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

IDENTIFICACIÓN DE CLIENTES CON PATRONES DE CONSUMO ELECTRICO FRAUDULENTO

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

NICOLE PEREIRA BIZAMA

PROFESOR GUÍA:
PABLO MARÍN VICUÑA

MIEMBROS DE LA COMISIÓN:
ALEJANDRA PUENTE CHANDÍA
LUIS ABURTO LAFOURCADE

SANTIAGO DE CHILE
AGOSTO 2014

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: NICOLE PEREIRA BIZAMA
FECHA: 22/08/2014
PROF. GUIA: PABLO MARÍN

IDENTIFICACIÓN DE CLIENTES CON PATRONES DE CONSUMO ELECTRICO FRAUDULENTO

La industria de distribución eléctrica en Chile sufre anualmente pérdidas, solo en el año 2012 la empresa en estudio registró pérdidas por más de 6 mil millones de pesos ya sea por robo o fallas en los equipos de medición, por lo cual existe un gran interés de parte de estas en buscar soluciones para mitigar esta problemática.

El presente trabajo tiene como objetivo la creación de modelos de minería de datos que logren identificar aquellos consumidores que poseen una alta propensión al hurto de electricidad. Para esto, se utilizó la información histórica disponible de los clientes desde enero de 2012 a marzo de 2014, tales como consumo mensual, inspecciones previas, cortes de suministro, entre otras fuentes. La información fue separada en dos bases de datos de acuerdo a si un cliente posee o no algún registro de inspección durante el periodo de estudio. Esta división se debe a que un cliente inspeccionado ya posee un filtro previo de inspección y a diferencia de un cliente no inspeccionado, se tiene la certeza de si ha cometido fraude o no.

Con la data de clientes inspeccionados, se construyeron tres modelos de clasificación: regresión logística, árbol de decisión y random forest. Además, debido a que se tiene una data desbalanceada con un 2.2% de casos fraude, se realizó de forma paralela un modelo de regresión logística ponderado que obtuvo resultados similares al modelo sin ponderar concluyendo que el desbalanceo de clases no afecta el problema.

Utilizando como métrica de evaluación una curva de ganancia, el modelo de random forest obtuvo los mejores resultados capturando un 39% del fraude en el primer decil de clientes versus un 35% alcanzado por el modelo de regresión. En cuanto al tiempo de ejecución, el modelo random forest tardo más de un día en su construcción mientras que el modelo de regresión y árbol de decisión tardaron entre 2 y 3 minutos. Debido a la simpleza en la interpretación de sus resultados y a su breve tiempo de ejecución se escoge el modelo de regresión logística (sin ponderar) para generar la probabilidad de fraude de cada cliente, el cual al ser aplicado a la data de clientes no inspeccionados logra una tasa esperada de fraude de un 8.6%, cifra que supera al 2.2% capturado en la realidad y que además se traduciría en una recuperación promedio mensual de más de $MM 7 si se realizasen la cantidad de inspecciones sugeridas.

De forma complementaria, con la data de clientes no inspeccionados, se construyó un modelo de clustering cuyo objetivo es agrupar clientes con similares características e identificar casos anómalos o más alejados de su grupo. Para establecer un punto de comparación entre los resultados obtenidos, se aplica el modelo de regresión al listado de casos anómalos, obteniendo una tasa esperada de fraude de un 3.1%.

Finalmente, como lineamiento futuro se espera la incorporación de otras fuentes de información que se cree serán de gran aporte en la detección de fraude energético, tales como información demográfica más detallada de los clientes y un análisis económico más preciso que permita mejores estimaciones de los beneficios a obtener.

Dedicatoria

Mamá, de manera muy especial quiero dedicarte este trabajo y más que este trabajo el haber llegado hasta aquí, ya que sin tu cariño, preocupación y el apoyo incondicional que siempre me has dado hoy no estaría escribiendo estas palabras.

Espero poder dedicarte muchos más logros y compartir más alegrías contigo…
Te amo mucho.

Agradecimientos

Habiendo culminado mi paso por la universidad me gustaría agradecer algunas personas que fueron parte de esta inolvidable y linda experiencia.

Primero agradezco profundamente a mis padres por el amor que siempre me han entregado, demostrándome día a día que puedo contar con ustedes en las buenas y en las malas. Gracias por cuidar de mí y por inculcarme siempre que este día tenía que llegar, espero se sientan orgullosos de mi porque yo estoy muy orgullosa de los padres que tengo. También a mi hermano juan pablo, aunque la mayoría del tiempo quiero matarte ocupas un lugar especial en mi corazón, gracias por tenerme tanta paciencia.

Dar las gracias también a mis profesores Pablo Marín, Luis Aburto y Alejandra Puente por toda la ayuda que me brindaron en este trabajo, ya que siempre estuvieron disponibles para cualquier pregunta o inquietud que tuviese, aprendí mucho de ustedes y se los agradezco con toda sinceridad.

Sin duda de mi paso por la universidad rescato muchas cosas y una de las más importantes son las personas que conocí, mención especial a mi amiguita de siempre Connie, que si bien nos conocemos desde mucho antes estoy muy feliz de que hayamos pasado juntas esta etapa, te agradezco mucho el saber que puedo contar siempre contigo y la paciencia que me tienes, te quiero mucho amiga.
Niñas rock: Naty, Pasita, Gilla, Dani, Nico, Connie y Mila gracias por ser como son, por quererme y porque hicieron de estos años en la universidad los más entretenidos! espero que estemos siempre unidas, las quiero mucho a todas.
También a ustedes Dani y Consu que aunque llegaron después a mi vida ocupan un lugar importante en mi corazón por las lindas personas que son.
No puedo no mencionarlos a ustedes mis queridos amigos mechones Pelao, Huaso, Scooby, Memo, gracias por todas las risas y los buenos momentos, sin ustedes definitivamente las clases no hubiesen sido lo mismo, los quiero mucho.
A ustedes amiguitos sección 3, Fabián, Pablo y Mati, que ni se imaginan el enorme cariño que les tengo, gracias por la compañía y el cariño que siempre me han demostrado.
Finalmente, a la pandilla por los lindos recuerdos que guardo de todas nuestras aventuras pandilleras.

Gracias a los chicos Penta (DM) por haberme acogido y ayudado siempre que lo necesité. Me sentí muy a gusto con ustedes y les deseo lo mejor a todos porque son excelentes personas.

Gracias a todos quienes de alguna u otra forma formaron parte de este proceso.

Tabla de contenido

Índice de tablas

Índice de figuras

Índice de ilustraciones

1. Introducción

1.1 Antecedentes generales

La industria de energía eléctrica en Chile, específicamente el sector distribución, por años se ha visto afectado por pérdidas en su suministro. Las pérdidas se definen como la diferencia entre la energía comprada (a empresas generadoras) y la energía vendida y habitualmente se clasifican en dos tipos: pérdidas técnicas y pérdidas no técnicas o comerciales.

Las pérdidas técnicas corresponden a la energía que se pierde durante la transmisión dentro la red y la distribución como consecuencia de fallas en los equipos que transportan la electricidad desde las plantas generadoras.

Las pérdidas no técnicas se pueden clasificar en tres tipos:
a) Accidentales, las cuales tienen su origen en el mal uso u operación de los elementos y equipos de los circuitos eléctricos, tal es el caso de una conexión errónea.
b) Administrativas, energía que por algún motivo no se contabiliza: usuarios sin medidores (toma directa), ferias, etc.
c) Fraudulentas, referidas a la energía que toman algunos consumidores evitando mediante algún mecanismo pasar por los medidores de la compañía de electricidad.

Es en esta última categoría donde se enmarca el presente trabajo de título, específicamente en el proceso de detección de casos fraudulentos.

1.2 Tratamiento de las pérdidas no técnicas

Para el proceso de detección de fraude eléctrico, la normativa [1] esclarece que en el caso de detectar fraude en instalaciones con medidor, el concesionario podrá facturar en perjuicio de las pérdidas de acuerdo a los valores de consumo promedio de los últimos 12 meses registrados anteriormente, por el plazo que se comprobase se estuviese cometiendo el ilícito, o considerar el ilícito perteneciente a un trimestre o dos bimestres. Para el caso de fraude en instalaciones sin medidor, el cálculo del consumo promedio mensual lo determina la empresa eléctrica de acuerdo a la potencia y horas de uso de artefactos asociados a la instalación.

2. Descripción del proyecto y justificación

2.1 Planteamiento del problema

El problema a desarrollar forma parte de uno de los servicios entregados por la empresa Penta Analytics S.A.[1] quienes entregan servicio de consultoría en inteligencia de negocios a diversas empresas que lo requieran. Dentro de esta categoría se encuentra un servicio de detección de fraude el cual ha sido focalizado a empresas distribuidoras de electricidad, donde con datos entregados por estas compañías es posible realizar predicciones de consumidores ilícitos, con la finalidad de que el proceso de inspección se focalice en clientes que tienen una mayor probabilidad de fraude o en palabras simples de estar hurtando energía.

En este contexto, el problema consiste en identificar clientes que presumiblemente estén hurtando energía, basado en patrones obtenidos desde la data de consumo de estos y otras variables tales como cortes de suministro, zona geográfica, entre otros. Sumado a esto se espera estandarizar el servicio de detección de fraude al interior de Penta Analytics generando una metodología y un detalle de las variables a utilizar, que facilite la realización de un proyecto de similares características a futuro.

2.2 Justificación del proyecto

Las compañías distribuidoras de electricidad presentan enormes pérdidas monetarias por concepto de robo de electricidad, según un estudio realizado durante el año 2012 la empresa en estudio, habría alcanzado pérdidas por más de 6 mil millones de pesos por robo de electricidad en la región que abastece, donde 2 mil corresponden al sector residencial y comercial.

Además de las pérdidas por concepto de electricidad no facturada, se suman otros costos asociados como lo son las inspecciones y en caso de detectar consumidores que estén hurtando energía, según lo estable la normativa, la empresa queda facultada para cobrar retroactivamente por un cierto período a determinar, procedimiento que también tiene recursos involucrados.

Según fuentes de la propia empresa, la tasa de fraude detectado alcanza en promedio un 2% anual por sobre el total de inspeccionados en este periodo (aproximadamente 100 mil clientes), un numero bastante bajo dada la dimensión de las pérdidas producidas por estos ilícitos, considerando que según fuentes de la misma empresa de acuerdo al estudio realizado, durante el año 2012 solo en los sectores residencial y comercial, las pérdidas habrían ascendido a más de 2 mil millones de pesos.

Actualmente para designar aquellos clientes que serán inspeccionados, la mayoría de las compañías utilizan el “juicio de expertos” a cargo de personal con años de

[1] http://www.analytics.cl/

experiencia quienes designan el grupo de clientes a inspeccionar basándose principalmente en quiebres de consumo[2]. En paralelo a esto, también se realizan las denominadas inspecciones de "barrido" que consisten en inspeccionar masivamente cierto sector sin focalizar las inspecciones a clientes específicos.

De acuerdo a los antecedentes previos es deseable detectar de forma temprana a consumidores que estén hurtando electricidad, y el principal objetivo ligado a esto es lograr un incremento en la tasa de ilícitos descubiertos (cercana a un 2% en la actualidad) que por supuesto se traduzca en una disminución de las pérdidas para las compañías.

3. Objetivos

3.1 Objetivo general

Identificar clientes con patrones de consumo altamente propensos a estar cometiendo fraude energético.

3.2 Objetivos específicos

- Identificar atributos relevantes para la detección de fraude energético (Consumos No Registrados).

- Construir y comparar resultados de modelos predictivos.

- Construir un listado estándar de las variables, junto a una metodología, a utilizar en el modelamiento de predicción de fraude.

4. Alcances

Los alcances que pretende abordar el proyecto se detallan a continuación:

- En la línea del tipo de consumidores a abordar, se trabajará solo con clientes del tipo tarifa BT1[3], vale decir clientes conectados a baja tensión con medidor simple de energía (denominado monofásico). Además se abordará solo a los clientes de tipo residencial y comercial, ya que constituyen un 98% del total de clientes. Si bien los clientes de otros rubros y tarifas (Por ejemplo, industrial, agrícola) podrían aportar una recuperación mayor a la empresa en caso de detectar un fraude, requieren otro tipo de análisis debido a que presentan un comportamiento totalmente distinto al de los segmentos escogidos y representan un porcentaje muy pequeño del universo de clientes.

[2] Se habla de quiebres de consumo, cuando un consumidor tiene un promedio de consumo más bien parejo en el tiempo y de un mes a otro presenta una baja significativa.
[3] Opción de tarifa simple en baja tensión. Medición de energía cuya potencia conectada sea inferior a 10 kW o la demanda sea limitada a 10 kW (residencial)

- Para la fase de modelamiento, se trabajará a nivel de cliente, considerando el período de datos de dos años desde enero de 2012 hasta diciembre de 2013. La data correspondiente al primer trimestre de 2014 se utilizará para analizar económicamente el desempeño de los modelos.

