cómo se va incrementando el número de aciertos a medida que el corte toma una mayor proporción del total de clientes. Esta forma de evaluación es conocida como la *curva de aciertos acumulados*.

En el Experimento 1 (Tabla 4.28) para el producto Contrato mes de Agosto (Figura 4.9) es posible observar que el modelo propuesto en base a la técnica Random Forest y balanceo under-sampling entrega resultados superiores al modelo base (Tabla 4.27)y mejores que las demás técnicas de clasificación. Este modelo logra identificar a un 53,1 % de los clientes de clase positiva dentro del 40 % superior del ranking de clientes, contra un 50,8 % que alcanza a encontrar el modelo base, lo que se traduce en la detección de 94 clientes más por parte del modelo social. En el caso de balanceo over-sampling, el modelo basado en Random Forest entrega resultados similares al modelo base de comparación.

Modelo Empresa - Agosto												
	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	16.11 %	1.97 %	3.51 %	28.97 %	1.77 %	3.34 %	41.69 %	1.70 %	3.26 %	50.78 %	1.55 %	3.01 %
CC	13.53 %	4.89 %	7.19 %	30.05 %	5.43 %	9.20 %	41.35 %	4.98 %	8.90 %	51.41 %	4.65 %	8.52 %
NEM	9.73 %	0.13 %	0.25 %	50.51 %	0.33 %	0.65 %	68.87 %	0.30 %	0.59 %	70.35 %	0.23 %	0.45 %

Modelo Empresa - Septiembre												
	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	36.96 %	4.51 %	8.05 %	53.74 %	3.28 %	6.19 %	62.91 %	2.56 %	4.92 %	69.45 %	2.12 %	4.12 %
CC	13.95 %	4.97 %	7.33 %	26.67 %	4.75 %	8.07 %	36.86 %	4.38 %	7.83 %	46.52 %	4.14 %	7.61 %
NEM	12.92 %	0.20 %	0.40 %	61.12 %	0.48 %	0.95 %	80.29 %	0.42 %	0.83 %	81.49 %	0.32 %	0.63 %

Tabla 4.27: Resultados obtenidos por el modelo que se usará como base de comparación

Es posible ver que los resultados siguen un comportamiento contínuo a través del tiempo ya que en el Experimento 2 (Tabla 4.27) para Contrato mes de Septiembre el modelo basado en la técnica Random Forest es el que nuevamente muestra un mejor desempeño, aunque muy similar al del modelo base. Ambos modelos son capaces de identificar a una gran parte de los clientes que adoptan el producto dentro del corte de 40 % superior del ranking, alcanzando el modelo social una tasa de aciertos de un 72,5 % y el modelo base una tasa del 69,5 % (sin importar el método de balanceo).

Modelo Propuesto Experimento 1 Contrato												
CTO	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	23.82 %	2.97 %	5.28 %	36.95 %	2.30 %	4.33 %	45.36 %	1.88 %	3.62 %	53.09 %	1.65 %	3.21 %
MLP	11.83 %	1.47 %	2.62 %	20.28 %	1.26 %	2.38 %	32.99 %	1.37 %	2.63 %	44.12 %	1.37 %	2.67 %
SVM	8.50 %	1.06 %	1.88 %	18.90 %	1.18 %	2.22 %	29.74 %	1.24 %	2.37 %	41.83 %	1.30 %	2.53 %

Modelo Propuesto Experimento 1 Cuenta Controlada												
CC	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	13.90 %	5.17 %	7.54 %	25.77 %	4.80 %	8.09 %	36.62 %	4.54 %	8.09 %	47.53 %	4.42 %	8.09 %
MLP	12.48 %	4.65 %	6.77 %	24.93 %	4.64 %	7.82 %	35.31 %	4.38 %	7.80 %	43.79 %	4.08 %	7.46 %
SVM	11.19 %	4.17 %	6.07 %	20.37 %	3.79 %	6.39 %	28.98 %	3.60 %	6.40 %	37.18 %	3.46 %	6.33 %

Modelo Propuesto Experimento 1 Navegación en el Móvil												
NEM	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	11.98 %	0.17 %	0.34 %	21.95 %	0.16 %	0.32 %	31.05 %	0.15 %	0.30 %	40.16 %	0.15 %	0.29 %
MLP	5.21 %	0.08 %	0.15 %	16.50 %	0.12 %	0.24 %	26.15 %	0.13 %	0.25 %	39.46 %	0.14 %	0.28 %
SVM	6.54 %	0.09 %	0.19 %	16.96 %	0.12 %	0.24 %	26.54 %	0.13 %	0.25 %	40.39 %	0.15 %	0.29 %

Tabla 4.28: Resultados obtenidos por el modelo propuesto en el experimento 1

Modelo Propuesto Experimento 2 Contrato												
CTO	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	32.33 %	3.95 %	7.04 %	55.40 %	3.38 %	6.38 %	67.19 %	2.74 %	5.26 %	72.53 %	2.22 %	4.30 %
MLP	27.76 %	3.39 %	6.04 %	43.07 %	2.63 %	4.96 %	61.39 %	2.50 %	4.80 %	66.73 %	2.04 %	3.96 %
SVM	9.37 %	1.14 %	2.04 %	19.24 %	1.18 %	2.22 %	31.57 %	1.29 %	2.47 %	46.01 %	1.41 %	2.73 %

Modelo Propuesto Experimento 2 Cuenta Controlada												
CC	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	15.27 %	5.44 %	8.03 %	47.66 %	8.50 %	14.42 %	61.83 %	7.35 %	13.13 %	72.50 %	6.46 %	11.87 %
MLP	14.06 %	5.01 %	7.39 %	27.56 %	4.91 %	8.34 %	41.24 %	4.90 %	8.76 %	52.96 %	4.72 %	8.67 %
SVM	9.43 %	3.36 %	4.96 %	19.65 %	3.50 %	5.95 %	29.29 %	3.48 %	6.22 %	35.68 %	3.18 %	5.84 %

Modelo Propuesto Experimento 2 Navegación en el Móvil												
NEM	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	29.96 %	0.53 %	1.03 %	66.91 %	0.59 %	1.16 %	83.02 %	0.49 %	0.97 %	85.02 %	0.37 %	0.74 %
MLP	9.72 %	0.17 %	0.34 %	20.24 %	0.18 %	0.35 %	29.03 %	0.17 %	0.34 %	37.75 %	0.17 %	0.33 %
SVM	10.32 %	0.18 %	0.36 %	20.97 %	0.18 %	0.36 %	31.23 %	0.18 %	0.36 %	41.08 %	0.18 %	0.36 %

Tabla 4.29: Resultados obtenidos por el modelo propuesto en el experimento 2

En el caso del producto Cuenta Controlada los resultados muestran que las tres técnicas de clasificación tienen un rendimiento por debajo de el logrado por el modelo base sin atributos sociales en el mes de Agosto, aunque Random Forest se sigue destacando por encima de Multilayer Perceptron y SVM. En el mes de Septiembre esto cambia ya que las técnicas Random Forest y Multilayer Perceptron obtienen mejores resultados que el modelo base, siendo la capacidad de predicción similar en ambas. Dentro del 40 % superior del ranking de clientes, el modelo base logra identificar a un 51,4 % de los clientes que adoptan en Agosto y un 46,5 % en Septiembre, mientras que el Random Forest social un 47,5 % en el caso under-sampling en Agosto, y un 53,2 % en Septiembre.

Así como para el producto Cuenta Controlada, para Navegación en el Móvil mes de Agosto los resultados obtenidos por los tres modelos de clasificación que incluyen atributos sociales se sitúan muy por debajo del desempeño del modelo base. Mientras el modelo social logra detectar a través de Random Forest con over-sampling a un 40,2 % de los clientes que adoptan el producto, el modelo base alcanza a un 70,4 % cuando se considera el 40 % superior del ranking de scores de clientes.

Similar a lo ocurrido en el mes de Agosto, para el mes de Septiembre nuevamente el desempeño del modelo social es inferior al del modelo base. Una vez más la técnica Random Forest logra superar los resultados de las demás técnicas de clasificación, tanto en el caso en que se usa el método de balanceo de clases under-sampling como el over-sampling. Con ambos métodos de balanceo, el modelo social logra detectar cerca de un 70 % de los clientes que adoptan, mientras que el modelo base consigue hacerlo para un 81,5 % al tomar en consideración al 40 % con scores más altos en el ranking.

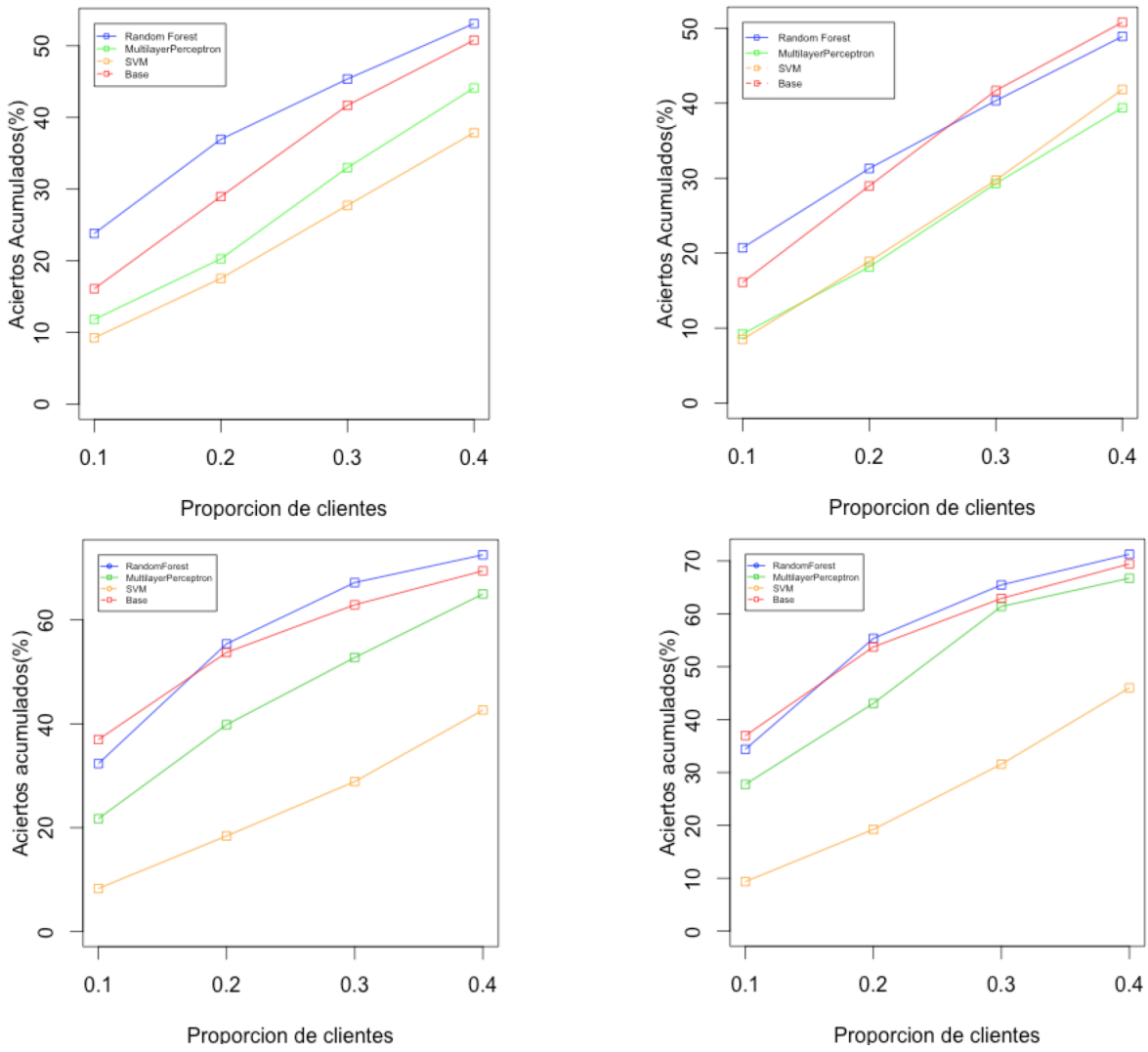


Figura 4.9: Aciertos acumulados Contrato Agosto(arriba) y Septiembre(abajo) con undersampling(izquierda) y SMOTE(derecha)

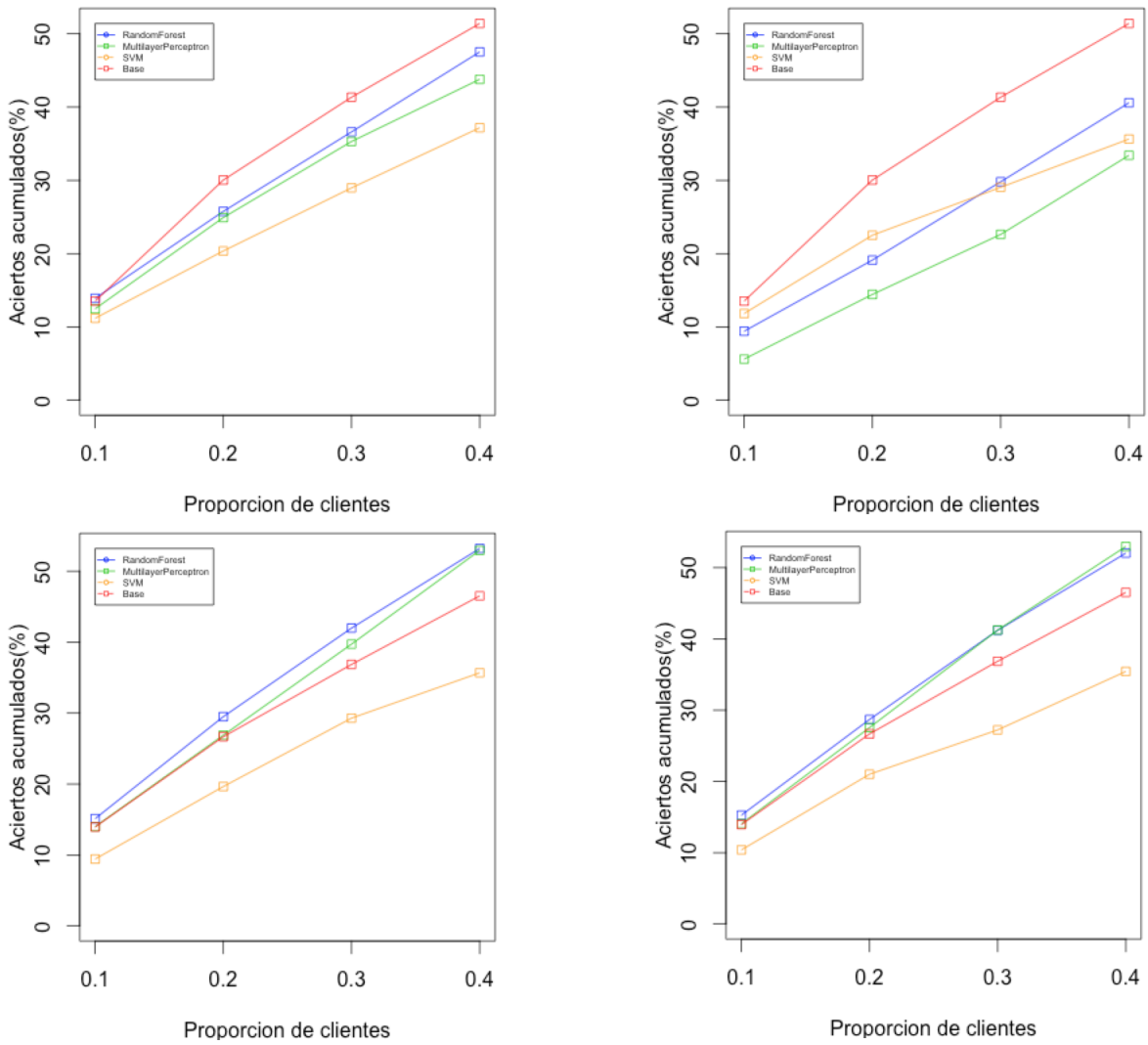


Figura 4.10: Aciertos acumulados Cuenta Controlada Agosto y Septiembre con undersampling y SMOTE

Un nuevo camino para evaluar los resultados es analizar el porcentaje de aciertos totales que se puede alcanzar al combinar los aciertos del modelo base y del modelo social. En la medida que ambos modelos logren identificar diferentes clientes de la clase que adopta el producto, el modelo combinado exhibirá un desempeño en la predicción muy por encima del desempeño de cada uno por sí sólo, logrando una sinergia importante en la clasificación.

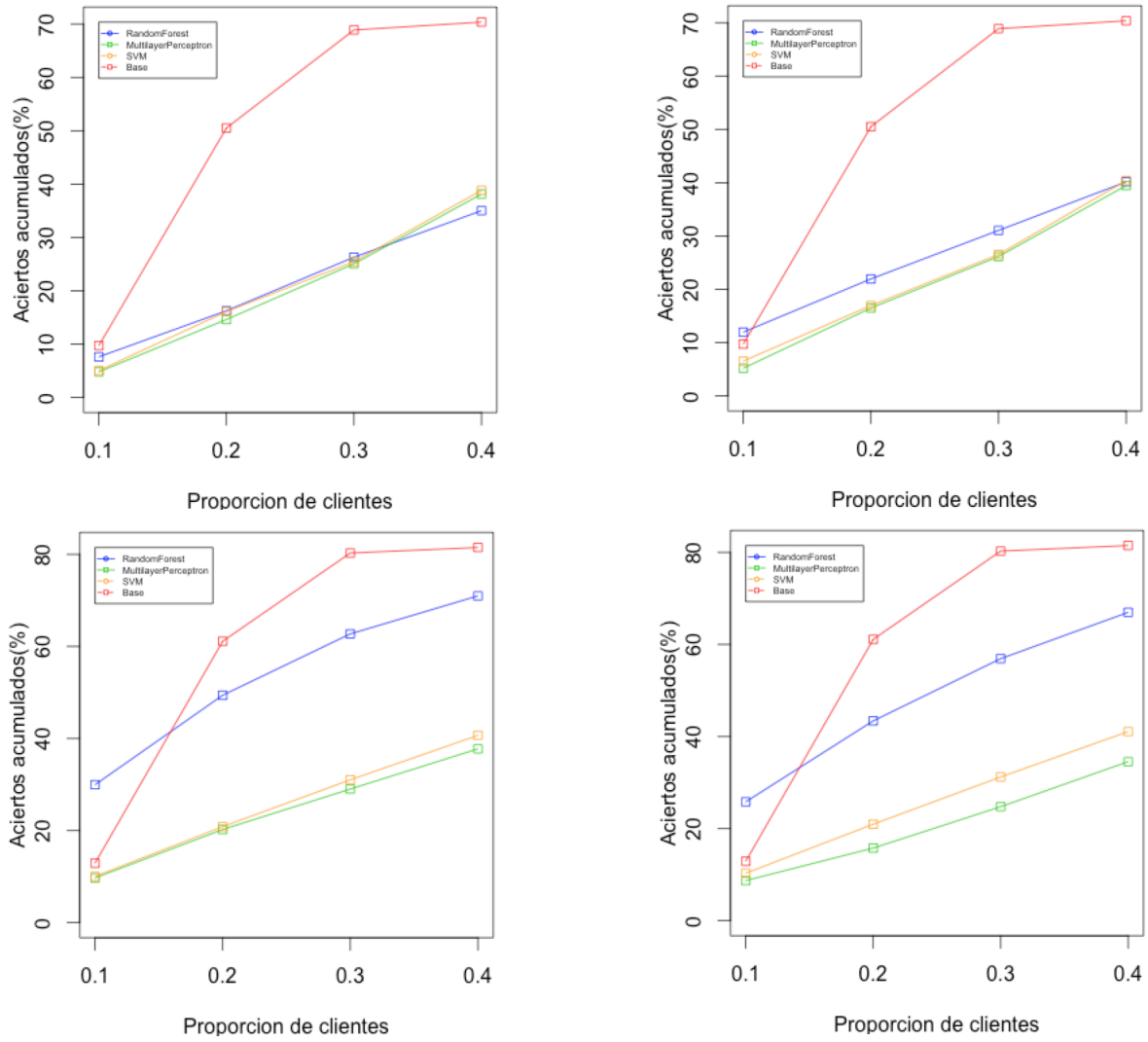


Figura 4.11: Aciertos acumulados NEM Agosto y Septiembre con under-sampling y SMOTE

Es posible observar que en los resultados exhibidos en la Figura 4.12 la curva de aciertos acumulados del producto Contrato del modelo combinado va por arriba de la curva de cada modelo individual. Si bien la mejora en la cantidad de clientes identificados dentro del 40 % superior del ranking es mucho mayor en Agosto (+18 % promedio[8]) que en Septiembre (+6 % promedio) -lo que se explica por la baja capacidad del modelo social de detectar clientes nuevos en dicho mes- al analizar el total de los clientes detectados, el mes de Septiembre está por encima con un promedio de aciertos cercano al 75 % contra un promedio de casi 70 % en Agosto (Tablas 4.30 y 4.31).

[8]Valor promedio de los resultados entregados por las tres técnicas de clasificación utilizadas a lo largo del estudio

Modelo Combinado Propuesto Experimento 1 - Contrato												
CTO	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	32.18 %	4.01 %	7.13 %	47.18 %	2.94 %	5.53 %	60.56 %	2.52 %	4.83 %	70.61 %	2.20 %	4.27 %
MLP	23.95 %	2.98 %	5.31 %	41.19 %	2.57 %	4.83 %	57.60 %	2.39 %	4.59 %	68.74 %	2.14 %	4.15 %
SVM	22.69 %	2.83 %	5.03 %	42.79 %	2.67 %	5.02 %	59.18 %	2.46 %	4.72 %	71.37 %	2.22 %	4.31 %

Modelo Combinado Propuesto Experimento 1 - Cuenta Controlada												
CC	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	20.95 %	7.80 %	11.37 %	41.60 %	7.74 %	13.06 %	56.13 %	6.97 %	12.39 %	68.34 %	6.36 %	11.64 %
MLP	18.48 %	6.88 %	10.03 %	40.40 %	7.52 %	12.68 %	55.19 %	6.85 %	12.19 %	67.94 %	6.32 %	11.57 %
SVM	23.44 %	8.73 %	12.72 %	45.08 %	8.39 %	14.15 %	57.97 %	7.19 %	12.80 %	69.42 %	6.46 %	11.82 %

Modelo Combinado Propuesto Experimento 1 - Navegación en el Móvil												
NEM	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	17.20 %	0.25 %	0.49 %	61.63 %	0.45 %	0.89 %	80.70 %	0.39 %	0.78 %	84.82 %	0.31 %	0.61 %
MLP	25.76 %	0.37 %	0.74 %	68.40 %	0.50 %	0.98 %	77.59 %	0.37 %	0.75 %	81.87 %	0.30 %	0.59 %
SVM	15.56 %	0.23 %	0.44 %	58.99 %	0.43 %	0.85 %	77.59 %	0.37 %	0.75 %	81.87 %	0.30 %	0.59 %

Tabla 4.30: Resultados obtenidos por el modelo combinado propuesto en el experimento 1

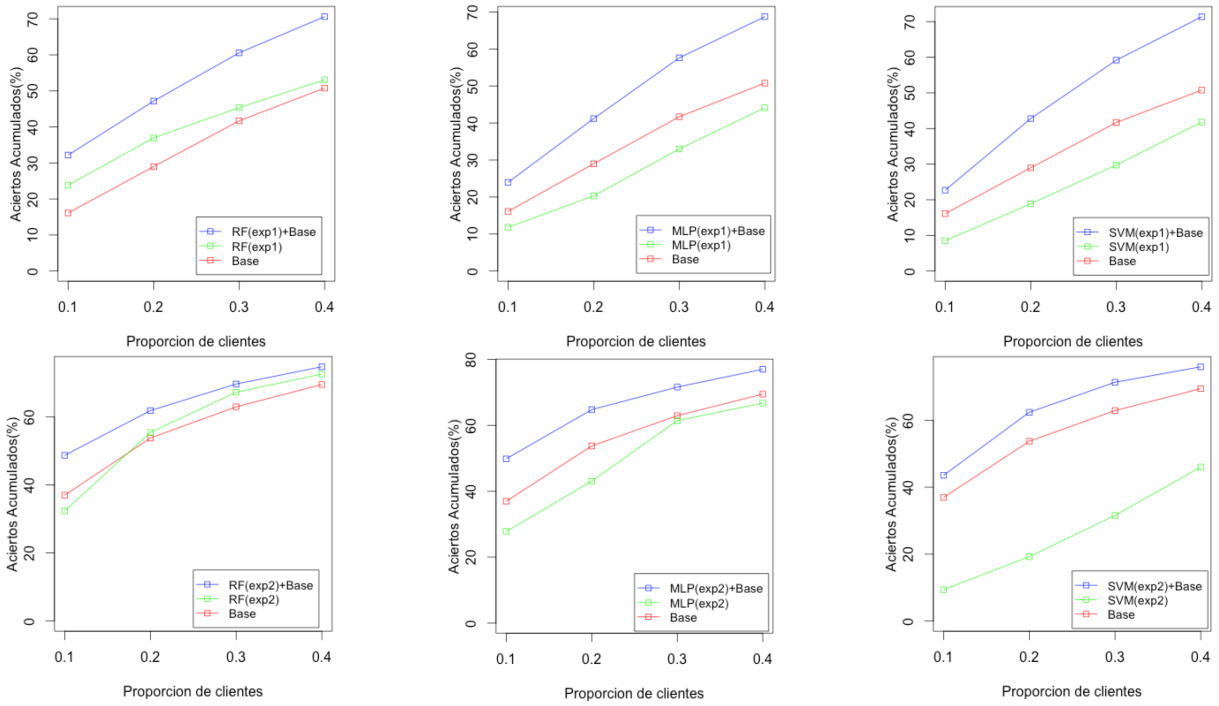


Figura 4.12: Comparación de aciertos acumulados entre el modelo social, el modelo base y la combinación de ambos para Contrato Agosto (Exp1) y Septiembre (Exp2)

Los resultados presentados en la Figura 4.13 muestran nuevamente que las curvas de aciertos acumulados del producto Cuenta Controlada se encuentra por sobre la curva de cada modelo si se analizan individualmente, marcando un claro incremento en la cantidad de clientes identificados (+25 % promedio para ambos meses). Es posible destacar que para ambos meses, el total de clientes detectados es muy alto, alcanzando un nivel en torno al 71 % promedio de las tres técnicas para los meses de Agosto y Septiembre (Tablas 4.30 y 4.31).

Para el producto Navegación en el Móvil los resultados de la Figura 4.14 muestran una vez más que la curva de aciertos acumulados se mueve por sobre la de los modelos individuales, aunque en esta ocasión con una mejora no tan sustancial para el mes de Septiembre (+3 % promedio contra +12 % en Agosto), debido principalmente a la poca capacidad del modelo social de identificar clientes que el modelo base no haya conseguido detectar. Esto quiere decir que tanto el modelo base como el modelo social coinciden ampliamente en los clientes identificados como adoptadores del producto. Sin embargo, debe destacarse que el nivel de precisión que alcanza el modelo combinado llega a una proporción de aciertos promedio de las tres técnicas para ambos meses cercano al 83 % (Tablas 4.30 y 4.31).

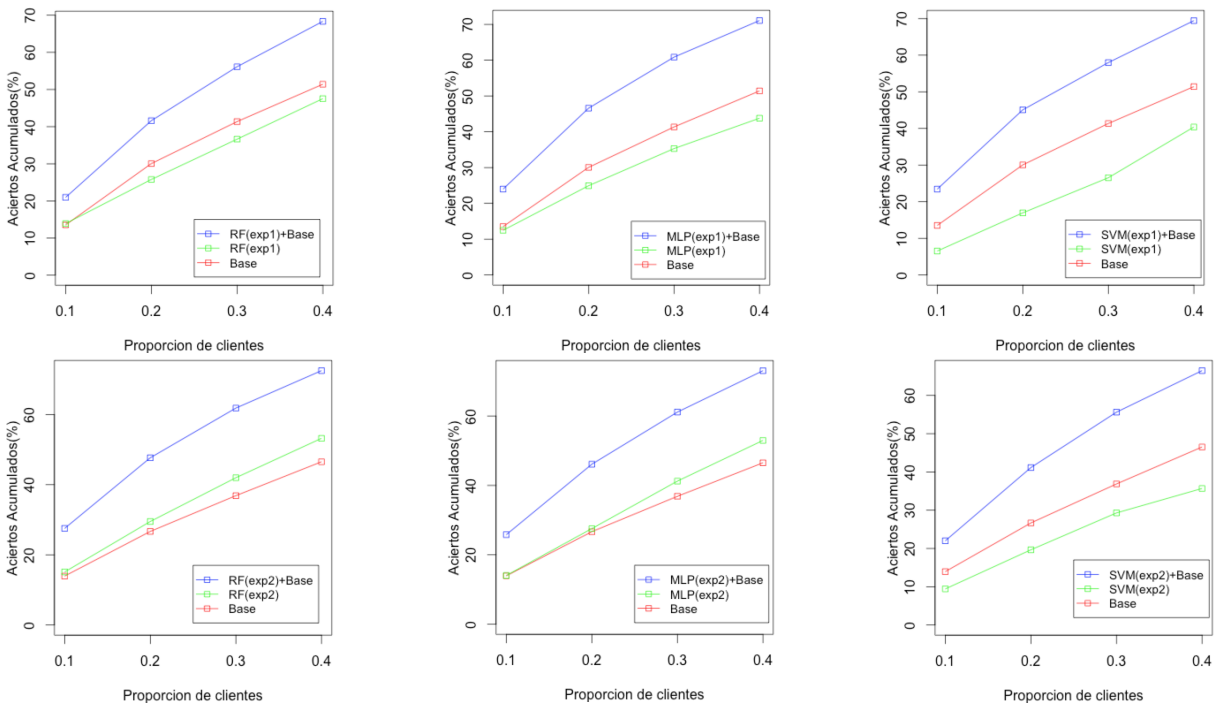


Figura 4.13: Comparación de aciertos acumulados entre el modelo social, el modelo base y la combinación de ambos para Cuenta Controlada Agosto (Exp1) y Septiembre (Exp2)

Modelo Combinado Propuesto Experimento 2 - Contrato												
CTO	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	48.66 %	5.95 %	10.60 %	61.80 %	3.78 %	7.12 %	69.62 %	2.84 %	5.45 %	74.68 %	2.28 %	4.43 %
MLP	49.84 %	6.09 %	10.85 %	64.75 %	3.96 %	7.46 %	71.59 %	2.92 %	5.60 %	76.99 %	2.35 %	4.56 %
SVM	43.56 %	5.32 %	9.49 %	62.40 %	3.81 %	7.19 %	71.37 %	2.91 %	5.59 %	75.97 %	2.32 %	4.50 %

Modelo Combinado Propuesto Experimento 2 - Cuenta Controlada												
CC	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	27.54 %	9.82 %	14.48 %	47.66 %	8.50 %	14.42 %	61.83 %	7.35 %	13.13 %	72.50 %	6.46 %	11.87 %
MLP	25.81 %	9.20 %	13.57 %	46.11 %	8.22 %	13.95 %	61.13 %	7.26 %	12.99 %	73.05 %	6.51 %	11.96 %
SVM	22.01 %	7.85 %	11.57 %	41.14 %	7.33 %	12.45 %	55.62 %	6.61 %	11.82 %	66.47 %	5.92 %	10.88 %

Modelo Combinado Propuesto Experimento 2 - Navegación en el Móvil												
NEM	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	34.49 %	0.61 %	1.19 %	66.91 %	0.59 %	1.16 %	83.02 %	0.49 %	0.97 %	85.02 %	0.37 %	0.74 %
MLP	20.97 %	0.37 %	0.72 %	65.85 %	0.58 %	1.15 %	82.82 %	0.48 %	0.96 %	83.82 %	0.37 %	0.73 %
SVM	22.17 %	0.39 %	0.77 %	66.98 %	0.59 %	1.17 %	83.02 %	0.49 %	0.97 %	84.95 %	0.37 %	0.74 %

Tabla 4.31: Resultados obtenidos por el modelo combinado propuesto en el experimento 2