- Respecto a la información extra a utilizar, CASEN y SERVEL, luego de evaluar la factibilidad de éstas, no se utilizará la base de datos CASEN 2011 debido a que esta contiene información solo de muestras y no desagregada a nivel hogar, por lo cual la información no puede ser asociada a un conjunto de hogares de una zona geográfica en específico. No se utilizará el padrón electoral obtenido de SERVEL debido a la complejidad en la extracción de datos (los cuales se encuentran disponibles en un archivo formato XML[4] de gran tamaño) los cuales no significan un real aporte al trabajo, ya que de esta fuente de información solo se podría obtener la cantidad de adultos mayores de 18 años (registrados en el servicio electoral) asociados a una dirección particular, es decir, una vaga estimación de cuantas personas efectivamente viven en aquella dirección. Ambas fuentes se abordarán en la sección de recomendaciones futuras del trabajo.

- Para la construcción de modelos basados en aprendizaje supervisado (regresión logística, árbol de decisión, random forest) se trabajará solo con información correspondiente a clientes inspeccionados previamente, debido a que no se tiene la certeza de que no existan fraudes (consumos no registrados) en la base de clientes no inspeccionados. De manera complementaria, se incorporará un modelo exclusivo para estos datos, denominado modelo de aprendizaje no supervisado, que permita la detección de casos anómalos[5] dentro de este grupo de clientes.

[4] XML: siglas en inglés de eXtensible Markup Language, es un lenguaje de marcas utilizado para almacenar datos en forma legible.
[5] Caso anómalo: Corresponde a un consumidor que en alguna variable utilizada se encuentre lejos del grupo o clúster al cual pertenece.

5. Análisis descriptivo

Clientes

Dentro de la región los clientes se encuentran divididos en zonas, debido a que cada una es alimentada de energía por un grupo diferente de subestaciones eléctricas[6]. La distribución de los clientes por zona se aprecia en la siguiente tabla:

Zona	N° clientes	% de total
1	110987	21.7%
2	97419	19.0%
3	93094	18.2%
4	86961	17.0%
5	79736	15.6%
6	43617	8.5%
Total	**511814**	**100.0%**

Tabla 1: Distribución clientes por zona

Además de esta clasificación por zona los clientes pueden pertenecer a distintos rubros, siendo predominante la presencia de clientes residenciales seguido de comerciales según muestra la siguiente gráfica:

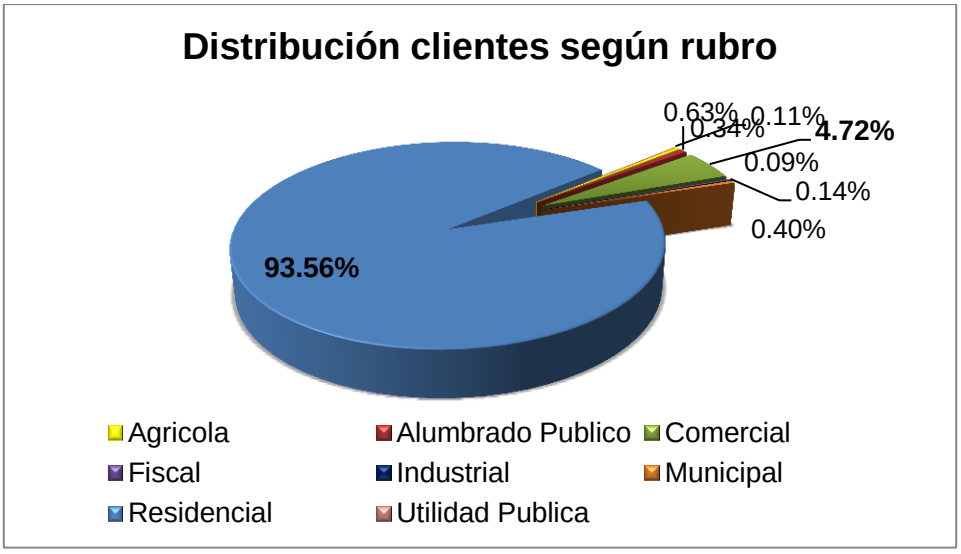

Figura 1: Gráfico de distribución clientes por rubro

Inspecciones

Del total de clientes abastecidos por la empresa, un porcentaje de éstos se inspecciona mensualmente para ver el estado de su suministro, entre los meses de enero 2012 y enero 2014 se han realizado en total 270 mil inspecciones, abarcando

[6] Una subestación eléctrica es una instalación destinada a modificar y establecer los niveles de tensión de una infraestructura eléctrica, para facilitar el transporte y distribución de la energía eléctrica.

aproximadamente 230 mil clientes, pudiendo así haber inspeccionado algún cliente en más de una ocasión. En la figura 2, se aprecia que del total de inspecciones realizadas se obtienen una tasa de 1.6% de CNR en el período estudiado.

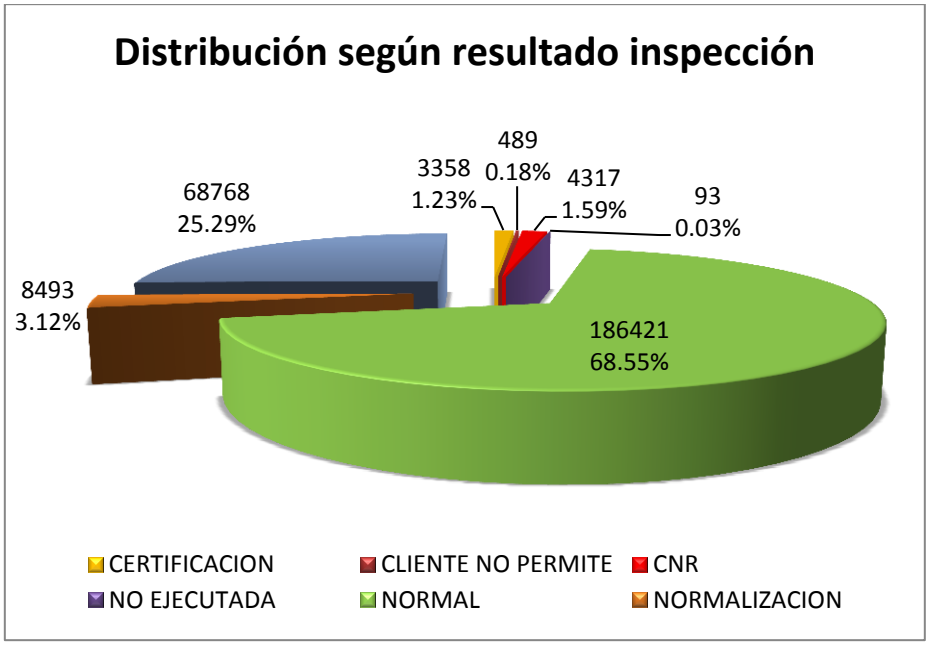

Figura 2: Gráfico de distribución de resultados inspección

El detalle del significado de cada resultado se adjunta en la siguiente tabla:

Código	Resultado	Definición
20	NORMAL	Suministro sin fraude ni falla.
30	OBSERVACION	No se lleva a cabo la inspección por diversas causas entre ellas: “terreno dificulta realización de inspección”, “casa no ubicada”, “casa cerrada”, incluso “cliente no permite”. Se deja un registro del motivo.
40	NO EJECUTADA	No se lleva a cabo inspección, no se estable motivo como registro.
50	CNR	Consumo No Registrado, intervención del medidor, dolo, fraude.
60	NORMALIZACION	Se detecta irregularidad en suministro, cuya causa de origen no implica fraude o dolo, es decir, medidor no registra los consumos de manera correcta por causas ajenas al cliente.
70	CERTIFICACION	Similar a Normalización pero no es evidente si hubo intervención o no, es decir se debe llevar a cabo otro proceso que indique si la falla es intencional o no.
80	CLIENTE NO PERMITE	Cliente no permite ingreso de fiscalizadores.

Tabla 2: Descripción resultados inspecciones

Fraudes

Todo cliente que haya sido inspeccionado y como resultado se encontrase alguna manipulación y/o alteración en el medidor, es considerado un Consumo No Registrado, en adelante CNR.

Ubicación de los CNR

En el mapa de la figura 3 se esclarece donde se encuentran ubicados los fraudes en la zona de estudio. Se observan algunas zonas de mayor concentración tales como la zona 2, 4 y 5. Lo anterior puede dar luces de que la zona geográfica a la cual el cliente pertenece puede explicar en parte (sumado a otras variables) el comportamiento fraudulento de un cliente.

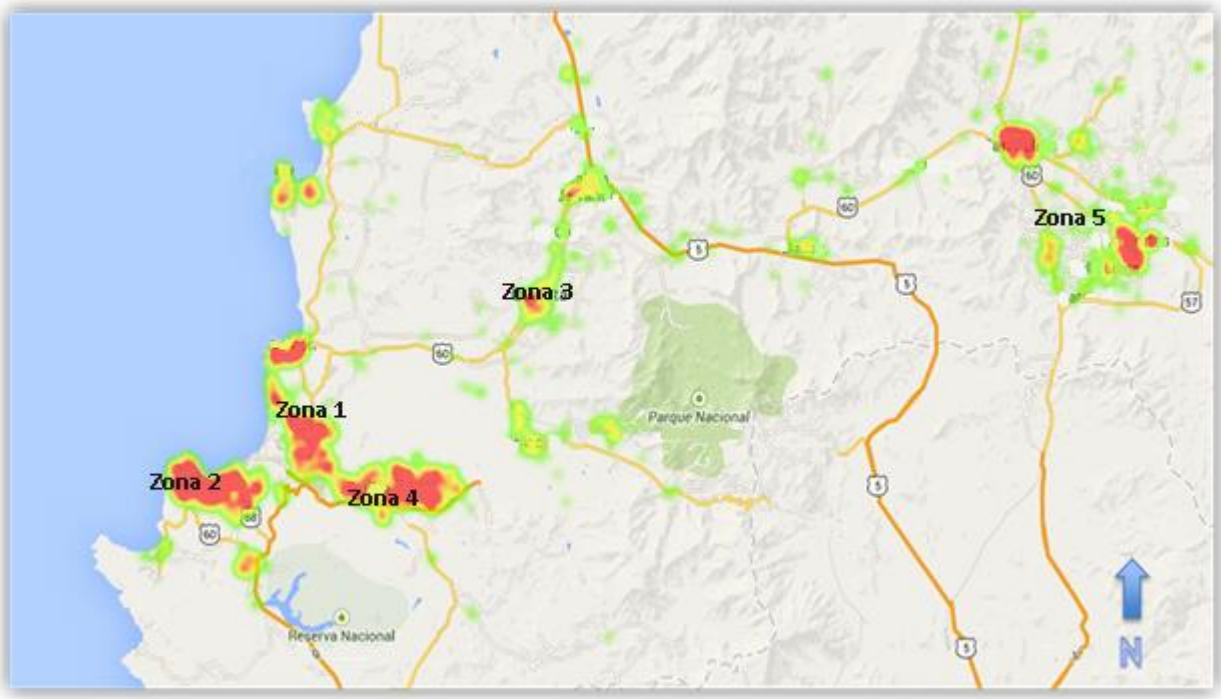

Figura 3: HeatMap de fraudes en la región de estudio

Para evaluar si este efecto de las denominadas “zonas rojas” no se debe a una mayor cantidad de inspecciones, se realiza un análisis de éstas versus los fraudes encontrados en cada zona.

De acuerdo al gráfico de la figura 4, se aprecia que el sector más inspeccionado en la actualidad corresponde a la zona 2, sin embargo la localidad con la mayor tasa de fraude detectado corresponde a zona 5 con un 3% de CNR identificados en el período considerado (Enero 2012-Enero 2014).

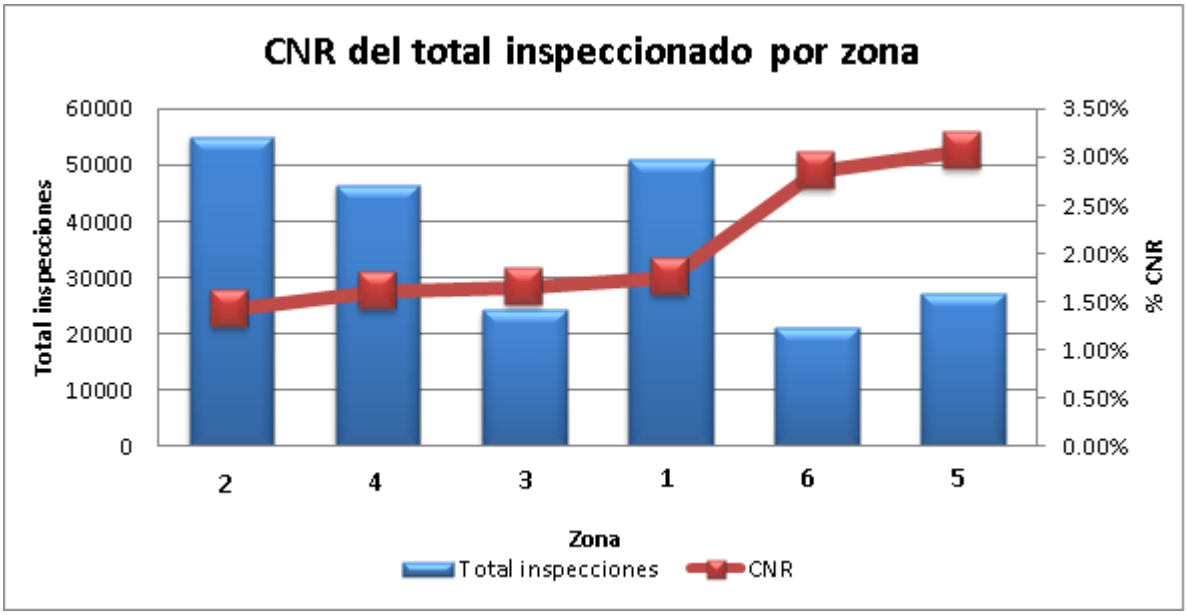

Figura 4: CNR detectados por cantidad de inspecciones por zona

El que una zona posea una mayor concentración de fraude respecto a otras, puede explicarse en que existen diferencias en el proceso de inspección de cada zona. En las zonas 1,2 y 4, periódicamente se realizan las denominadas inspecciones de “barrido” donde se selecciona algún sector de la zona y se inspecciona de forma masiva. En las zonas 3,5 y 6, se realizan inspecciones más detalladas basadas en sospechas de fraudes previas de los clientes, lo cual implicaría, según las cifras encontradas, inspecciones de mayor efectividad.