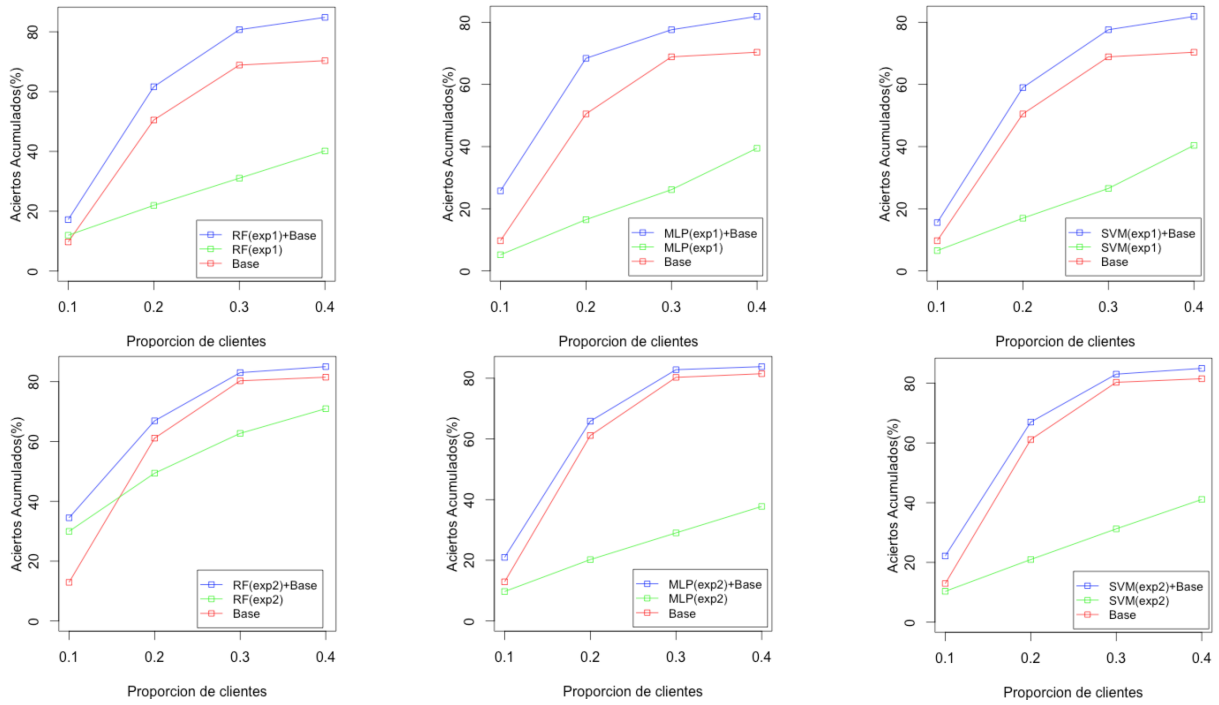


Figura 4.14: Comparación de aciertos acumulados entre el modelo social, el modelo base y la combinación de ambos para NEM Agosto (Exp1) y Septiembre (Exp2)

Una vez obtenidos los resultados de las predicciones realizadas para los meses de Agosto y Septiembre, en donde para ambos casos se toman como conjunto de entrenamiento los datos del mes inmediatamente anterior (t-1), se propone evaluar la capacidad de predicción de los modelos para una ventana de tiempo de dos meses (t-2), con el fin de determinar si el factor tiempo influye de alguna manera en el desempeño de los clasificadores. Dado esto, a través del **experimento 3** se procede a realizar una nueva predicción sobre el mes de Septiembre, pero utilizando esta vez los datos del mes de Julio como conjunto de entrenamiento (Tabla 4.32).

Modelo Combinado Propuesto Experimento 3 - Contrato												
CTO	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	50.13 %	6.12 %	10.92 %	66.16 %	4.04 %	7.62 %	74.93 %	3.05 %	5.86 %	82.33 %	2.51 %	4.88 %
MLP	46.83 %	5.50 %	9.84 %	66.78 %	3.79 %	7.18 %	73.48 %	2.99 %	5.75 %	81.86 %	2.50 %	4.85 %
SVM	41.93 %	5.12 %	9.13 %	63.80 %	3.90 %	7.35 %	77.68 %	3.16 %	6.08 %	85.17 %	2.60 %	5.05 %

Modelo Combinado Propuesto Experimento 3 - Cuenta Controlada												
CC	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	25.77 %	9.19 %	13.55 %	45.22 %	8.06 %	13.68 %	60.42 %	7.18 %	12.84 %	72.47 %	6.46 %	11.86 %
MLP	27.29 %	6.79 %	10.88 %	51.44 %	7.01 %	12.35 %	56.55 %	6.72 %	12.01 %	69.85 %	6.23 %	11.43 %
SVM	22.83 %	8.14 %	12.00 %	42.13 %	7.51 %	12.75 %	55.45 %	6.59 %	11.78 %	65.90 %	5.87 %	10.79 %

Modelo Combinado Propuesto Experimento 3 - Navegación en el Móvil												
NEM	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	22.10 %	0.39 %	0.76 %	69.37 %	0.61 %	1.21 %	87.35 %	0.51 %	1.02 %	88.68 %	0.39 %	0.78 %
MLP	20.11 %	0.34 %	0.67 %	65.78 %	0.57 %	1.13 %	82.76 %	0.48 %	0.96 %	83.62 %	0.37 %	0.73 %
SVM	22.37 %	0.39 %	0.77 %	66.18 %	0.58 %	1.15 %	82.76 %	0.48 %	0.96 %	83.82 %	0.37 %	0.73 %

Tabla 4.32: Resultados obtenidos por el modelo combinado propuesto en el experimento 3

Para evaluar si el factor tiempo influye en la precisión de los modelos de clasificación, es necesario comparar los resultados de esta nueva configuración con los entregados por el experimento 2, en el cual se predice el mismo mes de Septiembre, pero con los datos de Agosto como conjunto de entrenamiento. Esta comparación se realizará contrastando el desempeño de los modelos combinados presentados anteriormente, tomando en consideración sólo el método de balanceo de clases que mejores resultados entregue.

La Figura 4.15 muestra que para el producto Contrato los resultados obtenidos por el modelo de predicción de Septiembre construido a partir de una ventana de tiempo de dos meses (curva azul) se encuentran por encima de los entregados por el modelo original con ventana de un mes (curva verde) independientemente de la técnica de clasificación usada, y aún más cuando se compara contra el modelo base (curva roja). La mejora en la capacidad de predicción del modelo se refleja en una proporción de aciertos incrementada en un 7,5 % promedio para las tres técnicas de clasificación.

En el caso del producto Cuenta Controlada, los resultados que se presentan difieren de los del producto anterior ya que si bien el modelo diseñado en el Experimento 3 también es superior al modelo base, éste no logra superar de manera importante al del Experimento 2, reflejando mejoras cercanas al 0,4 % usando las técnicas de clasificación Random Forest y Support Vector Machines.

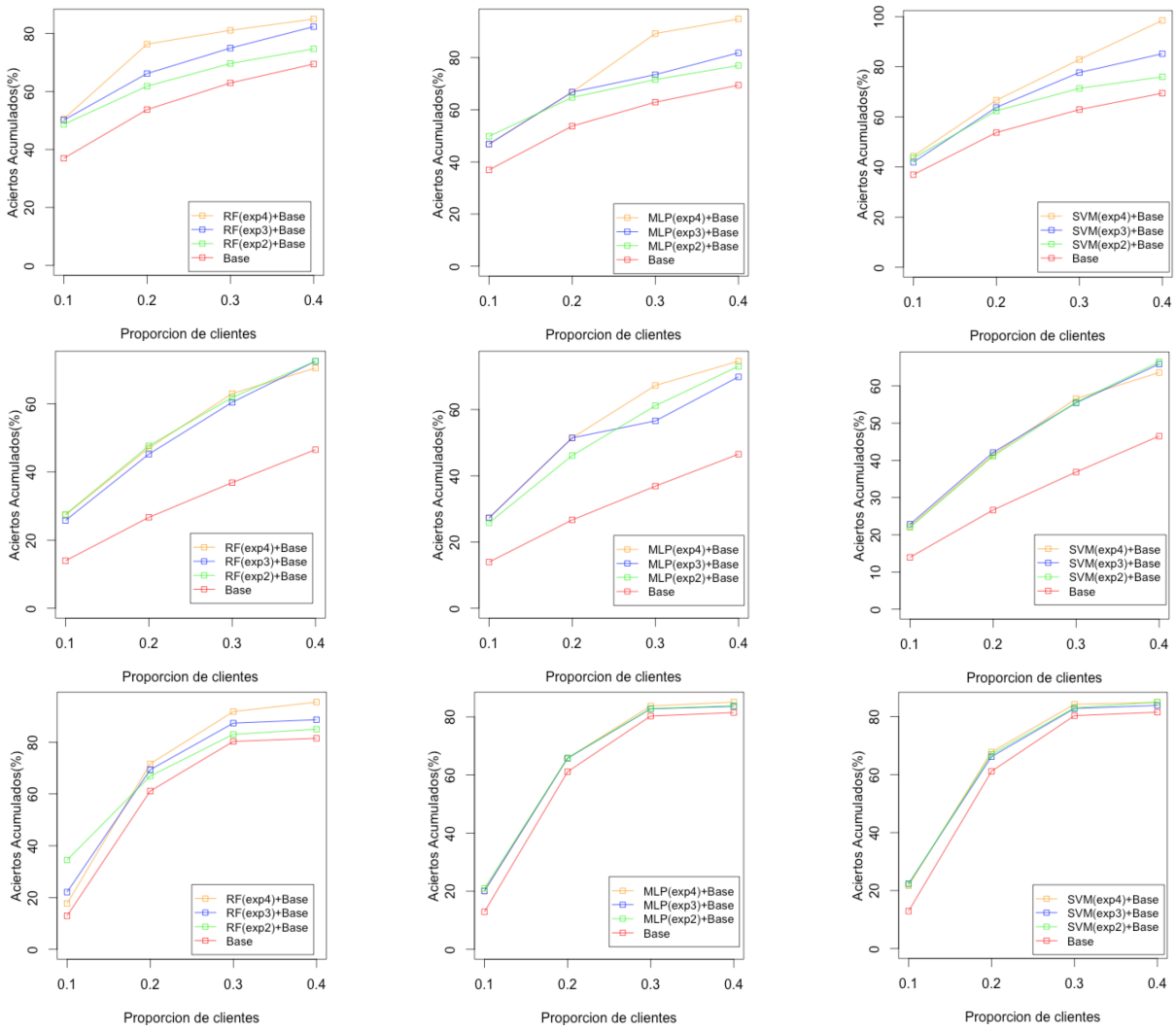


Figura 4.15: Comparación de aciertos acumulados entre el modelo base y los modelos sociales combinados de los experimentos 2, 3 y 4 para los productos Contrato(superior), Cuenta Controlada(medio) y Navegación en el Móvil(inferior)

Para el producto Navegación en el Móvil, los resultados varían dependiendo de la técnica de clasificación utilizada para construir el modelo de predicción social con ventana de tiempo de dos meses. En el caso en que se usa la técnica Random Forest, este nuevo modelo logra mejorar el nivel de aciertos cerca de un 4 %. Sin embargo, en el caso de las técnicas Multilayer Perceptron y SVM, no existen diferencias importantes entre predecir utilizando la configuración del experimento 2 o del experimento 3.

Luego de evaluar los resultados obtenidos al predecir con una ventana de tiempo de dos meses entre el conjunto de entrenamiento y el de prueba, se lleva a cabo el **experimento 4**, el que tiene por objetivo analizar el impacto que tiene sobre los resultados de la predicción de Septiembre la construcción del modelo en base a un conjunto de entrenamiento que agrupe los datos de Julio y de Agosto (Tabla 4.33).

Modelo Combinado Propuesto Experimento 4 - Contrato												
CTO	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	50.53 %	6.17 %	11.00 %	76.31 %	4.66 %	8.79 %	81.07 %	3.30 %	6.35 %	84.95 %	2.59 %	5.04 %
MLP	46.83 %	5.72 %	10.20 %	66.78 %	4.08 %	7.69 %	89.25 %	3.63 %	6.99 %	94.84 %	2.90 %	5.62 %
SVM	44.35 %	5.42 %	9.66 %	66.63 %	4.07 %	7.67 %	82.87 %	3.38 %	6.49 %	98.46 %	3.01 %	5.84 %

Modelo Combinado Propuesto Experimento 4 - Cuenta Controlada												
CC	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	27.40 %	9.77 %	14.40 %	47.01 %	8.38 %	14.22 %	62.98 %	7.48 %	13.38 %	70.49 %	6.28 %	11.54 %
MLP	27.29 %	9.73 %	4.34 %	51.44 %	9.17 %	15.57 %	67.26 %	7.99 %	14.29 %	74.60 %	6.65 %	12.21 %
SVM	22.31 %	7.95 %	11.73 %	41.49 %	7.40 %	12.55 %	56.59 %	6.73 %	12.02 %	63.62 %	5.67 %	10.41 %

Modelo Combinado Propuesto Experimento 4 - Navegación en el Móvil												
NEM	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	17.71 %	0.31 %	0.61 %	71.64 %	0.63 %	1.25 %	91.81 %	0.54 %	1.07 %	95.47 %	0.42 %	0.83 %
MLP	20.11 %	0.35 %	0.69 %	65.78 %	0.58 %	1.14 %	83.75 %	0.49 %	0.97 %	85.09 %	0.37 %	0.74 %
SVM	21.70 %	0.38 %	0.75 %	67.78 %	0.59 %	1.18 %	84.22 %	0.49 %	0.98 %	84.89 %	0.37 %	0.74 %

Tabla 4.33: Resultados obtenidos por el modelo combinado propuesto en el experimento 4

Los resultados presentados en la Figura 4.15 indican que el hecho de utilizar un conjunto de entrenamiento que agrega data histórica de dos meses inmediatamente anteriores (t-1 y t-2) al mes sobre el que se realiza la predicción (t) permite mejorar aún más los resultados obtenidos en los experimentos previos (curva amarilla). Al igual que antes, los gráficos comparan los resultados entregados por los modelos sociales combinados calculando una tasa de aciertos conjunta.

Es posible observar que para el producto Contrato el mayor alza en la capacidad de predicción se obtiene bajo la utilización de las técnicas MLP y SVM, ambas con un incremento cercano al 13 % cuando se analiza el 40 % superior del ranking de clientes (+9,6 % promedio para las tres técnicas), sin embargo el nivel de aciertos acumulados más alto lo alcanza SVM con un 98 %. En el caso de Cuenta Controlada sólo se aprecia un alza cuando el modelo se construye bajo la técnica MLP, logrando un incremento cercano a un 5 % en el último corte (+0,2 % promedio), alcanzando un 75 % de aciertos acumulados. Y finalmente para el producto Navegación en el Móvil los resultados del nuevo modelo generan la mayor mejora en la cantidad de aciertos acumulados al usar la técnica Random Forest alcanzando un aumento aproximado de un 7 % en el corte del 40 %, llegando al un nivel del 96 %, a pesar de que para las tres téncnicas la mejora promedio llegó a sólo un 3 %.

4.8. Resultados de incluir la adopción social de un cliente

Tal como se explicó en la sección 3.5, los clientes poseen un valor asociado al comportamiento de adopción de los integrantes de la red social de teléfonos móviles que los rodean, valor que puede modificar el perfil que estos tienen de cara al proceso de selección de los clientes objetivo de una oferta o campaña de marketing lanzada por una compañía. En esta sección se presentarán los resultados obtenidos al incluir el valor de adopción social en los rankings de clientes calculados en los experimentos previos, los que serán comparados con los resultados generados en el experimento 4.

Modelo Combinado Propuesto con Valor Social - Contrato												
CTO	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	49.94 %	6.10 %	10.87 %	64.21 %	3.92 %	7.39 %	71.67 %	2.92 %	5.61 %	78.84 %	2.41 %	4.67 %
MLP	46.94 %	5.73 %	10.22 %	61.09 %	3.73 %	7.03 %	69.86 %	2.85 %	5.47 %	77.36 %	2.36 %	4.59 %
SVM	44.53 %	5.44 %	9.70 %	64.13 %	3.92 %	7.38 %	73.10 %	2.98 %	5.72 %	78.19 %	2.39 %	4.63 %

Modelo Combinado Propuesto con Valor Social - Cuenta Controlada												
CC	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	24.67 %	8.80 %	12.97 %	39.66 %	7.07 %	12.00 %	44.03 %	5.23 %	9.35 %	47.26 %	4.21 %	7.74 %
MLP	24.66 %	8.79 %	12.96 %	39.92 %	7.12 %	12.08 %	45.46 %	5.40 %	9.66 %	47.23 %	4.21 %	7.73 %
SVM	21.10 %	7.52 %	11.09 %	35.27 %	6.29 %	10.67 %	42.05 %	5.00 %	8.93 %	47.07 %	4.19 %	7.70 %

Modelo Combinado Propuesto con Valor Social - Navegación en el Móvil												
NEM	10 %			20 %			30 %			40 %		
	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
RF	22.90 %	0.40 %	0.79 %	72.24 %	0.63 %	1.26 %	84.75 %	0.63 %	1.26 %	92.01 %	0.40 %	0.80 %
MLP	24.03 %	0.42 %	0.83 %	81.62 %	0.72 %	1.42 %	85.49 %	0.57 %	1.13 %	89.88 %	0.39 %	0.79 %
SVM	24.77 %	0.43 %	0.85 %	81.29 %	0.71 %	1.41 %	86.286 %	0.57 %	1.14 %	90.01 %	0.40 %	0.79 %

Tabla 4.34: Resultados obtenidos por el modelo propuesto que incorpora el valor de adopción social del cliente

Es posible observar que para los productos Contrato y Cuenta Controlada (Tabla 4.34) el hecho de incluir el valor de adopción social en el ranking de clientes (curva negra) genera una caída en la cantidad de aciertos acumulados a lo largo de los cortes realizados, lo que ocurre independiente de la técnica de clasificación utilizada. Para Contrato, el porcentaje de aciertos cae alrededor de un 15 % en promedio para las tres técnicas de clasificación, donde la caída más fuerte se da bajo la utilización de la técnica SVM (-20,27 %). En el caso del producto Cuenta Controlada la caída promedio en la capacidad de predicción alcanza un 34 %, donde el descenso más acentuado nuevamente se registra al usar la técnica SVM (-51,4 %).

Para el producto Navegación en el Móvil, los resultados muestran que en este caso la alteración realizada en el ranking de clientes por medio del valor de adopción social permite igualar e incluso mejorar los resultados obtenidos en el experimento 4. Específicamente, al utilizar la técnica Ranfom Forest los resultados de ambos modelos son prácticamente los mismos, presentando una pequeña ventaja el del experimento 4. Sin embargo, al basarse en las técnicas MLP y SVM la ventaja se da por el lado del modelo que incorpora el valor de adopción social de los clientes, logrando identificar cerca de un 5 % más de clientes.

Cabe destacar que para los tres productos y las tres técnicas de clasificación los resultados en el corte de 10 % son los mismos tanto para el modelo del experimento 4 como para el modelo con ranking de clientes modificado a través del valor de adopción social, pero a medida que los cortes incorporan una mayor proporción de clientes la diferencia entre los resultados de ambos modelos se va incrementando.

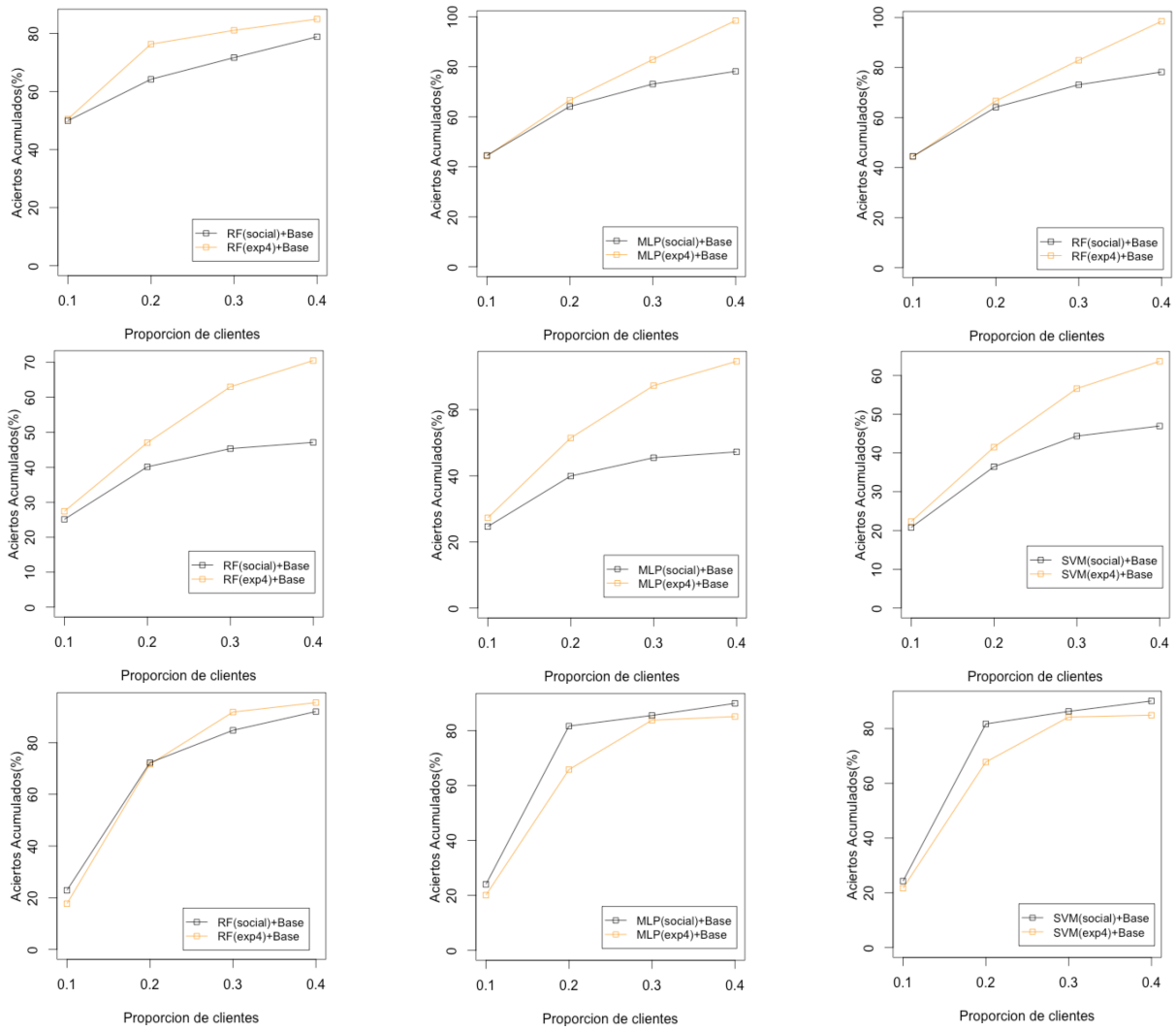


Figura 4.16: Comparación de aciertos acumulados entre el modelo entrenado con la combinación de los meses Julio y Agosto y el modelo que incluye el valor social del cliente para los productos Contrato(arriba), Cuenta Controlada(al medio) y Navegación en el Móvil(abajo)

4.9. Evaluación Económica

Incrementar la capacidad de predección de los modelos que buscan determinar qué clientes adoptarán un producto en un mes determinado permite seleccionar a un grupo de clientes adecuado para una campaña de marketing directo haciendo que ésta tenga un grado mayor de efectividad. En otras palabras, una compañía de telecomunicaciones que sea capaz de identificar con mayor certeza a sus clientes más propensos a aceptar una oferta logra un mayor número de ventas y por lo tanto logra un beneficio económico extra importante para el negocio.

Los resultados obtenidos por los modelos de targeting propuestos muestran que con su utilización es posible incrementar las ventas de los productos ya que son capaces de elevar la capacidad de predecir si un cliente adoptará o no un producto en relación al modelo de base con que se compara cuando se analiza el 40 % más alto del ranking de clientes. El detalle se muestra en la Tabla 4.35.

Producto	Aciertos Base	Aciertos Propuesto	Diferencia
Contrato	4.136	5.059	923
Cuenta Controlada	3.305	5.008	1.703
NEM	1.224	1.434	210

Tabla 4.35: Número de clientes identificados por los modelos de targeting en el 40 % más alto del ranking

Tomando en consideración estos incrementos potenciales en las ventas que se logran gracias a la selección más precisa de clientes para las campañas de marketing y el márgen de ganancias que cada producto deja por cada venta realizada es posible calcular el monto que la compañía puede obtener como ganancia extra por el hecho de utilizar la metodología y los modelos propuestos en este trabajo. El producto up-selling de Contrato deja un margen de $2.999, el up-selling de Cuenta Controlada un margen de $1.700 y el cross-selling de NEM un margen de $2.000, con lo que mensualmente se podrían generar ganancias promedio extra por aproximadamente $6.000.000 por los tres productos. Considerando este promedio mensual, se puede estimar que podrían lograrse ingresos extra anuales carcanos a $72.000.000 por los mismos tres productos. La Tabla 4.36 contiene todo el detalle de los cálculos.

Producto	Margen	Ganancia Mensual	Ganancia Anual
Contrato	$2.999	$2.770.000	$33.217.000
Cuenta Controlada	$1.700	$2.895.000	$34.741.000
NEM	$2.000	$420.000	$5.040.000

Tabla 4.36: Ganancias extra obtenidas por el incremento en las ventas dada la mejora de los modelos de targeting

Capítulo 5

Conclusiones y Trabajo Futuro

En un escenario tan competitivo como el de las industria de las telecomunicaciones es necesario que las compañías diseñen diversas estrategias para diferenciarse unas de otras. Una de las más reconocidas es la estrategia CRM, la cual desde su creación ha buscado establecer una relación rentable y de largo plazo con los clientes, sobretodo considerando que para las empresas es más costoso conseguir un nuevo cliente que retener uno activo. En este sentido, los esfuerzos deben enfocarse en seleccionar a los clientes adecuados que permitan elevar su valor a través de diferentes ofertas de up-selling o cross-selling de productos, y que por lo tanto sea necesario retener.

La propuesta clásica de las compañías de telecomunicaciones apunta a seleccionar a los clientes a través de diversas técnicas de clasificación basadas en los atributos más utilizados para perfilar a un cliente: atributos sociodemográficos y atributos comerciales. Sin embargo, los resultados obtenidos bajo este enfoque reflejan una baja tasa de respuesta positiva frente a la oferta por un producto. Dado lo anterior, se propone resolver el problema incorporando una nueva perspectiva del problema, la cual busca construir los mismos modelos de clasificación, pero esta vez incluyendo ciertos *atributos sociales* que describan la participación del cliente en su red social de teléfonos móviles, de manera de que a partir del impacto que puedan tener la red de contactos en las decisiones de los clientes, los modelos sean capaces de identificar de mejor manera a aquellos más propensos a aceptar una oferta por un producto.

En línea con lo anterior, se determinó que el objetivo central de este trabajo de memoria fuera el mejoramiento de un modelo de targeting de clientes para la adopción de productos a través de la incorporación de información extraída de una red social de teléfonos móviles, para lo cual se utilizaron diferentes herramientas de análisis de redes sociales y minería de datos, con el fin de realizar una serie de experimentos que buscaran establecer la configuración del modelo de targeting que mejores resultados conseguía. De esta manera es cómo a través de este trabajo se fueron estableciendo hitos y cumpliendo con los objetivos específicos propuetos en la sección 1.4 del primer Capítulo.

En el Capítulo 2 se construyó un breve reporte que resume los antecedentes más importantes en los que se basa este trabajo. En la sección 2.2 se presentaron las técnicas de clasificación más comunes junto con una serie de ejemplos en los cuáles éstas son utilizadas con éxito, para luego en la sección 2.3 describir cómo basándose en ellas es posible diseñar un modelo de targeting. Posteriormente, en la sección 2.4 se introduce el enfoque de Análisis de Redes Sociales y las distintas aristas que éste puede alcanzar, además de algunas aplicaciones que recientemente se le han podido dar a las ideas que plantea SNA. Finalmente, se hace una revisón de algunas estrategias que permiten combinar los conceptos que están detrás del targeting de clientes y de SNA para generar un nuevo modelo de targeting social, cuya metodología asociada es detallada a lo largo del capítulo 3.

Los primeros resultados que se obtuvieron reflejan mejoras en la identificación de los clientes que adoptan los productos ofrecidos por medio de una campaña de marketing, aunque no de la manera que se esperaba inicialmente. Si bien en un comienzo se pensaba que el modelo de targeting lograría mejores resultados por el sólo hecho de recibir más información a través de los atributos sociales, los dos primeros experimentos llevados a cabo demostraron que ésto no se cumplía ya que el poder de predicción del modelo social propuesto era similar o incluso peor que el del modelo base, fenómeno que se repitió para los tres productos en estudio, independiente de la técnica de clasificación utilizada. Sin embargo, al analizar los resultados desde el punto de vista de qué y no cuántos clientes estaban siendo identificados, se logró el primer descubrimiento clave del trabajo: el modelo de targeting social y el modelo base son capaces de seleccionar clientes con perfiles distintos. Esta importante conclusión permite que combinando los aciertos predichos por ambos modelos de targeting se logre incrementar de manera significativa el porcentaje de clientes que son identificados como adoptadores.

Considerando que la variación en el diseño de un modelo de targeting permite identificar clientes de diferentes perfiles, podría darse el caso que bajo la utilización de varios modelos -donde cada uno apuntara a un perfil distinto- se logre la identificación de la totalidad de los clientes que tienen propensión a la adopción de productos de telefonía móvil, lo que significaría un avance importante en el desempeño de las campañas de marketing directo ya que éstas tendrían un nivel de acierto muy cercano a 100 %, lo que se traduciría en un ahorro significativo para las compañías. Para que lo anterior se logre de manera eficiente, la intersección entre los conjuntos de clientes encontrados por los diferentes modelos debería ser lo más pequeña posible.

Una vez que se logra corroborar que existe un incremento en la capacidad de predicción al combinar los aciertos de los modelos de targeting social independiente de la técnica de clasificación utilizada, se quiso buscar una manera de lograr un nivel de aciertos aún mayor, por lo que se propuso la realización de una serie de experimentos que modifican la configuración del modelo de trageting respecto de los datos que se utilizan como conjunto de entrenamiento. Un segundo descubrimiento clave para este trabajo se logra al analizar los resultados de estos experimentos: configurando el conjunto de entrenamiento a partir de la combinación de datos del cliente correspondientes a meses previos al mes sobre el cual se desea predecir se generan resultados que reflejan una nueva mejora en el poder de predicción de los modelos

de targeting, lo que nuevamente se consigue para los tres productos estudiados, sin importar la técnica de clasificación utilizada.

A partir de lo anterior se puede desprender que el tiempo es un factor importante al momento de construir los modelos de targeting, ya que el hecho de incluir datos históricos consecutivos del cliente al conjunto de entrenamiento permite que el algoritmo de clasificación encuentre un patrón que representa una tendencia que se repite a lo largo del tiempo. A través de este patrón tendencial es que se puede llegar a definir el perfil del cliente y por ende inferir la decisión que éste podría tomar frente a una oferta por un producto.

Es importante que los grupos de datos que se combinan para formar el conjunto de entrenamiento agregado correspondan a periodos consecutivos ya que este requisito es el que permite que el modelo logre identificar el patrón tendencial que se esconde en los atributos asociados al cliente, de los contrario podría ocurrir lo que se observó en el experimento 3, donde al construirse el conjunto de entrenamiento a partir de un grupo de datos con dos meses de antigüedad (t-2) sin considerar los datos del mes siguiente a ese (t-1), se obtuvo una mejora no tan significativa como en el caso del experimento 4.

Los modelos de targeting estudiados en este trabajo fueron diseñados en base a tres técnicas de clasificación distintas: Random Forest, algoritmo perteneciente a la familia de los árboles de decisión; Multilayer Perceptron, una red neuronal con capacidad de aprendizaje; y Support Vector Machines, un método de aprendizaje supervisado que se basa en la idea de un hiperplano separador de clases. El objetivo de realizar pruebas usando esta variedad de técnicas era determinar cuál de ellas lograba el mayor poder de predicción de la adopción de productos, ante lo cual se obtuvieron importantes conclusiones respecto de los algoritmos predominantes en un problema como el presentado en esta memoria.

Si bien Random Forest es la técnica que en general presenta mejores resultados -en un 61 % de los casos estudiados es la que tiene el mayor poder de predicción- este comportamiento prevalece sólo cuando se consideran los aciertos del modelo social y no la combinación con el modelo base, lo que se evidencia en los experimentos 1 y 2, donde Random Forest resultó ser el algoritmo que destaca por sobre los otros en los tres productos estudiados cuando los resultados se analizaron de manera aislada. Sin embargo, cuando se toma en consideración la combinación de los aciertos, en los cuatro experimentos comienzan a lograr mejores resultados las técnicas MLP y SVM, fenómeno que demuestra que si bien estos algoritmos no logran identificar el mayor número de clientes, sí consiguen encontrar más clientes de un perfil distinto que Random Forest y el modelo base.