Evolución histórica inspecciones y CNR

La tasa de CNR varía según el período de estudio a considerar, esto debido a que se realizan inspecciones de forma mensual, se observa en la figura 5 que el total de inspecciones realizadas ha disminuido con el tiempo, sin embargo la tasa de CNR mantiene su fluctuación entre un 1% y un 2%. Entre enero de 2013 y enero de 2014, el número promedio de inspecciones mensual ronda entorno a las diez mil, de las cuales se logra identificar fraude (cliente identificado como CNR) en promedio para 150 casos, tal como se muestra en la figura 5.

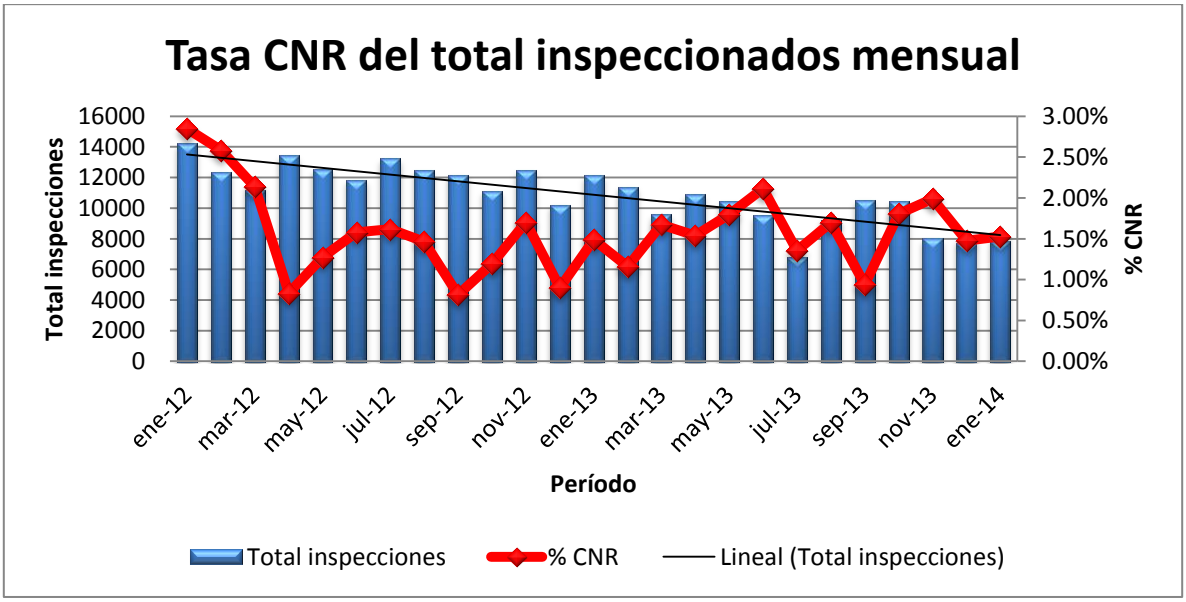

Figura 5: Tasa CNR mensual respecto a total de inspeccionados mensual

La principal hipótesis tras la disminución de las inspecciones es el alto costo que estas conllevan (más de $5 mil pesos por inspección) asociado a una tasa de fraude que no logra recuperar los gastos realizados, por lo que de no ser posible incrementar la tasa de fraude obtenida, una forma de mitigar perdidas es disminuir la cantidad de inspecciones a realizar.

Variables relevantes a estudiar

Como fase inicial del trabajo, se realiza un análisis exploratorio de ciertas variables que a criterio del alumno en colaboración con la mirada experta del negocio de personal de la empresa eléctrica, podrían estar correlacionadas con el comportamiento de clientes identificados como CNR.

- ***Consumo***

La variable consumo es la principal fuente de estudio en la que la empresa basa las inspecciones en la actualidad, es habitual que en los consumidores existan denominados “patrones de consumo” ya que un hogar por ejemplo, no debiera presentar variaciones importantes respecto a sus vecinos o entre estaciones del año.

La figura 6 muestra como debiese ser el comportamiento “normal” de un consumidor residencial en las cuatro estaciones del año, si bien su promedio de consumo puede ser más alto o bajo no debiesen existir variaciones significativas entre un mes y otro, ya que esto puede ser signo de que alguna anomalía está ocurriendo.

Según el propio personal de la empresa, el consumo promedio mensual de un cliente debiese ser de aproximadamente 150KWh con fluctuaciones máximas de ±20KWh, por lo tanto cifras mayores (menores) a ésta debiesen ser tema de análisis.

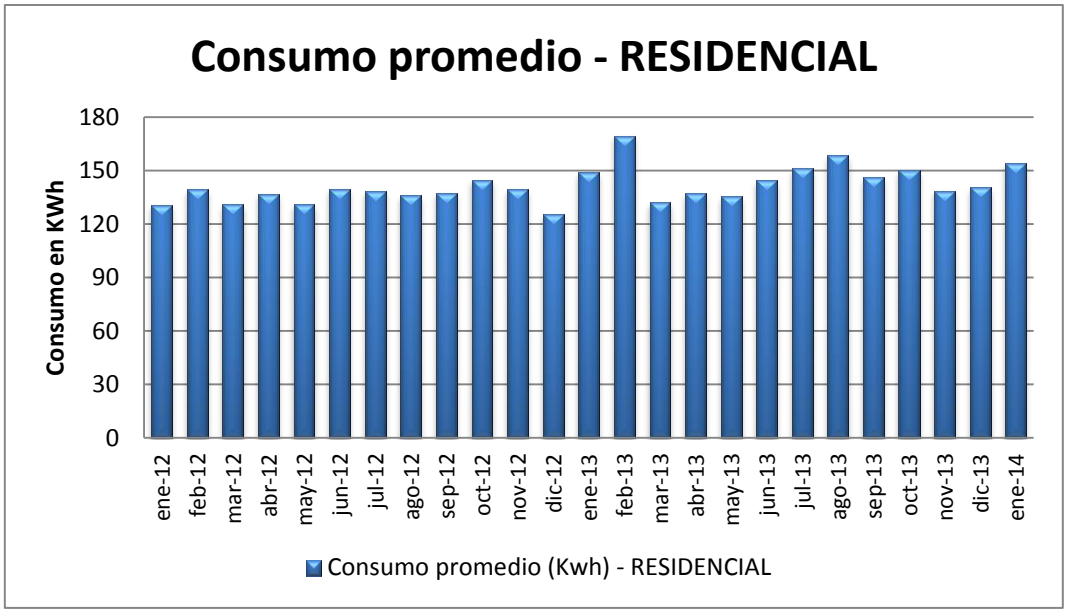

Figura 6: Consumo promedio mensual en rubro RESIDENCIAL

Además se muestra en la figura 7, el comportamiento de los rubros comercial, agrícola e industrial para ejemplificar que difieren en su comportamiento de consumo respecto a un cliente residencial.

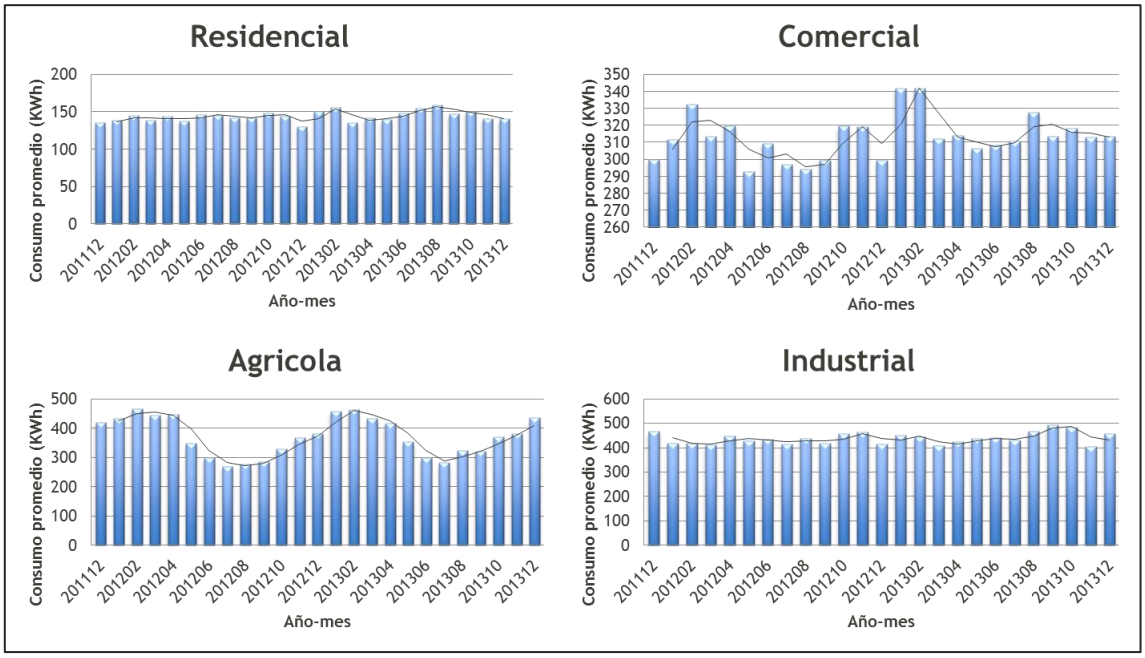

Figura 7: Consumo promedio mensual según rubro

En el caso de clientes identificados como fraudulentos una situación muy distinta se da en su patrón de consumo previo a la detección del fraude. La figura 8 muestra cuatro ejemplos de clientes residenciales identificados como CNR en el mes de diciembre de 2013. La gráfica señala su consumo mensual previo a este suceso. Se aprecia claramente un comportamiento irregular que para la empresa se traduce en una alerta de que alguna anomalía está ocurriendo.

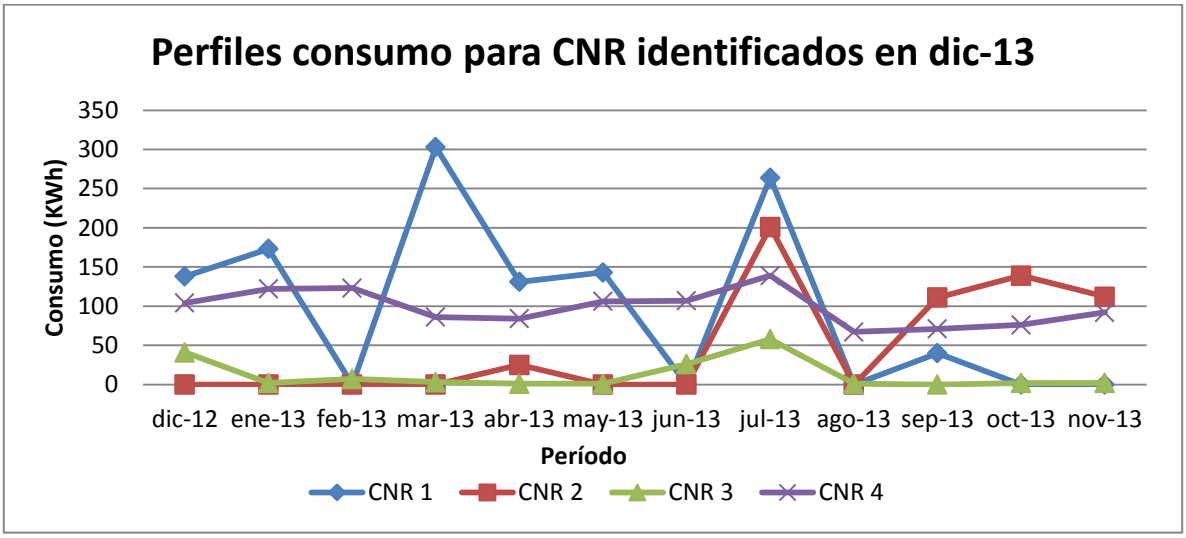

Figura 8: Perfil de consumo para casos CNR

- **Saldo adeudado**

Otra variable interesante de estudiar es el saldo adeudado por el cliente al momento de detectar un CNR. Para el análisis de esta variable se crearon distintos intervalos de saldo dentro de los clientes inspeccionados y dentro de estos grupos se identificó el porcentaje de clientes identificados con CNR. Como se observa en la figura 9 y 10, existe una correlación positiva entre tasas mayores de CNR y altos montos en saldo adeudado.