Lo anterior refuerza la idea de que es necesario trabajar con varios modelos complementarios que se enfoquen en clientes con perfiles distintos para lograr una predicción más eficaz y eficiente. Combinando modelos de targeting diseñados a partir de diferentes algoritmos de clasificación se podría construir un modelo agregado que elevara su capacidad de predicción a niveles cercanos al 100 %, permitiendo a la compañías incrementar el desempeño de sus campañas de marketing y sus planes comerciales.

Trabajo Futuro

Gracias a los resultados obtenidos a lo largo de la investigación se abren una serie de caminos por los cuales dar continuidad en el futuro a la búsqueda de mejores soluciones para el problema planteado en este trabajo.

Una de las direcciones más simples que se pueden adoptar para seguir desarrollando este estudio a futuro es la utilización de nuevas técnicas de clasificación que puedan adaptarse aún mejor a los datos disponibles para construir los modelos de targeting. El hecho de haber basado la investigación en los resultados entregados por sólo tres técnicas de clasificación abre grandes posibilidades de encontrar un nuevo método con el que se logre configurar un modelo de selección de clientes que consiga resultados óptimos para el problema.

Una nueva arista que también se podría explorar es la agregación de más meses de datos para entrenar los modelos de targeting con el fin de buscar mejores resultados que los entregados por el modelo que entrena con dos meses de datos. Incluso se podrían incorporar ponderadores que dieran mayor importancia a los conjuntos de datos de los meses más cercanos al mes que se busca predecir para que éstos influyan más al momento de ejecutar los modelos de selección.

El uso de los scores generados por los modelos de targeting como una variable más de los conjuntos de entrenamiento y testeo es otro tópico interesante para trabajar en el futuro. Esta modificación entregaría mayor información al modelo lo que podría permitirle mejorar de manera considerable su capacidad de predicción, lo que traería consigo la posibilidad de identificar a los clientes adoptadores más rápido de manera más eficaz.

Finalmente, un desafío futuro interesante sería el diseñar e incorporar una nueva metodología que permita incluir el valor social en el ranking de clientes distinta a la utilizada en este trabajo, ya que dado los resultados obtenidos queda claro que aún hay mucho camino por recorrer en el estudio de la influencia que puede llegar a tener el círculo cercano de un cliente sobre la decisiones comerciales que éste toma.

Bibliografía

[1] Bart Baesens, Rudy Setiono, Christophe Mues, and Jan Vanthienen. Using neural network rule extraction and decision tables for credit-risk evaluation. *Management science*, 49(3):312–329, 2003.

[2] Abraham Bagherjeiran and Rajesh Parekh. Combining behavioral and social network data for online advertising. In *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*, pages 837–846. IEEE, 2008.

[3] J.A Barnes. Human relations. *Class and committees in a Norwegian island parish, 7:39Đ58*, 1954.

[4] Francisco Barrientos and Sebastián A Ríos. Aplicación de minería de datos para predecir fuga de clientes en la industria de las telecomunicaciones.

[5] Rashmi Dutta Baruah and Plamen Angelov. Evolving social network analysis: A case study on mobile phone data. In *Evolving and Adaptive Intelligent Systems (EAIS), 2012 IEEE Conference on*, pages 114–120. IEEE, 2012.

[6] Mirta Bensic, Natasa Sarlija, and Marijana Zekic-Susac. Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance and Management*, 13(3):133–150, 2005.

[7] Michael J. Berry and Gordon Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Support*. John Wiley & Sons, Inc., New York, NY, USA, 1997.

[8] Rushi Bhatt, Vineet Chaoji, and Rajesh Parekh. Predicting product adoption in large-scale social networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1039–1048. ACM, 2010.

[9] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[10] Edgar F Borgatta. Jacob l. moreno and"sociometry": A mid-century reminiscence. *Social Psychology Quarterly*, pages 330–332, 2007.

[11] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[12] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification*

and regression trees. CRC press, 1984.

[13] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[14] Dragana Ćamilović, Dragana Bečejski-Vujaklija, and Nataša Gospić. A call detail records data mart: Data modeling and olap analysis. *Computer Science and Information Systems*, 6(2):87–110, 2009.

[15] Dorwin Cartwright and Frank Harary. Structural balance: a generalization of heider's theory. *Psychological review*, 63(5):277, 1956.

[16] Linda S Chan and Olive Jean Dunn. The treatment of missing values in discriminant analysisÑi. the sampling experiment. *Journal of the American Statistical Association*, 67(338):473–477, 1972.

[17] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000.

[18] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *arXiv preprint arXiv:1106.1813*, 2011.

[19] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004.

[20] Yetian Chen. Learning classifiers from imbalanced, only positive and unlabeled data sets. *Department of Computer Science, Iowa State University*, 2009.

[21] Yoon Ho Cho, Jae Kyeong Kim, and Soung Hie Kim. A personalized recommender system based on web usage mining and decision tree induction. *Expert Systems with Applications*, 23(3):329–342, 2002.

[22] Victoria J Cook, Sumi J Sun, Jane Tapia, Stephen Q Muth, D Fermín Argüello, Bryan L Lewis, Richard B Rothenberg, and Peter D McElroy. Transmission network analysis in tuberculosis contact investigations. *Journal of Infectious Diseases*, 196(10):1517–1527, 2007.

[23] Banco Central de Chile. Cuentas nacionales de chile, evolución de la actividad económica en el año 2013. 2013.

[24] Wouter De Nooy, Andrej Mrvar, and Vladimir Batagelj. *Exploratory social network analysis with Pajek*, volume 27. Cambridge University Press, 2011.

[25] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

[26] Harris Drucker, S Wu, and Vladimir N Vapnik. Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on*, 10(5):1048–1054, 1999.

[27] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Mobile phone data for inferring social network structure. In *Social computing, behavioral modeling, and prediction*, pages 79–88. Springer, 2008.

[28] David Easley and Jon Kleinberg. Networks, crowds, and markets. *Cambridge Univ Press*, 6(1):6–1, 2010.

[29] Tom Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine learning*, 31:1–38, 2004.

[30] Patrik Finne, Ralf Finne, Anssi Auvinen, Harri Juusela, Jussi Aro, Liisa Määttänen, Matti Hakama, Sakari Rannikko, Teuvo LJ Tammela, and Ulf-Håkan Stenman. Predicting the outcome of prostate biopsy in screen-positive men by a multilayer perceptron network. *Urology*, 56(3):418–422, 2000.

[31] Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In *ICML*, volume 99, pages 124–133, 1999.

[32] Mark A Friedl and Carla E Brodley. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3):399–409, 1997.

[33] Terrence S Furey, Nello Cristianini, Nigel Duffy, David W Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[34] Wendy Gersten, Rüdiger Wirth, and Dirk Arndt. Predictive modeling in automotive direct marketing: tools, experiences and open issues. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 398–406. ACM, 2000.

[35] Giorgio Giacinto and Fabio Roli. Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, 19(9):699–707, 2001.

[36] Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R Sveinsson. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006.

[37] Noah D Goodman, Joshua B Tenenbaum, Jacob Feldman, and Thomas L Griffiths. A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154, 2008.

[38] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[39] Creighton TR Hager. *Statistical Analysis of ATM Call Detail Records*. PhD thesis, Citeseer, 2000.

[40] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

[41] Frank Harary. Graph theory. 1969.

[42] Ernst Haselsteiner and Gert Pfurtscheller. Using time-dependent neural networks for eeg classification. *Rehabilitation Engineering, IEEE Transactions on*, 8(4):457–463, 2000.

[43] Fritz Heider. The gestalt theory of motivation. In *Nebraska symposium on motivation*, volume 8, pages 145–172, 1960.

[44] Cesar A Hidalgo and C Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024, 2008.

[45] Shawndra Hill, Foster Provost, and Chris Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, pages 256–276, 2006.

[46] Xiaohua Hu. A data mining approach for retailing bank customer attrition analysis. *Applied Intelligence*, 22(1):47–60, 2005.

[47] Bing Quan Huang, T-M Kechadi, Brian Buckley, G Kiernan, E Keogh, and Tarik Rashid. A new feature set with new window techniques for customer churn prediction in land-line telecommunications. *Expert Systems with Applications*, 37(5):3657–3665, 2010.

[48] Bingquan Huang, Mohand Tahar Kechadi, and Brian Buckley. Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1):1414–1425, 2012.

[49] C Huang, LS Davis, and JRG Townshend. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23(4):725–749, 2002.

[50] Earl B Hunt, Janet Marin, and Philip J Stone. Experiments in induction. 1966.

[51] Thorsten Joachims. *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[52] Judith K Jones. The role of data mining technology in the identification of signals of possible adverse drug reactions: value and limitations. *Current therapeutic research*, 62(9):664–672, 2001.

[53] Wagner A Kamakura, Michel Wedel, Fernando De Rosa, and Jose Afonso Mazzon. Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in marketing*, 20(1):45–65, 2003.

[54] Dimitrios Katsaros, Nikos Dimokas, and Leandros Tassiulas. Social network analysis concepts in the design of wireless ad hoc network protocols. *Network, IEEE*, 24(6):23–29, 2010.

[55] YongSeog Kim and W Nick Street. An intelligent system for customer targeting: a data

mining approach. *Decision Support Systems*, 37(2):215–228, 2004.

[56] Sotiris B Kotsiantis. Supervised machine learning: a review of classification techniques. *Informatica (03505596)*, 31(3), 2007.

[57] Roland Kuhn and Renato De Mori. The application of semantic classification trees to natural language understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(5):449–460, 1995.

[58] Edward O Laumann. Networks of collective action: A perspective on community influence systems (quantitative studies in social relations). 1976.

[59] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *Neural Networks, IEEE Transactions on*, 8(1):98–113, 1997.

[60] Raymond Ling and David C Yen. Customer relationship management: An analysis framework and implementation strategies. *Journal of Computer Information Systems*, 41(3):82–97, 2001.

[61] Lautaro Cuadra Lobos. Metodologia de busqueda de sub-comunidades mediante analisis de redes sociales y mineria de datos. 2012.

[62] Ilias G Maglogiannis. *Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies*, volume 160. Ios Press, 2007.

[63] Francesco Martino and Andrea Spoto. Social network analysis: A brief theoretical review and further perspectives in the study of information technology. *PsychNology Journal*, 4(1):53–86, 2006.

[64] Philipp Michel and Rana El Kaliouby. Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264. ACM, 2003.

[65] Ricardo Munoz. Diseesarrollo y evaluacion de un algortimo para detectar subcomunidades traslapadas usando analisis de redes sociales y mineria de datos. 2013.

[66] Costas Neocleous and Christos Schizas. Artificial neural network learning: A comparative review. In *Methods and Applications of Artificial Intelligence*, pages 300–313. Springer, 2002.

[67] Eric WT Ngai, Li Xiu, and Dorothy CK Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602, 2009.

[68] Lucila Ohno-Machado, Ronilda Lacson, and Eduardo Massad. Decision trees and fuzzy logic: a comparison of models for the selection of measles vaccination strategies in brazil. In *Proceedings of the AMIA Symposium*, page 625. American Medical Informatics

Association, 2000.

[69] Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, Gábor Szabó, M Argollo De Menezes, Kimmo Kaski, Albert-László Barabási, and János Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179, 2007.

[70] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[71] Philippa Pattison. *Algebraic models for social networks*. Number 7. Cambridge University Press, 1993.

[72] Adrian Payne and Pennie Frow. A strategic framework for customer relationship management. *Journal of marketing*, 69(4):167–176, 2005.

[73] F.C. Perkins. System and method for processing call detail records, May 28 2002. US Patent 6,396,913.

[74] Vili Podgorelec, Peter Kokol, Bruno Stiglic, and Ivan Rozman. Decision trees: an overview and their use in medicine. *Journal of medical systems*, 26(5):445–463, 2002.

[75] Werner Reinartz, Manfred Krafft, and Wayne D Hoyer. The customer relationship management process: its measurement and impact on performance. *Journal of marketing research*, 41(3):293–305, 2004.

[76] Irina Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[77] Lior Rokach. *Data mining with decision trees: theory and applications*, volume 69. World scientific, 2008.

[78] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[79] Alberto Maria Segre and Geoffrey J. Gordon. Sholom m. weiss and casimir a. kulikowski, computer systems that learn. *Artif. Intell.*, 62(2):363–378, 1993.

[80] Ministerior de Transportes y Telecomunicaciones Subsecretaría de Telecomunicaciones de Chile. Informe de análisis del sector telecomunicaciones. Diciembre 2013.

[81] Christine L Tsien, HS Fraser, William J Long, and R Lee Kennedy. Using classification tree and logistic regression methods to diagnose myocardial infarction. *Studies in health technology and informatics*, (1):493–497, 1998.

[82] Ivan F Videla-Cavieres and Sebastián A Ríos. Extending market basket analysis with graph mining techniques: A real case. *Expert Systems with Applications*, 41(4):1928–1936, 2014.

[83] Stanley Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[84] Art Weinstein. *Market Segmentation: Using Demographics, Psychographics and Other Niche Marketing Techniques to Predict and Model Customer Behavior*. Probus Chicago, IL, 1994.

[85] Barry Wellman. Structural analysis: From method and metaphor to theory and substance. *Contemporary Studies in Sociology*, 15:19–61, 1997.

[86] David West. Neural network credit scoring models. *Computers & Operations Research*, 27(11):1131–1152, 2000.

[87] Ji Young Woo, Sung Min Bae, and Sang Chan Park. Visualization method for customer targeting using customer map. *Expert Systems with Applications*, 28(4):763–772, 2005.

[88] Jennifer J Xu and Hsinchun Chen. Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks. *Decision Support Systems*, 38(3):473–487, 2004.

[89] B Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

[90] Guoqiang Zhang, Michael Y Hu, B Eddy Patuwo, and Daniel C Indro. Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European journal of operational research*, 116(1):16–32, 1999.

[91] Boris Zmazek, Ljupčo Todorovski, Sašo Džeroski, Janja Vaupotič, and Ivan Kobal. Application of decision trees to the analysis of soil radon data for earthquake prediction. *Applied Radiation and Isotopes*, 58(6):697–706, 2003.

Anexos

a. Detalle de resultados de modelos de targeting

Experimento 1.1: Modelo entrenado con Julio y testeado con Agosto[1]

Decision Tree			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	967	654	279	1.342
		S	841		189	1.306
	CC	U	1.022	995	318	1.699
		S	693		147	1.541
	NEM	U	98	125	2	221
		S	154		12	267
20 %	CTO	U	1.500	1176	674	2.002
		S	1.270		531	1.915
	CC	U	1.895	2210	943	3.162
		S	1.407		558	3.059
	NEM	U	209	649	66	792
		S	282		139	792
30 %	CTO	U	1.841	1692	1.057	2.476
		S	1.637		871	2.458
	CC	U	2.693	3041	1.595	4.139
		S	2.193		1.106	4.128
	NEM	U	338	885	186	1.037
		S	399		269	1.015
40 %	CTO	U	2.155	2061	1.399	2.817
		S	1.985		1.180	2.866
	CC	U	3.495	3781	2.298	4.978
		S	2.985		1.740	5.026
	NEM	U	450	904	264	1.090
		S	516		353	1.067

[1]Nomenclatura de las tablas: CTO corresponde al producto Contrato. CC a Cuenta Controlada y NEM a Navegación en el Móvil; U corresponde al método de balanceo under-sampling y S al método SMOTE.

Neural Network			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	480	654	97	1.037
		S	375		57	972
	CC	U	918	995	148	1.765
		S	413		49	1.359
	NEM	U	62	125	7	306
		S	67		6	331
20 %	CTO	U	823	1176	328	1.671
		S	737		241	1.672
	CC	U	1.833	2210	617	3.426
		S	1.062		301	2.971
	NEM	U	188	649	90	881
		S	212		106	879
30 %	CTO	U	1.339	1692	717	2.314
		S	1.189		543	2.338
	CC	U	2.597	3041	1.162	4.476
		S	1.664		646	4.059
	NEM	U	322	885	214	993
		S	336		224	997
40 %	CTO	U	1.791	2061	1.072	2.780
		S	1.599		870	2.790
	CC	U	3.220	3781	1.774	5.227
		S	2.458		1.243	4.996
	NEM	U	490	904	342	1.052
		S	507		359	1.052

Support Vector Machines			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	376	654	105	925
		S	345		78	921
	CC	U	823	995	134	1.684
		S	871		142	1.724
	NEM	U	65	125	7	183
		S	84		9	200
20 %	CTO	U	712	1176	224	1.664
		S	767		206	1.737
	CC	U	1.498	2210	537	3.171
		S	1.656		551	3.315
	NEM	U	207	649	102	754
		S	218		109	758
30 %	CTO	U	1.126	1692	508	2.310
		S	1.207		497	2.402
	CC	U	2.131	3041	980	4.192
		S	2.137		915	4.263
	NEM	U	327	885	217	995
		S	341		229	997
40 %	CTO	U	1.537	2061	807	2.791
		S	1.698		862	2.897
	CC	U	2.734	3781	1.465	5.050
		S	2.621		1.297	5.105
	NEM	U	499	904	351	1.052
		S	519		371	1.052

	Modelo Propuesto - Random Forest												
	RF	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	23.82 %	2.97 %	5.28 %	36.95 %	2.30 %	4.33 %	45.36 %	1.88 %	3.62 %	53.09 %	1.65 %	3.21 %
	SMOTE	20.72 %	2.58 %	4.59 %	31.29 %	1.95 %	3.67 %	40.33 %	1.68 %	3.22 %	48.90 %	1.52 %	2.95 %
CC	Under	13.90 %	5.17 %	7.54 %	25.77 %	4.80 %	8.09 %	36.62 %	4.54 %	8.09 %	47.53 %	4.42 %	8.09 %
	SMOTE	9.42 %	3.51 %	5.11 %	19.13 %	3.56 %	6.01 %	29.82 %	3.70 %	6.58 %	40.59 %	3.78 %	6.91 %
NEM	Under	7.63 %	0.11 %	0.22 %	16.26 %	0.12 %	0.23 %	26.30 %	0.13 %	0.25 %	35.02 %	0.13 %	0.25 %
	SMOTE	11.98 %	0.17 %	0.34 %	21.95 %	0.16 %	0.32 %	31.05 %	0.15 %	0.30 %	40.16 %	0.15 %	0.29 %

	Modelo Propuesto - Multilayer Perceptron												
	MLP	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	23.64 %	2.89 %	5.15 %	30.71 %	1.88 %	3.54 %	49.29 %	2.01 %	3.86 %	56.66 %	1.73 %	3.36 %
	SMOTE	21.06 %	2.57 %	4.59 %	33.84 %	2.07 %	3.90 %	46.55 %	1.90 %	3.64 %	56.76 %	1.73 %	3.36 %
CC	Under	16.27 %	5.80 %	8.55 %	30.88 %	5.50 %	9.34 %	43.72 %	5.20 %	9.29 %	55.50 %	4.95 %	9.08 %
	SMOTE	12.72 %	4.54 %	6.69 %	24.43 %	4.36 %	7.39 %	35.00 %	4.16 %	7.44 %	46.24 %	4.12 %	7.57 %
NEM	Under	11.38 %	0.20 %	0.39 %	25.63 %	0.23 %	0.45 %	36.55 %	0.21 %	0.43 %	53.26 %	0.23 %	0.47 %
	SMOTE	11.58 %	0.20 %	0.40 %	25.83 %	0.23 %	0.45 %	36.42 %	0.21 %	0.42 %	53.40 %	0.23 %	0.47 %

	Modelo Propuesto - Support Vector Machines												
	SVM	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	9.03 %	1.10 %	1.97 %	18.69 %	1.14 %	2.15 %	28.25 %	1.15 %	2.21 %	41.56 %	1.27 %	2.46 %
	SMOTE	10.45 %	1.28 %	2.27 %	20.91 %	1.28 %	2.41 %	32.39 %	1.32 %	2.54 %	46.58 %	1.42 %	2.76 %
CC	Under	10.05 %	3.58 %	5.28 %	20.96 %	3.74 %	6.34 %	30.60 %	3.64 %	6.50 %	37.11 %	3.31 %	6.07 %
	SMOTE	10.77 %	3.84 %	5.66 %	19.28 %	3.44 %	5.83 %	25.76 %	3.06 %	5.47 %	36.85 %	3.28 %	6.03 %
NEM	Under	11.85 %	0.21 %	0.41 %	26.36 %	0.23 %	0.46 %	37.22 %	0.22 %	0.43 %	54.13 %	0.24 %	0.47 %
	SMOTE	12.32 %	0.22 %	0.43 %	25.90 %	0.23 %	0.45 %	37.95 %	0.22 %	0.44 %	54.26 %	0.24 %	0.47 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Random Forest												
	RF	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	49.94 %	6.10 %	10.87 %	64.21 %	3.92 %	7.39 %	71.67 %	2.92 %	5.61 %	78.84 %	2.41 %	4.67 %
	SMOTE	50.70 %	6.19 %	11.04 %	63.81 %	3.90 %	7.35 %	70.78 %	2.88 %	5.54 %	77.36 %	2.36 %	4.59 %
CC	Under	25.12 %	8.96 %	13.21 %	40.11 %	7.15 %	12.14 %	45.32 %	5.39 %	9.63 %	47.11 %	4.20 %	7.71 %
	SMOTE	24.67 %	8.80 %	12.97 %	39.66 %	7.07 %	12.00 %	44.03 %	5.23 %	9.35 %	47.26 %	4.21 %	7.74 %
NEM	Under	22.90 %	0.40 %	0.79 %	72.24 %	0.63 %	1.26 %	107.86 %	0.63 %	1.26 %	92.01 %	0.40 %	0.80 %
	SMOTE	23.70 %	0.42 %	0.82 %	71.84 %	0.63 %	1.25 %	102.86 %	0.60 %	1.20 %	89.81 %	0.39 %	0.78 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Multilayer Perceptron												
	MLP	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	46.94 %	5.73 %	10.22 %	61.09 %	3.73 %	7.03 %	69.86 %	2.85 %	5.47 %	77.36 %	2.36 %	4.59 %
	SMOTE	46.30 %	5.66 %	10.08 %	61.07 %	3.73 %	7.03 %	69.84 %	2.84 %	5.47 %	77.16 %	2.36 %	4.57 %
CC	Under	24.66 %	8.79 %	12.96 %	39.92 %	7.12 %	12.08 %	45.46 %	5.40 %	9.66 %	47.23 %	4.21 %	7.73 %
	SMOTE	22.93 %	8.17 %	12.05 %	37.76 %	6.73 %	11.43 %	43.57 %	5.18 %	9.26 %	47.08 %	4.20 %	7.71 %
NEM	Under	23.83 %	0.42 %	0.82 %	81.49 %	0.72 %	1.42 %	97.34 %	0.57 %	1.13 %	89.88 %	0.39 %	0.79 %
	SMOTE	24.03 %	0.42 %	0.83 %	81.62 %	0.72 %	1.42 %	97.14 %	0.57 %	1.13 %	89.88 %	0.39 %	0.79 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Support Vector Machines												
	SVM	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	44.53 %	5.44 %	9.70 %	64.13 %	3.92 %	7.38 %	73.10 %	2.98 %	5.72 %	78.19 %	2.39 %	4.63 %
	SMOTE	45.02 %	5.50 %	9.80 %	63.09 %	3.85 %	7.26 %	71.52 %	2.91 %	5.60 %	76.20 %	2.33 %	4.52 %
CC	Under	20.79 %	7.41 %	10.93 %	36.45 %	6.50 %	11.03 %	44.35 %	5.27 %	9.42 %	46.95 %	4.18 %	7.68 %
	SMOTE	21.10 %	7.52 %	11.09 %	35.27 %	6.29 %	10.67 %	42.05 %	5.00 %	8.93 %	47.07 %	4.19 %	7.70 %
NEM	Under	24.30 %	0.43 %	0.84 %	81.69 %	0.72 %	1.42 %	97.94 %	0.57 %	1.14 %	90.08 %	0.40 %	0.79 %
	SMOTE	24.77 %	0.43 %	0.85 %	81.29 %	0.71 %	1.41 %	98.14 %	0.57 %	1.14 %	90.01 %	0.40 %	0.79 %

Experimento 2: Modelo entrenado con Agosto y testeado con Septiembre[2]

Decision Tree			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	1.925	2.201	1.228	2.898
		S	2.048		1.289	2.960
	CC	U	1.074	991	165	1.900
		S	1.085		119	1.957
	NEM	U	450	194	126	518
		S	388		102	480
20 %	CTO	U	3.299	3.200	2.819	3.680
		S	3.296		2.802	3.694
	CC	U	2.096	1.895	636	3.355
		S	2.039		548	3.386
	NEM	U	742	918	655	1.005
		S	652		603	967
30 %	CTO	U	4.001	3.746	3.601	4.146
		S	3.899		3.574	4.071
	CC	U	2.983	2.619	1.251	4.351
		S	2.927		1.153	4.393
	NEM	U	942	1.206	901	1.247
		S	855		846	1.215
40 %	CTO	U	4.319	4.136	4.008	4.447
		S	4.244		3.973	4.407
	CC	U	3.780	3.305	1.965	5.120
		S	3.697		1.851	5.151
	NEM	U	1.066	1.224	1.013	1.277
		S	1.006		976	1.254

[2]Nomenclatura de las tablas: CTO corresponde al producto Contrato. CC a Cuenta Controlada y NEM a Navegación en el Móvil; U corresponde al método de balanceo under-sampling y S al método SMOTE.

Neural Network			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	1.294	2.201	708	2.787
		S	1.653		886	2.968
	CC	U	995	991	169	1.817
		S	999		156	1.834
	NEM	U	146	194	25	315
		S	131		20	305
20 %	CTO	U	2.372	3.200	1.870	3.702
		S	2.565		1.909	3.856
	CC	U	1.910	1.895	569	3.236
		S	1.958		577	3.276
	NEM	U	304	918	233	989
		S	237		181	974
30 %	CTO	U	3.142	3.746	2.754	4.134
		S	3.656		3.139	4.263
	CC	U	2.822	2.619	1.128	4.313
		S	2.930		1.206	4.343
	NEM	U	436	1.206	398	1.244
		S	372		336	1.242
40 %	CTO	U	3.870	4.136	3.490	4.516
		S	3.974		3.525	4.585
	CC	U	3.761	3.305	1.903	5.163
		S	3.763		1.878	5.190
	NEM	U	567	1.224	532	1.259
		S	519		498	1.245

Support Vector Machines			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	494	2.201	101	2.594
		S	558		153	2.606
	CC	U	670	991	97	1.564
		S	739		104	1.626
	NEM	U	151	194	16	329
		S	155		16	333
20 %	CTO	U	1.096	3.200	580	3.716
		S	1.146		664	3.682
	CC	U	1.396	1.895	368	2.923
		S	1.493		387	3.001
	NEM	U	313	918	219	1.012
		S	315		227	1.006
30 %	CTO	U	1.720	3.746	1.216	4.250
		S	1.880		1.415	4.211
	CC	U	2.081	2.619	748	3.952
		S	1.935		657	3.897
	NEM	U	466	1.206	426	1.246
		S	469		428	1.247
40 %	CTO	U	2.539	4.136	2.151	4.524
		S	2.740		2.397	4.479
	CC	U	2.535	3.305	1.117	4.723
		S	2.517		1.172	4.650
	NEM	U	611	1.224	565	1.270
		S	617		565	1.276

	Modelo Propuesto - Random Forest											
	RF	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	32.33 %	3.95 %	7.04 %	61.80 %	3.78 %	7.12 %	69.62 %	2.84 %	5.45 %	74.68 %	2.28 %	4.43 %
	SMOTE	34.39 %	4.20 %	7.49 %	62.03 %	3.79 %	7.14 %	68.36 %	2.78 %	5.35 %	74.01 %	2.26 %	4.39 %
CC	Under	15.12 %	5.39 %	7.95 %	47.22 %	8.42 %	14.29 %	61.24 %	7.28 %	13.01 %	72.06 %	6.42 %	11.79 %
	SMOTE	15.27 %	5.44 %	8.03 %	47.66 %	8.50 %	14.42 %	61.83 %	7.35 %	13.13 %	72.50 %	6.46 %	11.87 %
NEM	Under	29.96 %	0.53 %	1.03 %	66.91 %	0.59 %	1.16 %	83.02 %	0.49 %	0.97 %	85.02 %	0.37 %	0.74 %
	SMOTE	25.83 %	0.45 %	0.89 %	64.38 %	0.57 %	1.12 %	80.89 %	0.47 %	0.94 %	83.49 %	0.37 %	0.73 %

	Modelo Propuesto - Multilayer Perceptron											
	MLP	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	21.73 %	2.65 %	4.73 %	39.83 %	2.43 %	4.59 %	52.76 %	2.15 %	4.13 %	64.99 %	1.99 %	3.85 %
	SMOTE	27.76 %	3.39 %	6.04 %	43.07 %	2.63 %	4.96 %	61.39 %	2.50 %	4.80 %	66.73 %	2.04 %	3.96 %
CC	Under	14.00 %	4.99 %	7.36 %	26.88 %	4.79 %	8.13 %	39.72 %	4.72 %	8.44 %	52.93 %	4.72 %	8.66 %
	SMOTE	14.06 %	5.01 %	7.39 %	27.56 %	4.91 %	8.34 %	41.24 %	4.90 %	8.76 %	52.96 %	4.72 %	8.67 %
NEM	Under	9.72 %	0.17 %	0.34 %	20.24 %	0.18 %	0.35 %	29.03 %	0.17 %	0.34 %	37.75 %	0.17 %	0.33 %
	SMOTE	8.72 %	0.15 %	0.30 %	15.78 %	0.14 %	0.27 %	24.77 %	0.14 %	0.29 %	34.55 %	0.15 %	0.30 %

	Modelo Propuesto - Support Vector Machines											
	SVM	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	8.30 %	1.01 %	1.81 %	18.40 %	1.12 %	2.12 %	28.88 %	1.18 %	2.26 %	42.64 %	1.30 %	2.53 %
	SMOTE	9.37 %	1.14 %	2.04 %	19.24 %	1.18 %	2.22 %	31.57 %	1.29 %	2.47 %	46.01 %	1.41 %	2.73 %
CC	Under	9.43 %	3.36 %	4.96 %	19.65 %	3.50 %	5.95 %	29.29 %	3.48 %	6.22 %	35.68 %	3.18 %	5.84 %
	SMOTE	10.40 %	3.71 %	5.47 %	21.01 %	3.75 %	6.36 %	27.23 %	3.24 %	5.79 %	35.43 %	3.16 %	5.80 %
NEM	Under	10.05 %	0.18 %	0.35 %	20.84 %	0.18 %	0.36 %	31.03 %	0.18 %	0.36 %	40.68 %	0.18 %	0.36 %
	SMOTE	10.32 %	0.18 %	0.36 %	20.97 %	0.18 %	0.36 %	31.23 %	0.18 %	0.36 %	41.08 %	0.18 %	0.36 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Random Forest											
	RF	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	48.66 %	5.95 %	10.60 %	61.80 %	3.78 %	7.12 %	69.62 %	2.84 %	5.45 %	74.68 %	2.28 %	4.43 %
	SMOTE	49.71 %	6.07 %	10.82 %	62.03 %	3.79 %	7.14 %	68.36 %	2.78 %	5.35 %	74.01 %	2.26 %	4.39 %
CC	Under	26.74 %	9.53 %	14.06 %	47.22 %	8.42 %	14.29 %	61.24 %	7.28 %	13.01 %	72.06 %	6.42 %	11.79 %
	SMOTE	27.54 %	9.82 %	14.48 %	47.66 %	8.50 %	14.42 %	61.83 %	7.35 %	13.13 %	72.50 %	6.46 %	11.87 %
NEM	Under	34.49 %	0.61 %	1.19 %	66.91 %	0.59 %	1.16 %	83.02 %	0.49 %	0.97 %	85.02 %	0.37 %	0.74 %
	SMOTE	31.96 %	0.56 %	1.10 %	64.38 %	0.57 %	1.12 %	80.89 %	0.47 %	0.94 %	83.49 %	0.37 %	0.73 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Multilayer Perceptron											
	MLP	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	46.80 %	5.72 %	10.19 %	62.17 %	3.80 %	7.16 %	69.42 %	2.83 %	5.43 %	75.84 %	2.32 %	4.50 %
	SMOTE	49.84 %	6.09 %	10.85 %	64.75 %	3.96 %	7.46 %	71.59 %	2.92 %	5.60 %	76.99 %	2.35 %	4.56 %
CC	Under	25.57 %	9.12 %	13.44 %	45.55 %	8.12 %	13.78 %	60.70 %	7.21 %	12.90 %	72.67 %	6.48 %	11.89 %
	SMOTE	25.81 %	9.20 %	13.57 %	46.11 %	8.22 %	13.95 %	61.13 %	7.26 %	12.99 %	73.05 %	6.51 %	11.96 %
NEM	Under	20.97 %	0.37 %	0.72 %	65.85 %	0.58 %	1.15 %	82.82 %	0.48 %	0.96 %	83.82 %	0.37 %	0.73 %
	SMOTE	20.31 %	0.36 %	0.70 %	64.85 %	0.57 %	1.13 %	82.69 %	0.48 %	0.96 %	82.89 %	0.36 %	0.72 %