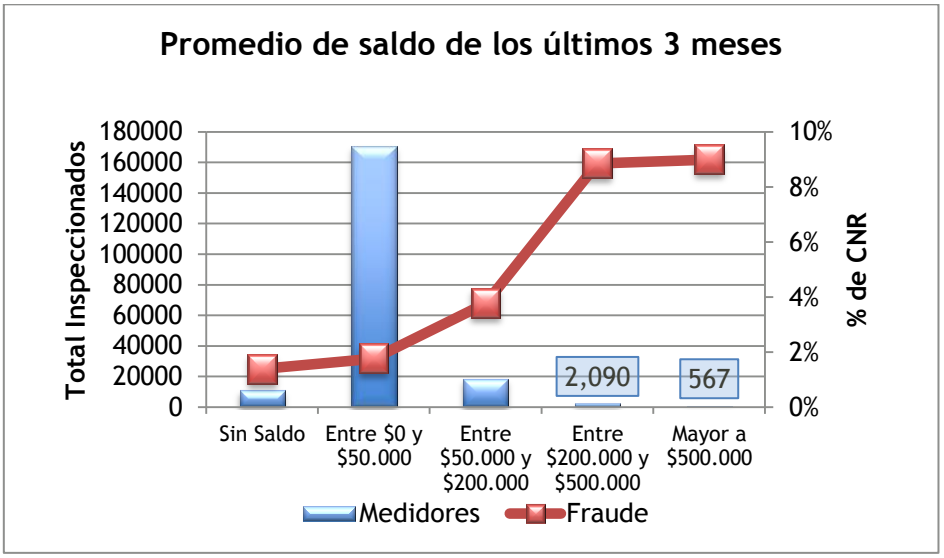

Figura 9: Porcentaje CNR respecto a saldo promedio últimos 3 meses

Figura 10: Porcentaje CNR respecto a saldo máximo últimos 12 meses

- **Rubro**

Según se puede apreciar en la figura 11, los clientes Residenciales, Agrícolas y Comerciales predominan en tasa de fraude. Cabe destacar que el rubro predominante de clientes corresponde a residenciales, sin embargo según fuentes de la empresa el descubrimiento de un fraude en un cliente comercial puede resultar en una mayor recuperación de dinero, por lo cual es importante considerarlos dentro del proyecto.

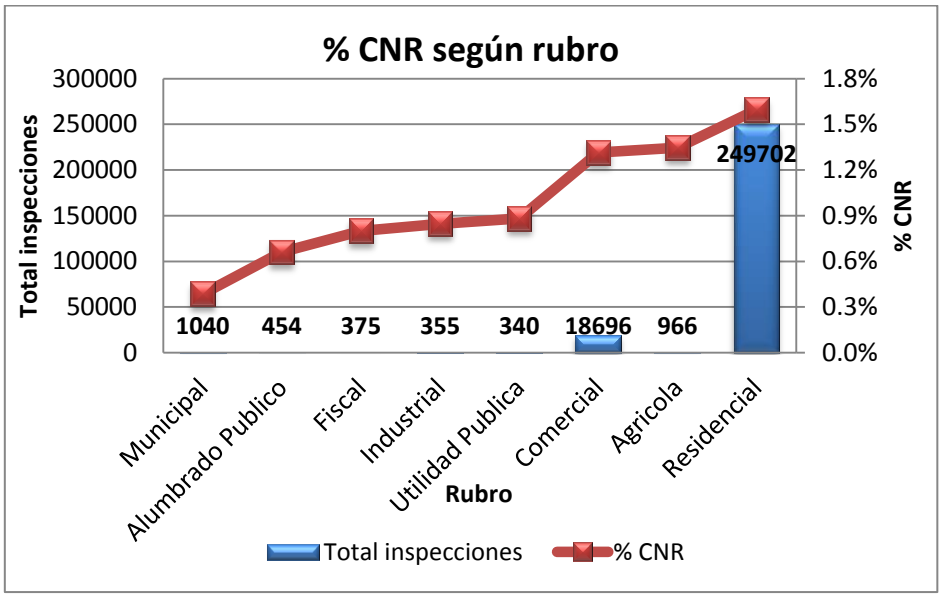

Figura 11: Porcentaje CNR según rubro

- **Cortes de suministro**

Otra información interesante de estudiar son los cortes y reposiciones de suministro, el análisis de datos evidencia que la presencia de cortes en el suministro se correlaciona positivamente con una mayor tasa de fraude. La figura 12 da muestra de cómo la cantidad de cortes en los últimos doce meses se relaciona con la tasa de fraude en el grupo de clientes inspeccionados.

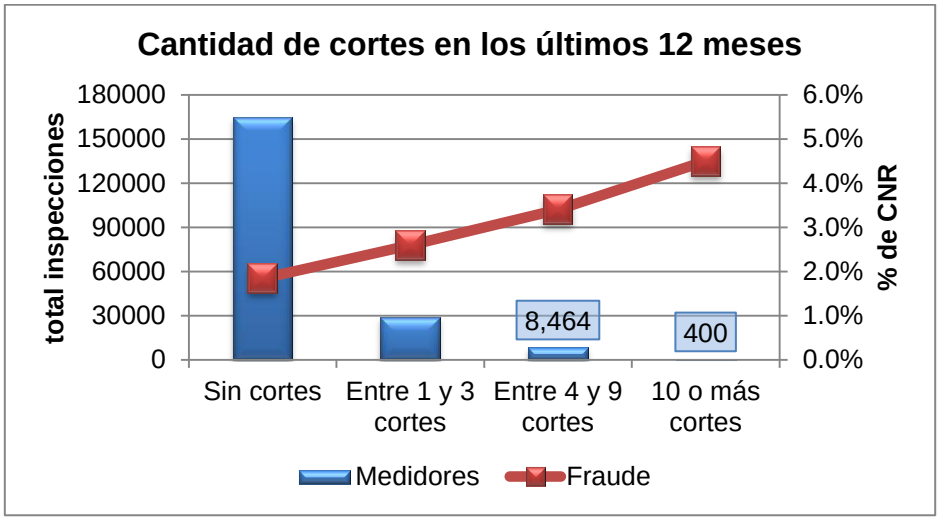

Figura 12: Porcentaje CNR respecto a cantidad de cortes de suministro

- **Claves de lectura**

Las claves de lectura corresponden a marcas registradas por el personal encargado de la lectura mensual de consumo de cada cliente en caso que existiese algún inconveniente en dicho procedimiento. A modo de ejemplo, en una casa deshabitada un inspector no podrá tomar registro del consumo por lo cual debe informar a la compañía de esta situación.

A modo de simplificar el uso de esta información y debido a la similitud que poseen entre si algunas claves, estas se agrupan en cuatro grandes grupos los cuales fueron definidos utilizando criterios entregados por la propia empresa:

Grupo Claves de lectura	Claves incorporadas					
sospecha fraude	1	10	48	68	NL	NP
falla medidor	4	12	50	53		
lectura estimada	26	27	LE	NM		
consumo atípico	13	30	42	45	60	

Tabla 3: Agrupación claves de lectura

El detalle del significado de las claves de lectura se adjunta a continuación:

Clave Observación	Observación
01	No permite leer
04	Medidor no registra consumos
10	Sospecha de Fraude
12	Medidor Desprendido
13	Valida lectura post Entrega
26	Lectura ingresada por analista.
27	Lectura estimada por analista.
30	Lectura Verificada
42	Sobre Consumo
45	Consumo Atípico
48	Repite Error Lectura Anterior
50	Cerrado por cambio medidor
53	Medidor Cifras Variadas
60	Verificado inspección de Contratista, No cerrar
68	Sin acceso al Medidor
LE	Lectura Estimada
NL	Cliente no leído
NP	Cliente no permite leer
NM	No medido

Tabla 4: Descripción claves de lectura

Síntesis

A partir de los análisis descritos se establecen primeras hipótesis que serán abordadas en el trascurso del trabajo:

- La variable zona posee un grado de dependencia con la existencia de fraude.
- Un cliente con cortes de suministro es más propenso al hurto de energía
- El saldo adeudado por un cliente está correlacionado positivamente con el hurto de energía, es decir, a mayor saldo mayor probabilidad de fraude.
- Un cliente normal, en promedio debiese mantener un consumo parejo durante todo el año y, según explica personal de la empresa, en caso de presentar alzas/bajas estas no debiesen superar los ±20KWh. Un cliente fraudulento posee un perfil de consumo variable en el tiempo.

6. Marco conceptual

El desarrollo del presente proyecto tiene como principal objetivo identificar o predecir clientes que estén hurtando energía. Para identificar a dichos clientes se requiere la utilización de técnicas de minería de datos que permitan extraer patrones a partir de la data disponible, para lograr dicho objetivo se utilizará como base la conocida metodología KDD (Knowledge Discovery in Databases), que como bien su nombre lo dice, tiene como objetivo extraer conocimiento desde bases de datos.

6.1 Metodología KDD

El objetivo principal de la metodología KDD es descubrir patrones desconocidos a priori dentro de un conjunto de datos por lo cual, se adecua bien al problema a resolver.

El proceso KDD consta de una serie de etapas consecutivas e iterativas, las cuales se describen a continuación:

- **Identificación del problema/objetivo de estudio**
  En esta etapa se debe definir cuál es el problema a resolver para de esta forma clarificar los objetivos.

- **Selección e integración**
  Se escoge el conjunto de datos desde donde se extraerá el conocimiento, dado que estos pueden venir desde distintas fuentes y formatos es importante la creación e integración del conjunto de datos que se utilizarán como base de estudio en el proceso.

- **Pre-procesamiento**
  Esta etapa corresponde a la limpieza y tratamiento de los datos, donde se produce el tratamiento de valores sin información (datos incompletos en el caso de que no se pueda extraer su valor original), ausentes (*missing*) y los valores fuera de rango (*outliers*).

- **Transformación**
  El objetivo principal de esta fase es reducir o agrupar los datos en las características de interés para el problema y en el formato apropiado para la fase posterior (minería de datos). Los datos son transformados de tal manera de generar nuevas variables o atributos que describan una característica determinada.
  Es en esta etapa donde se deben normalizar, discretizar o generar nuevas variables a partir de otras ya existentes.

  Ejemplos:
  1. A partir de la variable continua *consumo mensual (KWh)* se genera una variable categórica con *n* intervalos de *consumo mensual*.
  2. La principal variable de estudio FRAUDE se codifica en una variable dicotómica, donde 1 corresponde a presencia de fraude y 0 ausencia de este.

- **Minería de datos**
  En esta etapa se selecciona el modelo a ocupar, considerando los objetivos del proyecto planteados a priori. Es aquí donde los algoritmos "aprenden" a partir de los datos, generando patrones que se encontraban implícitos en estos.
  Para el caso particular de este proyecto se utilizarán los modelos supervisados de clasificación[7]: regresión logística, árbol de decisión y *random forest y* el modelo no supervisado de *clustering* con detección de anomalías.

- **Interpretación y evaluación**
  Se interpretan y evalúan los patrones encontrados en la fase anterior. Implica la selección de medidas de evaluación, que permitan decidir la confiabilidad y validez del modelo, así como también, el traspaso de los resultados de dichas medidas a conocimiento y acciones correctivas en el negocio, que permitan la solución al problema estudiado.

6.2 Técnicas de minería de datos

Las técnicas de minería de datos según tipo de aprendizaje se pueden dividir en dos tipos: modelos supervisados y no supervisados. En los modelos supervisados se trabaja con clases conocidas a piori, es decir, cada individuo se encuentra asociado a una clase (ej: fraude/No fraude, fumador/No fumador) con una etiqueta. El modelo escogerá un sub-conjunto de estos datos para "entrenarse" y así identificar patrones que permitan identificar a que clase debiera pertenecer un individuo. En los modelos no supervisados se desconoce a qué clase pertenece cada individuo, por lo cual se busca agrupar casos de forma tal que los objetos dentro de un grupo o clase posean un alto grado de semejanza, mientras que los pertenecientes a grupos diferentes sean poco semejantes entre sí.

6.2.1 Aprendizaje supervisado

Dentro de las técnicas supervisadas se utilizarán los modelos de regresión logística, árbol de decisión y finalmente *random forest*, el cual corresponde a una algoritmo basado en ensambles de clasificadores individuales, específicamente de árboles de decisión.

Regresión logística

La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores, por lo que vale la pena estudiar su aplicación en un modelo de detección de fraude, donde se pretende estudiar si un cliente es normal o fraudulento (dos categorías) dado un conjunto de datos con ambas clases.

[7] Ver sección 6.2.1 Marco conceptual - Aprendizaje supervisado

Según el número y la naturaleza de sus variables, existen distintos tipos de regresiones logísticas:

- Regresión Logística Univariante Simple: Solo una variable explicativa y una variable dicotómica dependiente.
- Regresión Logística Univariante Múltiple: Más de una variable explicativa (lineal, numérica, dicotómica, etc.) y solo una variable dicotómica dependiente.
- Regresión Logística Multivariante: Más de una variable dependiente de naturaleza dicotómica.