	Modelo Combinado (Intersección entre Propuesto y Empresa)- Support Vector Machines											
	SVM	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	43.56 %	5.32 %	9.49 %	62.40 %	3.81 %	7.19 %	71.37 %	2.91 %	5.59 %	75.97 %	2.32 %	4.50 %
	SMOTE	43.76 %	5.35 %	9.53 %	61.83 %	3.78 %	7.12 %	70.71 %	2.88 %	5.53 %	75.21 %	2.30 %	4.46 %
CC	Under	22.01 %	7.85 %	11.57 %	41.14 %	7.33 %	12.45 %	55.62 %	6.61 %	11.82 %	66.47 %	5.92 %	10.88 %
	SMOTE	22.89 %	8.16 %	12.03 %	42.24 %	7.53 %	12.78 %	54.85 %	6.52 %	11.65 %	65.45 %	5.83 %	10.71 %
NEM	Under	21.90 %	0.38 %	0.76 %	67.38 %	0.59 %	1.17 %	82.96 %	0.49 %	0.97 %	84.55 %	0.37 %	0.74 %
	SMOTE	22.17 %	0.39 %	0.77 %	66.98 %	0.59 %	1.17 %	83.02 %	0.49 %	0.97 %	84.95 %	0.37 %	0.74 %

Experimento 3: Modelo entrenado con Julio y testeado con Septiembre[3]

Decision Tree			Estudio	Compañía	Intersección	Únicos
10 %	CC	U	1.956	2.201	1.172	2.985
		S	1.579		951	2.829
	HIB	U	1.076	991	236	1.831
		S	925		211	1.705
	NEM	U	144	194	6	332
		S	184		26	352
20 %	CC	U	2.955	3.200	2.215	3.940
		S	2.480		1.976	3.704
	HIB	U	2.011	1.895	693	3.213
		S	1.848		611	3.132
	NEM	U	280	918	156	1.042
		S	341		238	1.021
30 %	CC	U	3.634	3.746	2.918	4.462
		S	3.127		2.732	4.141
	HIB	U	2.905	2.619	1.231	4.293
		S	2.710		1.149	4.180
	NEM	U	414	1.206	308	1.312
		S	495		442	1.259
40 %	CC	U	4.113	4.136	3.346	4.903
		S	3.660		3.236	4.560
	HIB	U	3.691	3.305	1.847	5.149
		S	3.503		1.753	5.055
	NEM	U	545	1.224	437	1.332
		S	682		600	1.306

[3]Nomenclatura de las tablas: CTO corresponde al producto Contrato. CC a Cuenta Controlada y NEM a Navegación en el Móvil; U corresponde al método de balanceo under-sampling y S al método SMOTE.

Neural Network			Estudio	Compañía	Intersección	Únicos
10 %	CC	U	982	2.201	502	2.681
		S	664		306	2.559
	HIB	U	444	991	81	1.354
		S	796		156	1.631
	NEM	U	146	194	25	315
		S	119		20	293
20 %	CC	U	1.622	3.200	1.125	3.697
		S	1.363		938	3.625
	HIB	U	1.314	1.895	413	2.796
		S	1.528		450	2.973
	NEM	U	302	918	229	991
		S	281		221	978
30 %	CC	U	2.095	3.746	1.465	4.376
		S	1.929		1.476	4.199
	HIB	U	2.317	2.619	918	4.018
		S	2.176		825	3.970
	NEM	U	472	1.206	439	1.239
		S	447		410	1.243
40 %	CC	U	2.763	4.136	2.024	4.875
		S	2.647		2.115	4.668
	HIB	U	3.175	3.305	1.517	4.963
		S	2.964		1.400	4.869
	NEM	U	643	1.224	617	1.250
		S	605		573	1.256

Support Vector Machines			Estudio	Caompañía	Intersección	Únicos
10 %	CC	U	297	2201	1	2497
		S	402		118	2485
	HIB	U	710	991	79	1622
		S	770		87	1674
	NEM	U	142	194	22	314
		S	169		27	336
20 %	CC	U	612	3200	13	3799
		S	842		372	3670
	HIB	U	1.498	1895	400	2993
		S	1.377		348	2924
	NEM	U	296	918	214	1000
		S	316		240	994
30 %	CC	U	969	3746	89	4626
		S	1.297		682	4361
	HIB	U	2.041	2619	720	3940
		S	1.836		634	3821
	NEM	U	463	1206	422	1247
		S	483		446	1243
40 %	CC	U	1.476	4136	540	5072
		S	1.690		1021	4805
	HIB	U	2.526	3305	1149	4682
		S	2.371		1114	4562
	NEM	U	627	1224	596	1255
		S	654		619	1259

	Modelo Propuesto - Random Forest											
	RF	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	32.85 %	4.01 %	7.15 %	49.62 %	3.03 %	5.71 %	61.02 %	2.49 %	4.78 %	69.07 %	2.11 %	4.09 %
	SMOTE	26.52 %	3.24 %	5.77 %	41.65 %	2.54 %	4.80 %	52.51 %	2.14 %	4.11 %	61.46 %	1.88 %	3.64 %
CC	Under	15.14 %	5.40 %	7.96 %	28.30 %	5.05 %	8.56 %	40.89 %	4.86 %	8.69 %	51.95 %	4.63 %	8.50 %
	SMOTE	13.02 %	4.64 %	6.84 %	26.01 %	4.64 %	7.87 %	38.14 %	4.53 %	8.10 %	49.30 %	4.39 %	8.07 %
NEM	Under	9.59 %	0.17 %	0.33 %	18.64 %	0.16 %	0.32 %	27.56 %	0.16 %	0.32 %	36.28 %	0.16 %	0.32 %
	SMOTE	12.25 %	0.22 %	0.42 %	22.70 %	0.20 %	0.40 %	32.96 %	0.19 %	0.38 %	45.41 %	0.20 %	0.40 %

	Modelo Propuesto - Multilayer Perceptron											
	MLP	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	16.49 %	2.01 %	3.59 %	27.24 %	1.66 %	3.14 %	35.18 %	1.43 %	2.75 %	46.40 %	1.42 %	2.75 %
	SMOTE	11.15 %	1.36 %	2.43 %	22.89 %	1.40 %	2.64 %	32.39 %	1.32 %	2.54 %	44.45 %	1.36 %	2.63 %
CC	Under	6.25 %	2.23 %	3.28 %	18.49 %	3.30 %	5.60 %	32.61 %	3.88 %	6.93 %	44.69 %	3.98 %	7.31 %
	SMOTE	11.20 %	3.99 %	5.89 %	21.51 %	3.83 %	6.51 %	30.63 %	3.64 %	6.51 %	41.72 %	3.72 %	6.83 %
NEM	Under	9.72 %	0.17 %	0.34 %	20.11 %	0.18 %	0.35 %	31.42 %	0.18 %	0.37 %	42.81 %	0.19 %	0.37 %
	SMOTE	7.92 %	0.14 %	0.27 %	18.71 %	0.16 %	0.33 %	29.76 %	0.17 %	0.35 %	40.28 %	0.18 %	0.35 %

	Modelo Propuesto - Support Vector Machines											
	SVM	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	4.99 %	0.61 %	1.09 %	10.28 %	0.63 %	1.18 %	16.27 %	0.66 %	1.27 %	24.79 %	0.76 %	1.47 %
	SMOTE	6.75 %	0.82 %	1.47 %	14.14 %	0.86 %	1.63 %	21.78 %	0.89 %	1.70 %	28.38 %	0.87 %	1.68 %
CC	Under	9.99 %	3.56 %	5.25 %	21.08 %	3.76 %	6.38 %	28.73 %	3.41 %	6.10 %	35.55 %	3.17 %	5.82 %
	SMOTE	10.84 %	3.86 %	5.70 %	19.38 %	3.45 %	5.86 %	25.84 %	3.07 %	5.49 %	33.37 %	2.97 %	5.46 %
NEM	Under	9.45 %	0.17 %	0.33 %	19.71 %	0.17 %	0.34 %	30.83 %	0.18 %	0.36 %	41.74 %	0.18 %	0.36 %
	SMOTE	11.25 %	0.20 %	0.39 %	21.04 %	0.18 %	0.37 %	32.16 %	0.19 %	0.37 %	43.54 %	0.19 %	0.38 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Random Forest											
	RF	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	50.13 %	6.12 %	10.92 %	66.16 %	4.04 %	7.62 %	74.93 %	3.05 %	5.86 %	82.33 %	2.51 %	4.88 %
	SMOTE	47.51 %	5.80 %	10.34 %	62.20 %	3.80 %	7.16 %	69.54 %	2.83 %	5.44 %	76.57 %	2.34 %	4.54 %
CC	Under	25.77 %	9.19 %	13.55 %	45.22 %	8.06 %	13.68 %	60.42 %	7.18 %	12.84 %	72.47 %	6.46 %	11.86 %
	SMOTE	24.00 %	8.56 %	12.61 %	44.08 %	7.86 %	13.34 %	58.83 %	6.99 %	12.50 %	71.15 %	6.34 %	11.64 %
NEM	Under	22.10 %	0.39 %	0.76 %	69.37 %	0.61 %	1.21 %	87.35 %	0.51 %	1.02 %	88.68 %	0.39 %	0.78 %
	SMOTE	23.44 %	0.41 %	0.81 %	67.98 %	0.60 %	1.18 %	83.82 %	0.49 %	0.98 %	86.95 %	0.38 %	0.76 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Multilayer Perceptron											
	MLP	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	46.83 %	5.50 %	9.84 %	66.78 %	3.79 %	7.18 %	73.48 %	2.99 %	5.75 %	81.86 %	2.50 %	4.85 %
	SMOTE	46.03 %	5.25 %	9.43 %	73.30 %	3.72 %	7.08 %	70.51 %	2.87 %	5.52 %	78.39 %	2.39 %	4.65 %
CC	Under	27.29 %	6.79 %	10.88 %	51.44 %	7.01 %	12.35 %	56.55 %	6.72 %	12.01 %	69.85 %	6.23 %	11.43 %
	SMOTE	24.60 %	8.18 %	12.28 %	44.24 %	7.46 %	12.77 %	55.88 %	6.64 %	11.87 %	68.53 %	6.11 %	11.22 %
NEM	Under	20.31 %	0.37 %	0.72 %	65.31 %	0.58 %	1.15 %	82.49 %	0.48 %	0.96 %	83.22 %	0.37 %	0.73 %
	SMOTE	20.11 %	0.34 %	0.67 %	65.78 %	0.57 %	1.13 %	82.76 %	0.48 %	0.96 %	83.62 %	0.37 %	0.73 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Support Vector Machines											
	SVM	10 %			20 %			30 %			40 %		
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	41.93 %	5.12 %	9.13 %	63.80 %	3.90 %	7.35 %	77.68 %	3.16 %	6.08 %	85.17 %	2.60 %	5.05 %
	SMOTE	41.73 %	5.10 %	9.09 %	61.63 %	3.76 %	7.10 %	73.23 %	2.98 %	5.73 %	80.69 %	2.46 %	4.78 %
CC	Under	22.83 %	8.14 %	12.00 %	42.13 %	7.51 %	12.75 %	55.45 %	6.59 %	11.78 %	65.90 %	5.87 %	10.79 %
	SMOTE	23.56 %	8.40 %	12.38 %	41.15 %	7.34 %	12.45 %	53.78 %	6.39 %	11.42 %	64.21 %	5.72 %	10.51 %
NEM	Under	20.91 %	0.37 %	0.72 %	66.58 %	0.58 %	1.16 %	83.02 %	0.49 %	0.97 %	83.56 %	0.37 %	0.73 %
	SMOTE	22.37 %	0.39 %	0.77 %	66.18 %	0.58 %	1.15 %	82.76 %	0.48 %	0.96 %	83.82 %	0.37 %	0.73 %

Experimento 4: Modelo entrenado con Julio y Agosto agregados y testeado con Septiembre[4]

Decision Tree			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	2.188	2.201	1.431	2.958
		S	2.246		1.438	3.009
	CC	U	1.116	991	160	1.947
		S	1.022		134	1.879
	NEM	U	387	194	87	266
		S	293		79	288
20 %	CTO	U	3.327	3.200	2.215	4.312
		S	3.320		1.976	4.544
	CC	U	2.138	1.895	693	3.340
		S	2.014		611	3.298
	NEM	U	657	918	156	1.076
		S	535		238	1.003
30 %	CTO	U	3.874	3.746	2.918	4.702
		S	3.814		2.732	4.828
	CC	U	3.087	2.619	1.231	4.475
		S	2.894		1.149	4.364
	NEM	U	868	1.206	308	1.379
		S	752		442	1.254
40 %	CTO	U	4.268	4.136	3.346	5.058
		S	4.159		3.236	5.059
	CC	U	3.550	3.305	1.847	5.008
		S	3.369		1.753	4.921
	NEM	U	1.035	1.224	437	1.434
		S	925		600	1.282

[4]Nomenclatura de las tablas: CTO corresponde al producto Contrato. CC a Cuenta Controlada y NEM a Navegación en el Móvil; U corresponde al método de balanceo under-sampling y S al método SMOTE.