Para el presente trabajo se escoge un *modelo univariante múltiple*, ya que posee varias variables explicativas pero solo una dependiente que indicará la pertenencia del cliente a una u otra categoría (0 o 1).

Regresión logística binaria

Como se mencionó anteriormente, el modelo de regresión logística binaria posee como variable dependiente una variable cualitativa dicotómica y una o más variables explicativas independientes (covariables), que pueden ser cualitativas o cuantitativas.

Las covariables cualitativas deben ser dicotómicas, tomando valores 0 para su ausencia y 1 para su presencia (esta codificación es importante, ya que cualquier otra codificación provocaría modificaciones en la interpretación del modelo). Si la covariable cualitativa tuviera más de dos categorías, para su inclusión en el modelo, debe realizarse una transformación en varias covariables cualitativas dicotómicas ficticias (las llamadas variables dummy), de forma que una de las categorías se tomará como categoría de referencia, y así cada categoría entraría en el modelo de forma individual. En general, si la variable posee n categorías, habrá que realizar n−1 variables ficticias o *dummies*.

La ecuación de partida en los modelos de regresión logística es:

$$\mathbb{P}(y = 1|x) = \frac{exp(b_0 + \sum_{i=1}^{n} b_i x_i)}{1 + exp(b_0 + \sum_{i=1}^{n} b_i x_i)} \quad [1]$$

Donde $\mathbb{P}(\boldsymbol{y = 1}|\boldsymbol{x})$ es la probabilidad que y tome el valor 1 (presencia de fraude), dado el vector de covariables o variables independientes X y sus coeficientes b_i.

Si se divide la expresión [1] por su complemento, es decir, si se construye su odds(La probabilidad de estar cometiendo fraude entre la probabilidad de no estarlo), se obtiene la expresión:

$$\frac{\mathbb{P}(y=1|x)}{1-\mathbb{P}(y=1|x)} = exp(b_0 + \sum_{i=1}^{n} b_i x_i) \quad [2]$$

Y aplicando logaritmo natural se obtiene se obtiene una ecuación lineal:

$$\log\left(\frac{\mathbb{P}(y=1|x)}{1-\mathbb{P}(y=1|x)}\right) = b_0 + \sum_{i=1}^{n} b_i x_i \;\Leftrightarrow log\,\frac{p_i}{1+p_i} = \beta_0 + \beta_i x_i \;\; [3]$$

En la expresión [3] se ve a la izquierda de la igualdad el llamado *logit*, y se define como el logaritmo natural de la función *odds* de la variable dependiente (esto es, el logaritmo de la razón de proporciones de enfermar, de fallecer, de éxito, etc.). El término a la derecha de la igualdad es la expresión de una recta, idéntica a la del modelo general de regresión lineal:

$$y = b_0 + b_1 x_1 + \; b_2 x_2 + \cdots + b_n x_n \; [4]$$

Para la fase de clasificación, puede decidirse si un cliente pertenece a una categoría u otra si su probabilidad supera un umbral μ:

$$clasificación = \mathbb{P}(y = 1|x) > \mu \;\Rightarrow Fraude$$
$$\mathbb{P}(y = 1|x) < \mu \;\Rightarrow Normal$$

Ventajas

Además de clasificar a los clientes en las categorías establecidas, la principal ventaja del modelo de regresión logística es que cuantifica la importancia de la relación existente entre cada una de las covariables y la variable dependiente, permitiendo conocer qué variables son más relevantes para la variable categórica dependiente del presente trabajo.

Implementación

Para el desarrollo del modelo de regresión logística, se utilizarán los softwares RapidMiner y SPSS.

Árboles de decisión

Los árboles de decisión son una técnica no paramétrica basada en la generación de un modelo con estructura de árbol que permita explicar o predecir una determinada variable respuesta que puede ser tanto categórica (árboles de clasificación) como continua (árboles de regresión). Este modelo en forma de árbol se construye a partir de la división sucesiva de la muestra en subgrupos sobre el espacio de variables, de modo tal que cada grupo tenga una distinta proporción de casos positivos de la categoría en estudio.

La división de cada nodo se realiza a través de la variable independiente que sea más discriminante en cada caso y por el punto de corte más óptimo entre los posibles. Este proceso se realiza en forma recursiva hasta que se cumplan los criterios de parada establecidos.

Así una función que toma valores discretos (clase 1: Fraude, clase 0 No-fraude) puede representarse como un árbol donde:

- Cada **nodo** (no terminal) identifica alguna variable.
- Cara **rama** corresponde a un posible valor de la variable.
- Cada **nodo terminal** indica la clase en la que se clasifica una instancia.

De esta forma instancias no vistas se clasifican recorriendo el árbol: aplicándoles el test en cada nodo, por orden desde el nodo raíz hasta algún nodo hoja, que da su clasificación (ejemplo ilustración 1).

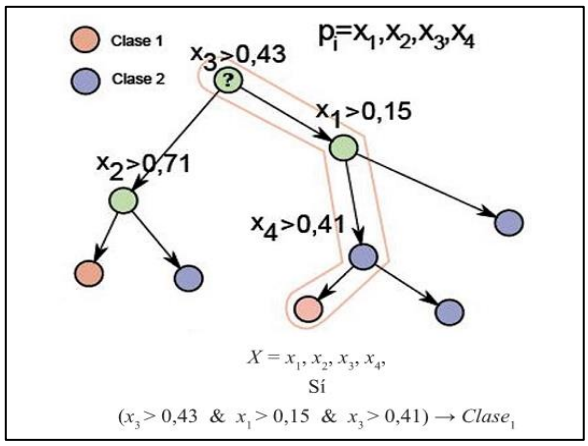

Ilustración 1: Ejemplo árbol de clasificación

Criterios de corte (Split)

- Ganancia de información (índice de entropía)
  La entropía es una medida de la variabilidad o el caos que posee la variable en cuanto a la cantidad de categorías que tiene, en este sentido una variable continua resulta altamente caótica debido a su multiplicidad de valores. Más precisamente es una medida de "información faltante".

  Dado un conjunto S con instancias que pertenecen a la clase j con probabilidad p_j. La entropía se define:

  $$Entropía(S) \ = \ -\sum p_j \ log2p_j \quad [5]$$

  Para el caso de una clasificación binaria la entropía se obtiene:

  $$Entropía = -p * log_2(p) - (1-p) * log_2(1-p) \ [6]$$

  Donde p corresponde a la proporción de casos positivos (clase=1)

Se calcula el índice de entropía para cada una de las variables: Si la variable es categórica, se obtiene sumando el índice de entropía de cada una de sus clases. En cambio, si es numérica, previamente se obtiene uno o varios puntos de corte por métodos iterativos. Se elegirá aquella variable que tenga menor índice de entropía.

- Ratio de ganancia

  Se define la *efectividad* de un atributo para subdividir un conjunto de instancias en n subconjuntos (uno por cada posible valor de X) como el valor esperado de la entropía tras efectuar la partición, calculándose como una suma ponderada de cada subconjunto S_i:

$$Efectividad(S,X) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropía(S_i) \quad [7]$$

  La ganancia de información mide cuanto disminuye en promedio la entropía al realizar cierta partición de la variable X. Se elige como nodo del árbol aquél que tenga mayor ganancia de información.

$$Ganancia(S,X) = Entropía(S) - Efectividad(E,X) \ [8]$$

- Índice de Gini

  Según su definición "mide el grado de impureza del nodo en cuestión con respecto a las clases", es decir, mide la probabilidad de no sacar dos registros con el mismo valor para la variable objetivo dentro del mismo nodo.

$$G = 1 - \sum_{i=1}^{r} p_i^2 \ [9]$$

  Donde p_i es la probabilidad de que la variable objetivo sea igual a 1 en cada uno de los grupos en los que se divide la variable explicativa. Cuanto menor es el índice de Gini mayor es la pureza del corte, por lo tanto, el primer corte propuesto será el que tenga menor valor del índice de gini.

Algunos criterios de parada

· Máxima profundidad del árbol (maximum depth): Se específica el número máximo de niveles que puede alcanzar el árbol.

· Mínimo de observaciones por nodo final (Leaf Size): Número mínimo de observaciones que tiene que tener un nodo final para que se construya la regla.

· Mínimo de observaciones para dividir un nodo (Split Size): Número mínimo de observaciones que debe tener un nodo para que se pueda dividir según la variable seleccionada.

En ocasiones es recomendable también podar (pruning) el árbol desarrollado, debido a que esto lo hace más sencillo dejando solo los nodos más relevantes y eliminando los que resultasen redundantes.

Tipos de árboles de clasificación:

Algoritmo ID3

Corresponde a una estrategia de búsqueda *voraz* (*greedy*) por un espacio de posibles arboles de clasificación, vale decir, se busca un árbol de decisión adecuado entre todos los posibles y no se considera más que una hipótesis (árbol) a lo largo del proceso. La búsqueda realizada sigue el principio de "divide y vencerás", que en este caso se centra en la división recursiva del árbol en sub-árboles en los que se busca una mayor homogeneidad en las clases existentes, de tal forma que el proceso se realiza hasta que cada partición contenga ejemplos que pertenezcan a un única clase o hasta que no haya posibilidad de realizar nuevas particiones.

Algoritmo a modo de resumen:

- Construir el árbol de arriba a abajo, preguntando: ¿Que atributo seleccionar como nodo raíz?
- Evaluar cada atributo (variable), utilizando el criterio "ganancia de información" para determinar cuan bien clasifica los ejemplos por sí solo.
- Seleccionar el mejor atributo como nodo, y se abre el árbol para cada posible valor del atributo y los ejemplos se clasifican y colocan en los nodos apropiados.
- Repetir todo el proceso usando los ejemplos asociados con el nodo en el que se esté (siempre hacia delante, buscando entre los atributos no usados en este camino)
- Parar cuando el árbol clasifica correctamente los ejemplos o cuando se han usado todos los atributos.
- Etiquetar el nodo hoja con la clase de los ejemplos.

Desventajas y limitantes

- Mediante la selección y posterior eliminación de un atributo en cada iteración, no se reconsidera la decisión en pasos sucesivos y no existe ningún tipo de retroceso en reconsideración de las opciones no elegidas, por lo cual puede que converja a una solución óptima local en vez de global.
- Matemáticamente se demuestra que favorece la elección de variables con mayor número de valores
- Complejidad crece linealmente con el número de instancias de entrenamiento y exponencialmente con el número de atributos
- Problemas por hacer crecer el árbol hasta que clasifique correctamente todos los ejemplos de entrenamiento: Sobreajuste
- Generación de grandes árboles de decisión que no representan garantía de reglas eficientes.
- Manejo solo de variables discretas (variables continuas requieren discretización previa).

Mejoras al algoritmo ID3

- Soluciones al sobreajuste:
  - Pre-poda: parar de aumentar el árbol antes de que alcance el punto en el que clasifica perfectamente los ejemplos de entrenamiento.
  - Post-poda: permitir que sobreajuste los datos, y después podarlo reemplazando subárboles por una hoja
- Otras medidas de selección de atributos (en vez de la ganancia de información): ej, Ratio de ganancia, Índice de gini.
- Manejo de datos continuos mediante la partición de estos en un conjunto discreto de intervalos, dentro de la construcción del árbol, buscando que el punto de corte para estos produzca la mayor ganancia de información.

Algoritmo C4.5

Implementando mejoras al algoritmo ID3 se crea el algoritmo C4.5 el cual escoge atributos maximizando el ratio de ganancia e incorporando post-poda de reglas: generar las reglas (1 por camino) y eliminar precondiciones (antecedentes) siempre que mejore o iguale el error de clasificación. Entre sus principales características se destacan:

- Permite trabajar con valores continuos para los atributos.
- Los árboles son menos frondosos, ya que cada hoja cubre una distribución de clases no una clase en particular.
- Utiliza el método “divide y vencerás” para generar el árbol de decisión inicial a partir de un conjunto de datos de entrenamiento.
- Se basan en la utilización del criterio de proporción de ganancia, lo cual consigue evitar que las variables con mayor número de categorías salgan beneficiadas en la selección.
- Es recursivo.

Implementación

Para el desarrollo de los modelos de árbol de decisión, se utiliza el software Rapid Miner el cual contiene entre sus funciones el modelo de clasificación ID3 y *decision tree,* correspondiente al algoritmo C4.5.

Random Forest

Random forest corresponde a una técnica multiclasificadora o también denominada técnica de ensamble. La idea es tomar *m* muestras aleatorias con reemplazo de los datos originales y luego aplicar en cada una de ellas un método predictivo para luego con algún criterio establecer un consenso de todos los resultados. Random forest genera una multitud de árboles de decisión distintos a partir de los cuales se establece la clasificación de los datos por votación, es decir, cada caso se clasifica según la categoría mayoritaria a partir de la clasificación de cada árbol.

Inconvenientes de árboles de decisión

- Cada nodo del árbol es fruto de una serie de divisiones y por tanto, las divisiones posteriores están afectadas por las divisiones precedentes.
- Es poco robusto. Pequeños cambios en los datos pueden originar árboles muy distintos.
- Para cada división puede existir un conjunto de variables con un rendimiento muy similar, información que se pierde al escoger sólo una de ellas.

Para solucionar estos inconvenientes, se trabaja con el algoritmo de random forest que entrega una solución a los problemas mencionados.

Algoritmo de random forest

1. Escoge N muestras de datos al azar (con reemplazo) creando subconjuntos del total de datos.
2. Para cada muestra genera un árbol de clasificación con la modificación de que en cada nodo en lugar de elegir la mejor división entre todas las variables predictoras (M), samplea aleatoriamente una muestra de m<M variables y escoge la mejor división entre estas.
   En el siguiente nodo, elije otras m variables al azar entre todas las variables predictoras y hace lo mismo.
3. La clasificación de los datos se realiza por votación, es decir, cada caso se clasifica según la categoría mayoritaria a partir de la clasificación de cada árbol.

Desventajas

Aplicado a los objetivos del presente trabajo, el principal inconveniente del algoritmo radica en ser un modelo denominado de “caja negra”, ya que no entrega en qué medida contribuye cada variable a la clasificación de los casos y solo es posible obtener el conjunto de árboles construidos aleatoriamente, por lo cual no es posible considerar representativo a ninguno de estos.

Implementación

Debido a la magnitud de la data a utilizar, se utilizará la extensión weka[8] del algoritmo implementada en RapidMiner, la cual muestra mejor desempeño en bases de datos de esta magnitud.

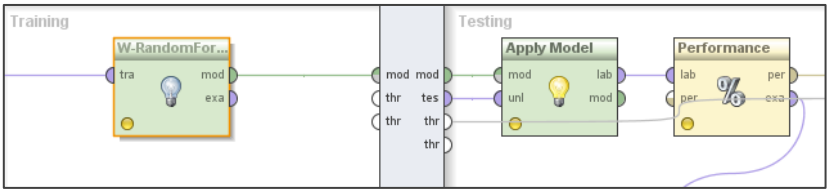

Ilustración 2: Ejecución modelo Random Forest en Rapid Miner

[8] Weka (Waikato Environment for Knowledge Analysis - Entorno para Análisis del Conocimiento de la Universidad de Waikato) es una plataforma de software para aprendizaje automático y minería de datos escrito en Java y desarrollado en la Universidad de Waikato.

6.2.2 Aprendizaje no supervisado

De manera complementaria a los modelos supervisados descritos anteriormente, se realiza un modelo de detección de anomalías, denominado de tipo no supervisado ya que utiliza como fuente de datos a clientes no inspeccionados, es decir, sin previa clasificación. El método para detectar datos anormales consiste en agrupar los casos en clústeres similares y luego buscar casos de datos dentro de cada clúster que sean diferentes en alguna medida a los otros casos dentro del mismo clúster.

Modelo de agrupamiento o clustering

La idea de los métodos de agrupamiento es clasificar un conjunto de objetos o individuos en grupos, de forma tal que los objetos dentro de un grupo posean un alto grado de semejanza, mientras que los pertenecientes a grupos diferentes sean poco semejantes entre sí. Esta clasificación por semejanza se lleva a cabo utilizando alguna *medida de similitud* definida a priori, usualmente definida como medida distancia. Algunas de las más habituales son la distancia euclidiana y la distancia de mahalanobis.

Corresponde a un método no supervisado, debido a que las clases a las que pertenecen las muestras no son conocidas.

Según como se forman los grupos, existen diferentes algoritmos de agrupamiento:

- Jerárquicos: Se generan sucesiones ordenadas (jerarquias) de clusters. Puede ser juntando pequeños clusters en uno más grande o dividiendo grandes clusters en otros más pequeños. Ejemplos de algoritmos jerárquicos: AGNES, DIANA.

- Particionales: El conjunto de casos es particionado en un número pre-especificado de grupos (K), y luego iterativamente se van reasignando las observaciones a los grupos hasta que algún criterio de parada se satisfaga. Ejemplos de algoritmos particionales: K-Means, X-means CLARA.

- Basados en densidad: Se emplea una función de densidad que divide la muestra en grupos, donde los casos tienden a agruparse en torno a la moda de una región. Ejemplo de algoritmo de densidad: DBSCAN

Para el presente proyecto se utilizará el algoritmo K-medias, por lo cual no se entrará en detalles respecto al resto de los algoritmos.

Algoritmo K-Medias

Entre los métodos particionales se encuentra el algoritmo *k-means* o *k-medias*, corresponde a uno de los algoritmos más conocidos de clustering. Particiona un conjunto de datos, D, en un determinado número de grupos, K, fijado a priori. Como salida se obtienen los grupos o *clústeres* junto a sus respectivos centroides, los cuales

son los elementos que representan a cada clúster. Éstos se calculan a partir de todos los elementos que pertenecen a un mismo clúster y definen el centro geométrico de un objeto. La función de similitud entre los elementos de este algoritmo es una métrica de distancia. Se escoge por sobre otros algoritmos, debido a que es muy útil cuando se quiere clasificar un gran número de casos.

El procedimiento de K-medias empieza con la selección de K objetos del conjunto de entrada. Estos K Objetos serán los centroides iniciales de los K-grupos. Luego se calculan las distancias de los casos a cada uno de los centroides y se asignan a aquellos grupos cuya distancia es mínima con respecto a todos los centroides, para luego actualizar las posiciones de los centroides a los valores medios de los casos en cada grupo. Estos pasos se repiten hasta que cualquier re-asignamiento de los casos haga que la variación de los grupos aumente.

Matemáticamente:
INPUT: Un conjunto de datos D y K número de clústeres a formar;
OUTPUT: L una lista de los clústeres en que caen los casos de D.

1. Seleccionar los centroides iniciales de los K clústeres: c_1, c_2,..., c_K.
2. Asignar cada observación x_i de D al clúster C(i) cuyo centroide c(i) está más cerca de x_i. Es decir, $C(i)=\text{argmin}_{1\leq k\leq K}||\mathbf{x}_i-\mathbf{c}_k||$
3. Para cada uno de los clusters se recalcula su centroide basado en los elementos que están contenidos en el clúster y minimizando la suma de cuadrados dentro del clúster. Es decir,

$$MIN\sum_{k=1}^{K}\sum_{C(i)=k}\| \mathbf{x}_i - \mathbf{c}_k \|^2$$

Ir al paso 2 hasta que se consiga convergencia.

Medida de distancia y normalización de datos
La distancia más utilizada en los métodos de agrupamiento (clustering) corresponde a la distancia euclidiana, donde la distancia entre los puntos X_i y X_j, para una dimensión de t variables está dada por:

$$D_{ij} = \sqrt{\sum_{k=1}^{t}(X_{ik} - X_{jk})^2}$$

Al utilizar las diferencias al cuadrado, las variables con grandes magnitudes ponderan mucho más que las diferencias más pequeñas.

Por ejemplo, en el problema actual se podría agrupar mediante las variables consumo promedio (KWh) y ratio consumo, donde es esperable que las diferencias en consumo

promedio sean valores más altos y por lo tanto influyan mucho más que las diferencias de la variable ratio consumo. Por este motivo es bastante útil normalizar dichos datos antes de la agrupación en clústeres.

Para normalizar los datos existen numerosas técnicas, una bastante común es la *transformación Z*, que consiste en calcular el promedio (μ) y la desviación estándar (σ) de cada variable, para luego para cada valor de la variable (x) se calcula un valor normalizado:

$$Z = \frac{x - \mu}{\sigma}$$

Validación de clustering

El procedimiento de evaluación de los resultados de un algoritmo de *clustering* es conocido como validación de *clustering.*

Debido a la dificultad de conseguir una definición precisa en la formulación de clústeres, la validación es una parte fundamental en la evaluación de un modelo de clustering. La validación de clustering se divide generalmente en dos tipos: interna y externa.
Para utilizar validación externa se debe conocer previamente la partición correcta de los datos. En general, en un contexto real a priori no se conoce la partición correcta, por lo cual este tipo de validación solo sirve en un contexto experimental para evaluar y comparar algoritmos de clustering.

La validación interna no precisa el conocimiento a priori de la partición correcta, ya que consiste en estudiar los datos y como se agrupan. Es decir, evalúa la partición en base a los datos y las distancias entre ellos. Dentro de los criterios internos, el índice Davies-Bouldin (DB) es uno de los más utilizados y con menor complejidad computacional.

Davies-Bouldin: Este índice tiene por objetivo identificar el conjunto de clústeres que son compactos y bien separados a través de una función de la proporción entre la suma de la dispersión intra-clúster y la separación inter-clúster. Para escoger el número de clústeres adecuado se toma el valor k que minimiza el índice de Davies-Bouldin porque esto significa que los clústeres son más compactos y están más separados. El índice Davies-Bouldin está definido como:

$$DB = \frac{1}{K}\sum_{i=1}^{K} máx_{j=1,\dots,K, j\neq i}\left(\frac{d(Q_i) + d(Q_j)}{d(c_i, c_j)}\right)$$

Donde K corresponde al número de clústeres, $d(Q_i)$ es la dispersión intra-cluster del cluster i y $d(c_i, c_j)$ es la separación entre los centroides del clúster i y j.

Detección de casos anómalos (outliers)
Un caso anómalo corresponde a una observación que se encuentra distanciada o alejada de un grupo de observaciones, por lo cual se consideran valores raros en comparación al normal de los datos (definiendo previamente qué es lo que se considera como “normal”).

La idea detrás de usar un método de clustering como base para la detección anomalías es, como muestra la ilustración 3, utilizar la distancia de cada caso a su correspondiente centroide como puntuación (score) de anomalía, así, mientras más se aleje un individuo de su grupo, más “raro” (*outlier*) será considerado.

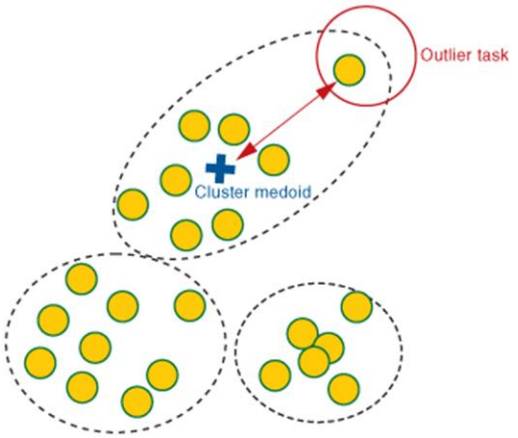

Ilustración 3: Detección de anomalías (*outliers*)

6.3 Desbalanceo de clases

Se habla de desbalanceo de clases cuando en una muestra de datos existe una clase que es representada por una gran cantidad de patrones mientras que otra es representada por muy pocos.

La mayoría de los algoritmos de aprendizaje, están diseñados para reducir el error de clasificación del clasificador construido, por tanto, si se tiene un problema donde, por ejemplo, la distribución de las clases es de 0.1%(evento positivo: clase 1) -99.9% (evento negativo: clase 0), un clasificador que identifique todos los casos como negativos, estará acertando en un 99.9%. Sin embargo, poco conocimiento se habrá extraído con dicha clasificación. Para resolver este tipo de problemas en la literatura se pueden encontrar dos tipos de enfoques: re-balanceo y ponderación.

El método de *re-balanceo* consiste en generar una nueva muestra de datos ya sea aumentando la cantidad de casos de la clase minoritaria (sobre-muestro) o extrayendo una muestra aleatoria de la clase mayoritaria (sub-muestreo) de forma tal que ambas clases tengan la misma cantidad de casos.

El método de *ponderación* introduce una nueva variable o atributo en los datos que toma el rol de "ponderador". Consiste en una variable numérica que asocia cada ejemplo con un factor de peso, permitiendo así otorgar más peso a la clase minoritaria.

Para el presente trabajo, se aborda la utilización de la segunda alternativa, la cual previo a la fase de modelamiento introduce un atributo especial en los datos que RapidMiner (ilustración 4) reconoce por su rol de ponderador y para el cual se cuenta con un operador automático que distribuye un peso dado sobre todos los ejemplos de la data, tal que la suma de los pesos por cada clase sea igual.

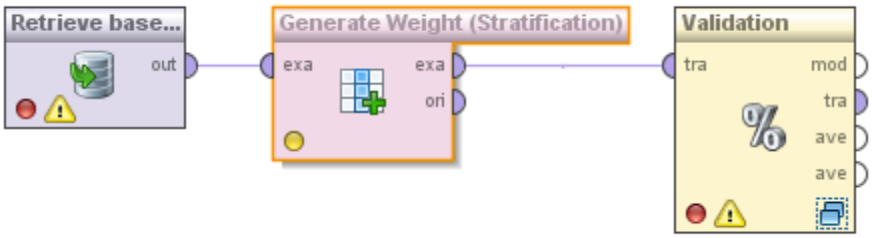

Ilustración 4: Ponderación de casos en Rapid Miner

7. Metodología

La metodología desarrollada posee como base la ya conocida knowledge-discovery in databases (KDD)[9]. Para resolver el problema se desarrollaron las fases descritas a continuación:

7.1 Identificación problema de estudio

El problema principal bajo el cual se fundamenta el proyecto, se enmarca en la necesidad de la empresa energética de aumentar su tasa de Consumos No Registrados (CNR) descubiertos. Como se mencionó en secciones previas, la tasa actual de CNR ronda en torno al 2%, lo cual es considerado como una baja efectividad del total de inspecciones realizadas mensualmente. En consecuencia, el objetivo principal del trabajo apunta a identificar un mayor número de consumos no registrados, sin considerar como variable de estudio la recuperación monetaria que éstos implican, pudiendo así descartar consumidores de impacto monetario mayor, como por ejemplo, consumidores industriales, agrícolas, entre otros.