Neural Network			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	1.456	2.201	868	2.789
		S	1.301		761	2.741
	CC	U	1.136	991	188	1.939
		S	885		128	1.748
	NEM	U	133	194	22	305
		S	129		21	302
20 %	CTO	U	1.902	3.200	1.125	3.977
		S	2.103		938	4.365
	CC	U	2.173	1.895	413	3.655
		S	1.698		450	3.143
	NEM	U	292	918	229	981
		S	291		221	988
30 %	CTO	U	3.034	3.746	1.465	5.315
		S	2.890		1.476	5.160
	CC	U	3.078	2.619	918	4.779
		S	2.499		825	4.293
	NEM	U	462	1.206	439	1.229
		S	462		410	1.258
40 %	CTO	U	3.536	4.136	2.024	5.648
		S	3.508		2.115	5.529
	CC	U	3.512	3.305	1.517	5.300
		S	2.916		1.400	4.821
	NEM	U	625	1.224	617	1.232
		S	627		573	1.278

Support Vector Machines			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	382	2.201	15	2.568
		S	552		112	2.641
	CC	U	693	991	99	1.585
		S	720		110	1.601
	NEM	U	159	194	27	326
		S	173		27	340
20 %	CTO	U	1.004	3.200	13	4.191
		S	1.140		372	3.968
	CC	U	1.453	1.895	400	2.948
		S	1.315		348	2.862
	NEM	U	314	918	214	1.018
		S	323		240	1.001
30 %	CTO	U	1.586	3.746	89	5.243
		S	1.871		682	4.935
	CC	U	2.122	2.619	720	4.021
		S	1.803		634	3.788
	NEM	U	481	1.206	422	1.265
		S	490		446	1.250
40 %	CTO	U	2.415	4.136	846	5.705
		S	2.748		1.021	5.863
	CC	U	2.364	3.305	1.149	4.520
		S	2.140		1.114	4.331
	NEM	U	647	1.224	596	1.275
		S	658		619	1.263

	Modelo Propuesto - Random Forest											
	RF	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	36.74 %	4.49 %	8.00 %	55.87 %	3.41 %	6.43 %	65.05 %	2.65 %	5.09 %	71.67 %	2.19 %	4.25 %
	SMOTE	37.72 %	4.61 %	8.21 %	55.75 %	3.41 %	6.42 %	64.05 %	2.61 %	5.01 %	69.84 %	2.13 %	4.14 %
CC	Under	15.71 %	5.60 %	8.26 %	30.09 %	5.36 %	9.11 %	43.45 %	5.16 %	9.23 %	49.96 %	4.45 %	8.18 %
	SMOTE	14.38 %	5.13 %	7.56 %	28.35 %	5.05 %	8.58 %	40.73 %	4.84 %	8.65 %	47.42 %	4.23 %	7.76 %
NEM	Under	25.77 %	0.45 %	0.89 %	43.74 %	0.38 %	0.76 %	57.79 %	0.34 %	0.67 %	68.91 %	0.30 %	0.60 %
	SMOTE	19.51 %	0.34 %	0.67 %	35.62 %	0.31 %	0.62 %	50.07 %	0.29 %	0.58 %	61.58 %	0.27 %	0.54 %

	Modelo Propuesto - Multilayer Perceptron											
	MLP	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	24.45 %	2.99 %	5.32 %	31.94 %	1.95 %	3.68 %	50.95 %	2.07 %	3.99 %	59.38 %	1.81 %	3.52 %
	SMOTE	21.85 %	2.67 %	4.76 %	35.31 %	2.16 %	4.07 %	48.53 %	1.98 %	3.80 %	58.91 %	1.80 %	3.49 %
CC	Under	15.99 %	5.70 %	8.40 %	30.58 %	5.45 %	9.25 %	43.32 %	5.15 %	9.20 %	49.43 %	4.41 %	8.09 %
	SMOTE	12.46 %	4.44 %	6.55 %	23.90 %	4.26 %	7.23 %	35.17 %	4.18 %	7.47 %	41.04 %	3.66 %	6.72 %
NEM	Under	8.85 %	0.16 %	0.31 %	19.44 %	0.17 %	0.34 %	30.76 %	0.18 %	0.36 %	41.61 %	0.18 %	0.36 %
	SMOTE	8.59 %	0.15 %	0.30 %	19.37 %	0.17 %	0.34 %	30.76 %	0.18 %	0.36 %	41.74 %	0.18 %	0.36 %

	Modelo Propuesto - Support Vector Machines											
	SVM	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	6.41 %	0.78 %	1.40 %	16.86 %	1.03 %	1.94 %	26.63 %	1.08 %	2.08 %	40.55 %	1.24 %	2.40 %
	SMOTE	9.27 %	1.13 %	2.02 %	19.14 %	1.17 %	2.20 %	31.42 %	1.28 %	2.46 %	46.15 %	1.41 %	2.74 %
CC	Under	9.75 %	3.48 %	5.13 %	20.45 %	3.65 %	6.19 %	29.87 %	3.55 %	6.34 %	33.27 %	2.97 %	5.45 %
	SMOTE	10.13 %	3.61 %	5.33 %	18.51 %	3.30 %	5.60 %	25.38 %	3.02 %	5.39 %	30.12 %	2.68 %	4.93 %
NEM	Under	10.59 %	0.19 %	0.37 %	20.91 %	0.18 %	0.36 %	32.02 %	0.19 %	0.37 %	43.08 %	0.19 %	0.38 %
	SMOTE	11.52 %	0.20 %	0.40 %	21.50 %	0.19 %	0.37 %	32.62 %	0.19 %	0.38 %	43.81 %	0.19 %	0.38 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Random Forest											
	RF	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	49.67 %	6.07 %	10.82 %	72.41 %	4.42 %	8.34 %	78.96 %	3.22 %	6.18 %	84.94 %	2.59 %	5.03 %
	SMOTE	50.53 %	6.17 %	11.00 %	76.31 %	4.66 %	8.79 %	81.07 %	3.30 %	6.35 %	84.95 %	2.59 %	5.04 %
CC	Under	27.40 %	9.77 %	14.40 %	47.01 %	8.38 %	14.22 %	62.98 %	7.48 %	13.38 %	70.49 %	6.28 %	11.54 %
	SMOTE	26.45 %	9.43 %	13.90 %	46.42 %	8.27 %	14.05 %	61.42 %	7.30 %	13.05 %	69.26 %	6.17 %	11.34 %
NEM	Under	17.71 %	0.31 %	0.61 %	71.64 %	0.63 %	1.25 %	91.81 %	0.54 %	1.07 %	95.47 %	0.42 %	0.83 %
	SMOTE	19.17 %	0.34 %	0.66 %	66.78 %	0.59 %	1.16 %	83.49 %	0.49 %	0.97 %	85.35 %	0.37 %	0.75 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Multilayer Perceptron											
	MLP	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	46.83 %	5.72 %	10.20 %	66.78 %	4.08 %	7.69 %	89.25 %	3.63 %	6.99 %	94.84 %	2.90 %	5.62 %
	SMOTE	46.03 %	5.62 %	10.02 %	73.30 %	4.48 %	8.44 %	86.65 %	3.53 %	6.78 %	92.85 %	2.84 %	5.50 %
CC	Under	27.29 %	9.73 %	14.34 %	51.44 %	9.17 %	15.57 %	67.26 %	7.99 %	14.29 %	74.60 %	6.65 %	12.21 %
	SMOTE	24.60 %	8.77 %	12.93 %	44.24 %	7.89 %	13.39 %	60.42 %	7.18 %	12.84 %	67.85 %	6.05 %	11.11 %
NEM	Under	20.31 %	0.36 %	0.70 %	65.31 %	0.57 %	1.14 %	81.82 %	0.48 %	0.95 %	82.02 %	0.36 %	0.72 %
	SMOTE	20.11 %	0.35 %	0.69 %	65.78 %	0.58 %	1.14 %	83.75 %	0.49 %	0.97 %	85.09 %	0.37 %	0.74 %

Experimento 5: Modelo entrenado con Julio y Agosto agregados y testeado con Septiembre[5] incluyendo el valor social del cliente

Decision Tree			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	2.077	2.201	1.304	2.974
		S	2.124		1.306	3.019
	CC	U	1.148	991	354	1.785
		S	1.051		289	1.753
	NEM	U	477	194	28	344
		S	400		23	356
20 %	CTO	U	3.147	3.200	2.523	3.824
		S	3.124		2.524	3.800
	CC	U	2.180	1.895	1.225	2.850
		S	2.056		1.133	2.818
	NEM	U	805	918	229	1.085
		S	702		228	1.079
30 %	CTO	U	3.671	3.746	3.149	4.268
		S	3.610		3.141	4.215
	CC	U	3.111	2.619	2.510	3.220
		S	2.880		2.371	3.128
	NEM	U	1.041	1.206	627	1.620
		S	956		617	1.545
40 %	CTO	U	4.052	4.136	3.493	4.695
		S	3.962		3.491	4.607
	CC	U	3.886	3.305	3.844	3.347
		S	3.725		3.672	3.358
	NEM	U	1.247	1.224	1.089	1.382
		S	1.171		1.046	1.349

[5]Nomenclatura de las tablas: CTO corresponde al producto Contrato. CC a Cuenta Controlada y NEM a Navegación en el Móvil; U corresponde al método de balanceo under-sampling y S al método SMOTE.

Neural Network			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	1.408	2.201	814	2.795
		S	1.254		698	2.757
	CC	U	1.156	991	395	1.752
		S	904		266	1.629
	NEM	U	171	194	7	358
		S	174		7	361
20 %	CTO	U	1.829	3.200	1.391	3.638
		S	2.015		1.578	3.637
	CC	U	2.194	1.895	1.253	2.836
		S	1.736		948	2.683
	NEM	U	385	918	79	1.224
		S	388		80	1.226
30 %	CTO	U	2.935	3.746	2.521	4.160
		S	2.772		2.359	4.159
	CC	U	3.106	2.619	2.495	3.230
		S	2.487		2.010	3.096
	NEM	U	549	1.206	293	1.462
		S	547		294	1.459
40 %	CTO	U	3.374	4.136	2.903	4.607
		S	3.380		2.921	4.595
	CC	U	3.943	3.305	3.892	3.356
		S	3.285		3.245	3.345
	NEM	U	800	1.224	674	1.350
		S	802		676	1.350

Support Vector Machines			Estudio	Compañía	Intersección	Únicos
10 %	CTO	U	538	2.201	87	2.652
		S	622		142	2.681
	CC	U	714	991	228	1.477
		S	765		257	1.499
	NEM	U	178	194	7	365
		S	185		7	372
20 %	CTO	U	1.113	3.200	494	3.819
		S	1.245		688	3.757
	CC	U	1.489	1.895	794	2.590
		S	1.370		759	2.506
	NEM	U	396	918	87	1.227
		S	389		86	1.221
30 %	CTO	U	1.682	3.746	1.075	4.353
		S	1.929		1.416	4.259
	CC	U	2.174	2.619	1.642	3.151
		S	1.830		1.461	2.988
	NEM	U	559	1.206	294	1.471
		S	570		302	1.474
40 %	CTO	U	2.475	4.136	1.955	4.656
		S	2.774		2.372	4.538
	CC	U	2.637	3.305	2.606	3.336
		S	2.618		2.579	3.344
	NEM	U	813	1.224	684	1.353
		S	815		687	1.352

	Modelo Propuesto - Random Forest											
	RF	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	34.88 %	4.26 %	7.59 %	52.85 %	3.23 %	6.08 %	61.65 %	2.51 %	4.82 %	68.04 %	2.08 %	4.03 %
	SMOTE	35.67 %	4.36 %	7.77 %	52.46 %	3.20 %	6.04 %	60.62 %	2.47 %	4.74 %	66.53 %	2.03 %	3.94 %
CC	Under	16.16 %	5.76 %	8.49 %	30.68 %	5.47 %	9.28 %	43.79 %	5.20 %	9.30 %	54.69 %	4.87 %	8.95 %
	SMOTE	14.79 %	5.27 %	7.78 %	28.94 %	5.16 %	8.76 %	40.53 %	4.82 %	8.61 %	52.43 %	4.67 %	8.58 %
NEM	Under	31.76 %	0.56 %	1.10 %	53.60 %	0.47 %	0.93 %	69.31 %	0.41 %	0.81 %	83.02 %	0.36 %	0.73 %
	SMOTE	26.63 %	0.47 %	0.92 %	46.74 %	0.41 %	0.81 %	63.65 %	0.37 %	0.74 %	77.96 %	0.34 %	0.68 %

	Modelo Propuesto - Multilayer Perceptron											
	MLP	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	23.64 %	2.89 %	5.15 %	30.71 %	1.88 %	3.54 %	49.29 %	2.01 %	3.86 %	56.66 %	1.73 %	3.36 %
	SMOTE	21.06 %	2.57 %	4.59 %	33.84 %	2.07 %	3.90 %	46.55 %	1.90 %	3.64 %	56.76 %	1.73 %	3.36 %
CC	Under	16.27 %	5.80 %	8.55 %	30.88 %	5.50 %	9.34 %	43.72 %	5.20 %	9.29 %	55.50 %	4.95 %	9.08 %
	SMOTE	12.72 %	4.54 %	6.69 %	24.43 %	4.36 %	7.39 %	35.00 %	4.16 %	7.44 %	46.24 %	4.12 %	7.57 %
NEM	Under	11.38 %	0.20 %	0.39 %	25.63 %	0.23 %	0.45 %	36.55 %	0.21 %	0.43 %	53.26 %	0.23 %	0.47 %
	SMOTE	11.58 %	0.20 %	0.40 %	25.83 %	0.23 %	0.45 %	36.42 %	0.21 %	0.42 %	53.40 %	0.23 %	0.47 %

	Modelo Propuesto - Support Vector Machines											
	SVM	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	9.03 %	1.10 %	1.97 %	18.69 %	1.14 %	2.15 %	28.25 %	1.15 %	2.21 %	41.56 %	1.27 %	2.46 %
	SMOTE	10.45 %	1.28 %	2.27 %	20.91 %	1.28 %	2.41 %	32.39 %	1.32 %	2.54 %	46.58 %	1.42 %	2.76 %
CC	Under	10.05 %	3.58 %	5.28 %	20.96 %	3.74 %	6.34 %	30.60 %	3.64 %	6.50 %	37.11 %	3.31 %	6.07 %
	SMOTE	10.77 %	3.84 %	5.66 %	19.28 %	3.44 %	5.83 %	25.76 %	3.06 %	5.47 %	36.85 %	3.28 %	6.03 %
NEM	Under	11.85 %	0.21 %	0.41 %	26.36 %	0.23 %	0.46 %	37.22 %	0.22 %	0.43 %	54.13 %	0.24 %	0.47 %
	SMOTE	12.32 %	0.22 %	0.43 %	25.90 %	0.23 %	0.45 %	37.95 %	0.22 %	0.44 %	54.26 %	0.24 %	0.47 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Random Forest											
	RF	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	49.94 %	6.10 %	10.87 %	64.21 %	3.92 %	7.39 %	71.67 %	2.92 %	5.61 %	78.84 %	2.41 %	4.67 %
	SMOTE	50.70 %	6.19 %	11.04 %	63.81 %	3.90 %	7.35 %	70.78 %	2.88 %	5.54 %	77.36 %	2.36 %	4.59 %
CC	Under	25.12 %	8.96 %	13.21 %	40.11 %	7.15 %	12.14 %	45.32 %	5.39 %	9.63 %	47.11 %	4.20 %	7.71 %
	SMOTE	24.67 %	8.80 %	12.97 %	39.66 %	7.07 %	12.00 %	44.03 %	5.23 %	9.35 %	47.26 %	4.21 %	7.74 %
NEM	Under	22.90 %	0.40 %	0.79 %	72.24 %	0.63 %	1.26 %	107.86 %	0.63 %	1.26 %	92.01 %	0.40 %	0.80 %
	SMOTE	23.70 %	0.42 %	0.82 %	71.84 %	0.63 %	1.25 %	102.86 %	0.60 %	1.20 %	89.81 %	0.39 %	0.78 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Multilayer Perceptron											
	MLP	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	46.94 %	5.73 %	10.22 %	61.09 %	3.73 %	7.03 %	69.86 %	2.85 %	5.47 %	77.36 %	2.36 %	4.59 %
	SMOTE	46.30 %	5.66 %	10.08 %	61.07 %	3.73 %	7.03 %	69.84 %	2.84 %	5.47 %	77.16 %	2.36 %	4.57 %
CC	Under	24.66 %	8.79 %	12.96 %	39.92 %	7.12 %	12.08 %	45.46 %	5.40 %	9.66 %	47.23 %	4.21 %	7.73 %
	SMOTE	22.93 %	8.17 %	12.05 %	37.76 %	6.73 %	11.43 %	43.57 %	5.18 %	9.26 %	47.08 %	4.20 %	7.71 %
NEM	Under	23.83 %	0.42 %	0.82 %	81.49 %	0.72 %	1.42 %	97.34 %	0.57 %	1.13 %	89.88 %	0.39 %	0.79 %
	SMOTE	24.03 %	0.42 %	0.83 %	81.62 %	0.72 %	1.42 %	97.14 %	0.57 %	1.13 %	89.88 %	0.39 %	0.79 %

	Modelo Combinado (Intersección entre Propuesto y Empresa) - Support Vector Machines											
	SVM	10 %			20 %			30 %			40 %	
		Recall	Precision	F	Recall	Precision	F	Recall	Precision	F	Recall	Precision	F
CTO	Under	44.53 %	5.44 %	9.70 %	64.13 %	3.92 %	7.38 %	73.10 %	2.98 %	5.72 %	78.19 %	2.39 %	4.63 %
	SMOTE	45.02 %	5.50 %	9.80 %	63.09 %	3.85 %	7.26 %	71.52 %	2.91 %	5.60 %	76.20 %	2.33 %	4.52 %
CC	Under	20.79 %	7.41 %	10.93 %	36.45 %	6.50 %	11.03 %	44.35 %	5.27 %	9.42 %	46.95 %	4.18 %	7.68 %
	SMOTE	21.10 %	7.52 %	11.09 %	35.27 %	6.29 %	10.67 %	42.05 %	5.00 %	8.93 %	47.07 %	4.19 %	7.70 %
NEM	Under	24.30 %	0.43 %	0.84 %	81.69 %	0.72 %	1.42 %	97.94 %	0.57 %	1.14 %	90.08 %	0.40 %	0.79 %
	SMOTE	24.77 %	0.43 %	0.85 %	81.29 %	0.71 %	1.41 %	98.14 %	0.57 %	1.14 %	90.01 %	0.40 %	0.79 %