7.2 Selección e integración

Se identifican las fuentes de información, las cuales posteriormente dan origen a las bases de datos a utilizar en la fase de modelamiento del problema.

Las principales fuentes de información se describen a continuación:

- Clientes: Cuenta con información de todos sus clientes, incluyendo el tipo de medidor, rubro al cual pertenece, zona geográfica, entre otros aspectos.

- Consumos y Saldos: Dividida en periodos mensuales, contiene información respecto al consumo (KWh), facturación y saldo adeudado en pesos (CLP) de cada cliente. Se cuenta con información desde el período enero 2012 hasta marzo 2014.

- Inspecciones: Detalla las inspecciones realizadas en el periodo de enero 2012 hasta marzo 2014, incluyendo el resultado de estas y la fecha en la cual se realizó.

- Cortes y Reposiciones: Incluye información de todos los cortes y reposiciones de suministro eléctrico, realizados entre enero de 2012 y marzo de 2014.

- Claves de lectura: En caso de que el proceso de lectura de medidor no pueda ser realizado correctamente, se toma registro de esta situación, asignándole a cada caso una denominada clave de lectura que contiene el motivo por el cual no pudo realizarse el procedimiento. Esta fuente de información contiene todas las claves registradas desde enero 2012 a marzo 2014.

[9] Ver sección 6.1 Marco conceptual – Metodología KDD

Toda la data disponible, es integrada usando una llave en común: código de medidor que identifica a un único cliente.

Integración de clientes inspeccionados y No inspeccionados

La mayoría de los modelos predictivos utilizan datos etiquetados con clases conocidas a priori. En el caso de detección de fraude se trabaja con dos clases correspondientes a:

$$1: Clientes\ identificados\ como\ CNR\ (fraudulentos)$$
$$0: Clientes\ normales$$

Esta clasificación presenta una problemática que debe ser abordada en el desarrollo del proyecto, la cual consiste en que dentro de la información disponible se cuenta con tres grupos importantes de consumidores:

1. *Consumidores que han cometido fraude*, para los cuales se tiene la certeza de que han estado hurtando energía, ya que fueron inspeccionados con resultado CNR (consumo no registrado).

2. *Consumidores inspeccionados con resultado de inspección distinto a CNR.* Dentro de este grupo de clientes, se encuentra un grupo que ha sido sujeto de inspección con resultado distinto a CNR[10]. La utilización de este grupo como la clase de clientes "normales" presenta un sesgo debido a que si bien se tiene la certeza de que estos clientes no han cometido hurto de energía, por algún motivo fueron inspeccionados, razones de esto podrían ser, por ejemplo, un patrón irregular de consumo lo que podría afectar negativamente el aprendizaje del modelo.

3. *Consumidores no inspeccionados.* También existe un grupo de clientes (aproximadamente 250 mil) que no han sido inspeccionados hasta la fecha, por lo cual no se tiene la certeza de si no han cometido fraude o si poseen un comportamiento normal exento de ilícitos.

División data de estudio: Clientes inspeccionados y no-inspeccionados

La división de la data en estos dos grandes grupos tiene su fundamento en que un cliente inspeccionado ya posee un filtro previo, denominado en adelante *"criterio de inspección".* Este criterio de inspección apunta a que actualmente la empresa inspecciona mayoritariamente a clientes que han mostrado bajas de consumo respecto a períodos anteriores. Para identificar este comportamiento se crea la variable *Ratio_consumo* que indica el consumo promedio del último semestre respecto al consumo promedio del semestre anterior. Como se observa en la figura 13, los clientes inspeccionados se concentran en los primeros deciles correspondiente a ratios de consumos cercanos a cero, vale decir, clientes que han tenido una baja en su consumo en el último semestre.

[10] Ver sección 5 Análisis descriptivo – Tabla 2: Descripción resultados inspecciones

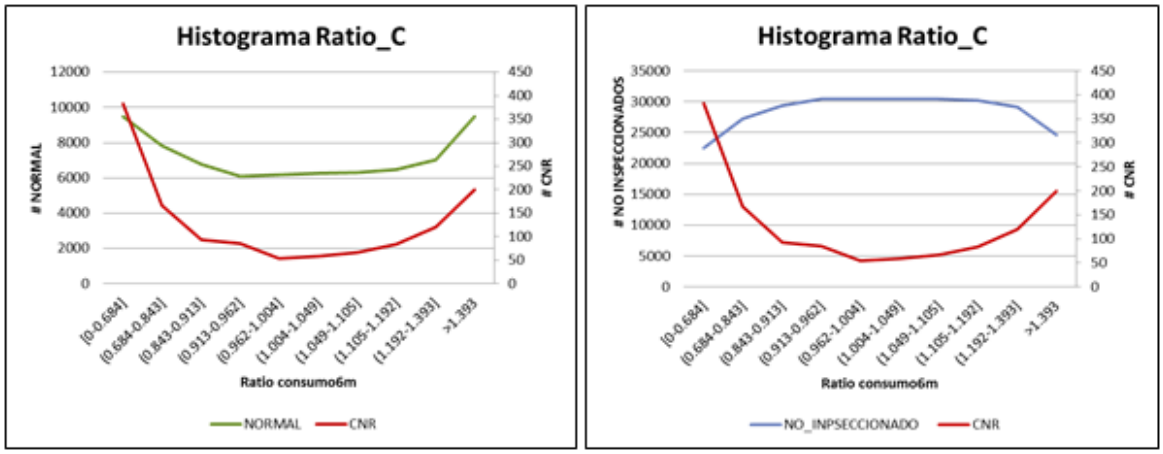

Figura 13: Histograma variable Ratio Consumo

Para hacerse cargo de la problemática asociada a trabajar con ambos grupos de datos, se incorpora la realización de dos modelos complementarios que serán comparados al finalizar el proyecto.

El primero consiste en utilizar modelos de aprendizaje supervisado (clasificación) utilizando solo la data de casos *inspeccionados* previamente, con dos clases etiquetadas, las cuales corresponden a:

Fraude=1: **Clientes inspeccionados e identificados como CNR**. Dentro de esta clase, para construir las variables a utilizar, como muestra la ilustración 5 se considerará la información registrada de todo el año previo al descubrimiento del fraude.

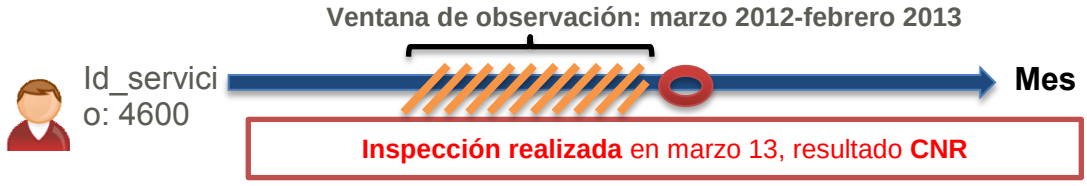

Ilustración 5: Ventana de información cliente CNR

***Fraude*=0: Clientes que fueron inspeccionados y se obtuvo como resultado inspección *normal*.** Dentro de esta clase, para construir las variables a utilizar, como muestra la ilustración 6, se considerará la información registrada de todo el año previo a la inspección.

Ilustración 6: Ventana de información cliente NORMAL

El segundo modelo, se realiza de manera complementaria al primero y cuenta con la incorporación de *clientes no inspeccionados* (utilizando información correspondiente al periodo enero 2013- diciembre 2013, según muestra la ilustración 7) construyendo un modelo no-supervisado, correspondiente a detección de anomalías en grupo similares de clientes (*clúster*) respecto a algunas variables de estudio consideradas como relevantes en la identificación de patrones del primer modelo.

Ilustración 7: Ventana de información cliente NO INSPECCIONADO

7.3 Pre-procesamiento

Aplicando la división de la data en estos dos grandes grupos, para la data inspeccionados se tiene un total de 111277 casos o inspecciones realizadas durante 2013, mientras que la data de clientes no inspeccionados cuenta con 279778 casos.

Limpieza de datos

Posterior a la división de la data se incorporan diversos filtros de limpieza en la de clientes *inspeccionados* detallados a continuación:

- Como primer paso se realiza una limpieza de datos correspondiente a eliminar aquellos consumidores que poseen un consumo total menor a 100KWh durante el primer semestre de estudio, esto con motivo de que se requiere una comparación del último semestre versus el anterior por lo cual estos casos extraños distorsionarían el análisis ya que en la mayoría de estos se tratan de viviendas deshabitadas, casas de verano, entre otras, que registran consumo cero en el primer semestre de estudio. El porcentaje de casos filtrados por esta condición corresponde a un 3.2% (3566 casos) del total de la base inspeccionados (111277). Quedando un total de 107711 casos correspondiente a un 96.8% de la base original.

- Luego corresponde el filtrado de datos correspondientes a clientes no residenciales ni comerciales. Como se mencionó en la sección de alcances, los clientes de otros rubros (agrícola, industrial, entre otros), si bien implican una recuperación monetaria mayor en caso de ser descubierto un CNR, corresponden a pocos casos y poseen un patrón de consumo distinto a clientes residenciales y/o comerciales[11]. Estos casos suman 1446 lo que significa un 1.3% de la base original.

- Como último filtro, se excluye del estudio los casos con resultado de inspección[12] distinto a CNR y NORMAL, estos casos en su mayoría corresponden a OBSERVACIÓN, lo cual significa que la inspección no fue realizada, por lo tanto no se tiene conocimiento del real resultado de la inspección. Luego de la exclusión de estos casos la base de datos a utilizar contiene 78221 registros, es decir, incorpora un 70.29% de la base original.

- Finalmente se aplican otros criterios de limpieza como eliminación de casos con facturaciones negativas y casos extremos de la variable ratio consumo. Esta última limpieza excluye 2182 casos (1.96%) incorporando en la base final un 68.33% de la base original de inspeccionados.

La base obtenida posterior a todos los filtros aplicados se compone según muestra la tabla a continuación:

Clase	Número casos	Porcentaje
Fraude = 1	1675	**2.2 %**
Fraude = 0	74364	**97.8 %**
TOTAL	**76039**	**100%**

Tabla 5: Distribución clases en base de estudio final - clientes *inspeccionados*

Cabe destacar que de los 111277 casos iniciales, 1817 casos correspondían a casos fraude, por lo cual la pérdida de información de estos casos se considera menor al aplicar los filtros aplicados mencionados anteriormente, ya que solo se eliminaron 142 casos fraude, es decir, un 7.8% del total de casos fraude iniciales.

Para la base de datos de clientes No inspeccionados, los únicos filtros aplicados fueron la eliminación de clientes que no pertenecieran al rubro residencial o comercial,

[11] Ver sección 5 Análisis descriptivo – consumos y saldos
[12] Ver sección 5 Análisis descriptivo – Tabla 2: Descripción resultados inspecciones

seleccionando un total de 279788 clientes. A partir de esta data, se extrajeron aquellos casos en que no se tuviera registro de consumo (consumo 0) en el primer semestre de estudio, excluyendo 1634 casos (0.58%), resultando en un total de 278144 clientes.

Clase	Número casos	Porcentaje
No inspeccionado	**278144**	**100%**

Tabla 6: Base de estudio clientes *No inspeccionados*

7.4 Transformación

Se generan los principales atributos a utilizar en la fase de modelamiento. Aquí los datos son transformados, agregados o normalizados, de forma tal que resulten apropiados para la fase de minería de datos.
Como primera etapa se crean distintas variables que a priori pueden resultar relevantes en la detección de un fraude, cada variable con índice *"6m"* es generada para cada semestre de estudio, vale decir desde el mes 1 al 6 y desde el mes 7 al 12.

Las variables creadas para la resolución del problema, se enlistan a continuación:

Nombre Variable	Descripción	Tipo variable	Base de origen
$\boldsymbol{Fraude}$	1: Fraude (CNR), 0: No-fraude (NORMAL)	Binaria	Inspecciones
$\boldsymbol{Rubro}$	Rubro cliente: Residencial, Comercial	Nominal (Categórica)	Cliente
$\boldsymbol{Zona}$	Zona donde habita cliente	Nominal (Categórica)	Cliente
$\boldsymbol{C_{6m}}$	Consumo promedio semestre	Continua	Consumo
$\boldsymbol{\sigma^{6m}_{consumo}}$	Desviación Estándar consumo (semestral)	Continua	Consumo
$\boldsymbol{\delta_{6m}}$	Coeficiente variación consumo (semestral)	Continua	Consumo
$\boldsymbol{C_{\max\_6m}}$	Consumo máximo semestre	Continua	Consumo
$\boldsymbol{C_{\min\_6m}}$	Consumo mínimo semestre	Continua	Consumo
$\boldsymbol{C_{\max\_dif}}$	Diferencia consumos máximos de ambos semestres	Continua	Consumo
$\boldsymbol{C_{\min\_dif}}$	Diferencia consumos máximos de ambos semestre		
$\boldsymbol{R^{6m}_{c}}$	Consumo promedio último semestre dividido Consumo primer semestre	Continua	Consumo
$\boldsymbol{R^{6m}_{stdev\_c}}$	Desviación Estándar consumo último semestre dividido consumo primer semestre	Continua	Consumo

$V\%_{consumo}^{6m}$	Variación % consumo promedio último semestre dividido consumo primer semestre	Continua	Consumo
$\rho_{C_{zona_{rubro}}}$	Coef.corrrelación consumo 12 meses respecto consumo promedio misma zona-rubro	Continua	Consumo
$Meses\_cero\_C_{6m}$	Meses sin consumo (semestral)	Continua	Consumo
$\overline{S_{6m}}$	Saldo promedio semestre	Continua	Facturación
$\overline{F_{6m}}$	Facturación promedio semestre	Continua	Facturación
$V\%_{facturación}^{6m}$	Variación % facturación promedio último semestre dividido Consumo primer semestre	Continua	Facturación
CR_{6m}	Cortes suministro 6m	Nominal	Corte y reposición
$Sospecha\_fraude_{6m}$	$\sum_{i=7}^{12} Sospecha_{fraude_{M_i}}$: Sospecha_fraude_M i=1 si en mes i,cliente fue marcado con clave de lectura relacionada a una sospecha de fraude	Nominal	Claves lectura
$Falla\_medidor_{6m}$	$\sum_{i=7}^{t+6} Falla\_medidor\_M_i$: Falla_medidor_i=1 si en mes i,cliente fue marcado con una clave de lectura relacionada a una falla de medidor	Nominal	Claves lectura
$Lectura\_estimada_{6m}$	$\sum_{i=7}^{12} Lectura\_estimada\_M_i$ Lectura_estimada_Mi=1 si en mes i,cliente fue facturado con lectura estimada y no directa del medidor	Nominal	Claves lectura
$Consumo\_atípico_{6m}$	$\sum_{i=7}^{12} Consumo\_atipico\_M_i$: Consumo_atipico_Mi=1 si en mes i cliente fue marcado con una clave de lectura relacionada a consumo atipico	Nominal	Claves lectura

Tabla 7: Variables creadas

Posteriormente, se procede a categorizar variables continuas generadas con el fin de incorporar posibles efectos de no linealidad, como por ejemplo el caso de la variable *Ratio consumo*. La discretización de las variables se lleva a cabo mediante la división del total de casos en grupos de igual tamaño.

7.5 Minería de datos

7.5.1 Selección de atributos

Se procede a realizar un ranking de importancia de las variables, a fin de conocer a priori que variables influyen en mayor medida frente a un caso clasificado como fraude. Para esto, se llevan a cabo dos análisis que entregan similares resultados permitiendo

simplificar la fase de modelamiento mediante el conocimiento de que sub-conjunto de variables posee una mayor capacidad predictiva.

Análisis n°1

Utilizando la construcción de árboles de decisión tipo CART[13], se realiza un ranking de importancia de las variables explicativas, evaluando la capacidad discriminante entre la variable objetivo (fraude) y cada variable independiente. Como resultado de este análisis, se obtiene el siguiente listado decreciente en orden de importancia[14]:

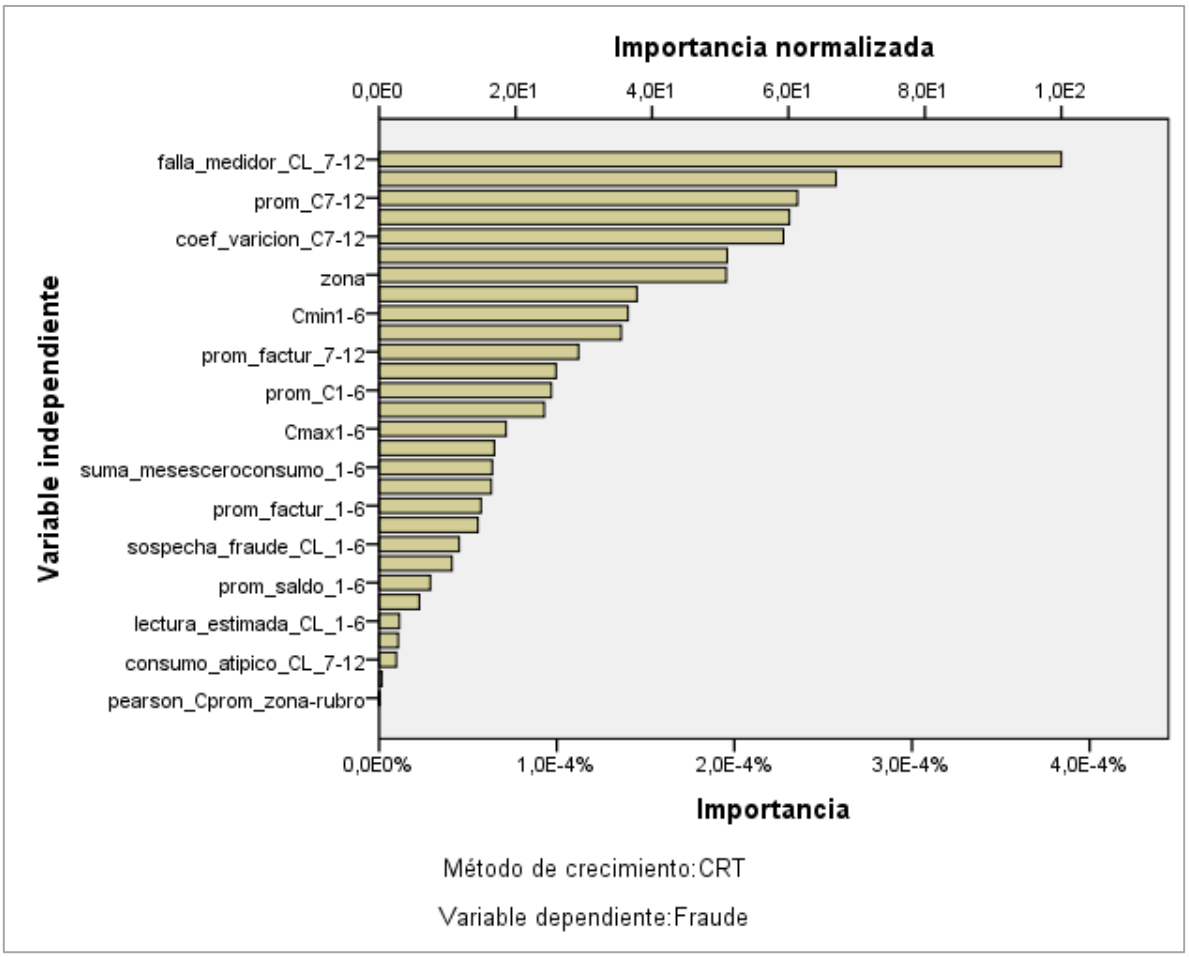

Figura 14: Gráfico de importancia de las variables independientes

[13] CART se caracteriza, fundamentalmente, por realizar divisiones basadas en el índice de gini.
[14] Ranking complete dirigirse a anexo A

Análisis n°2:
Para el análisis n°2, se construye una tabla de contingencia para cada variable generada y discretizada posteriormente. A partir de este análisis se pretende obtener que tanto poder discriminante poseen las variables independientes creadas, de modo tal de evaluar cuan valiosa es su incorporación en un modelo.

Con cada tabla de contingencia[15] se obtiene el cociente entre la probabilidad de que el intervalo i (decil n°i) sea fraude y la probabilidad de ser fraude dado el total de casos (2,2%), una vez obtenido este valor para cada intervalo se calcula el logaritmo natural de este cociente como medida de normalización. Luego de esto se obtiene el valor absoluto de cada logaritmo y se multiplica por la proporción de casos de cada intervalo (10%), la suma de esto para los diez intervalos entrega una probabilidad absoluta situada al final de la tabla (esquina inferior derecha), que indica el grado de dependencia entre la variable y la clase fraude. El objetivo de este análisis es escoger un subconjunto de variables que posean un mayor poder predictivo y de este modo comenzar integrando en orden de importancia variables al modelo predictivo.

Ratio_C6m	No fraude	fraude	Total decil	PR(fraude/total decil)	ODDS	LN(ODDS)	|LN(ODDS)|	Pr(decil)
[-inf - 0.58]	7206	398	7604	0.052	2.376	0.865	0.865	10%
[0.58 - 0.79]	7376	228	7604	0.030	1.361	0.308	0.308	10%
[0.79 - 0.88]	7451	153	7604	0.020	0.913	-0.091	0.091	10%
[0.88 - 0.94]	7480	123	7603	0.016	0.734	-0.309	0.309	10%
[0.94 - 0.99]	7508	97	7605	0.013	0.579	-0.546	0.546	10%
[0.99 - 1.04]	7515	89	7604	0.012	0.531	-0.632	0.632	10%
[1.04 - 1.11]	7505	99	7604	0.013	0.591	-0.526	0.526	10%
[1.11 - 1.21]	7478	125	7603	0.016	0.746	-0.293	0.293	10%
[1.21 - 1.43]	7434	170	7604	0.022	1.015	0.015	0.015	10%
[1.43 - inf]	7411	193	7604	0.025	1.152	0.142	0.142	10%
Total general	**74364**	**1675**	**76039**	**0.022**	**1**			37%

Tabla 8: Probabilidades condicionadas clase fraude - variable ratio consumo

De acuerdo al análisis descrito anteriormente, se obtienen doce variables relevantes[16], fijando un corte de 26% para la probabilidad absoluta. Comparando las variables para ambos métodos se observan claras similitudes, aun cuando el ordenamiento de estas no es del todo coincidente debido a los distintos criterios de decisión utilizados.

[15] Anexo B
[16] Se excluye la variable correspondiente a la *variación porcentual de consumo* debido a que resulta equivalente a la variable *ratio consumo.*

Orden importancia	Variable	Probabilidad absoluta {Ratio/Fraude}
1	**coef_varicion_C7-12**	54%
2	**Cmin_7-12**	53%
3	**coef_varicion_C1-6**	47%
4	**Ratio_C6m**	37%
5	**Cmax_dif**	36%
6	**Cmin_1-6**	36%
7	**Var%_factura**	31%
8	**Prom_C7-12**	32%
9	**Prom_ECM-mmovil_consumo**	27%
10	**Zona**	27%
11	**Prom_Factura_7-12**	26%
12	**Cmax_7-12**	26%

Tabla 9: Ranking variables discretas análisis n°2

Selección de variables predictivas para construcción modelos

Para el modelo de regresión se realizará una introducción manual de variables, utilizando como base los análisis previamente realizados. Para escoger un set optimo, se probarán distintas combinaciones de variables evaluando en cada iteración la mejora obtenida al complejizar el modelo, priorizando la elección de un modelo simple y explicativo.

Para el modelo de árbol de decisión se utilizará el mismo set de variables escogido para el modelo de regresión a fin de poder realizar comparaciones a posterior.

Para el modelo random forest se utilizará la totalidad de las variables construidas (anexo G), a fin de explorar una de las ventajas del modelo al construir diversos arboles aleatorios con distintos conjuntos de variables.

7.5.2 Modelos de clasificación (aprendizaje supervisado)

Para la etapa de modelamiento con aprendizaje supervisado, se incorporó la totalidad de los casos de la base inspeccionados (76,039 total), la cual contiene 1675 casos fraude que corresponden a un 2.2%. El objetivo de esta fase es la construcción de modelos predictivos que logren identificar con que probabilidad un cliente está cometiendo fraude.

Construcción modelo regresión logística

Se procede a utilizar la técnica de Regresión Logística incorporando como variable dependiente (clase) la variable *fraude* de tipo binaria, donde la clase fraude=1 implica la presencia de fraude y la clase fraude=0 la ausencia de este. La construcción del modelo se llevó a cabo utilizando el software RapidMiner en conjunto al software estadístico SPSS.

Para la validación del modelo, se utiliza la técnica de *Cross validation*[17] que incorpora en cada iteración un 90% de la muestra como data de entrenamiento y el 10% restante como data de prueba.

El método de introducción de variables se realizó de manera manual, mediante la construcción de varios modelos consecutivos incrementando y combinando en cada uno las variables incorporadas. Para cada uno de estos se evaluó la significancia de cada variable y la mejora en la curva de ganancia[18], de esta forma se escogió un set de 9 variables significativas que sin complejizar más el modelo mediante la incorporación de más variables, logra capturar el mayor porcentaje de fraude en el primer decil de clientes.

Cabe destacar que para mayor simplicidad en la interpretación del modelo, y para capturar efectos de no linealidad en las variables continuas incorporadas, estas se discretizaron en cuartiles.

[17] La validación cruzada o cross-validation consiste en particionar la muestra de datos en K sub-conjuntos de forma tal que en cada iteración se utilizan K-1particiones como data de entrenamiento y 1 como data de prueba, realizando en total K iteraciones para la construcción del modelo.

[18] Ver sección 7.6.1 Metodología - Evaluación de resultados

El listado completo de las variables incorporadas se detalla a continuación:

Nombre[19]	Tipo	Transformación
Cortes*	binominal	Cuenta de cortes de suministro_7-12>1 = true
Consumo Atípico (CL)*	binominal	Cuenta de claves de lectura consumo_atipico_CL_7-12>1 =true
Meses sin consumo*	binominal	meses_cero_C > 1 = true
Coeficiente variación consumo*	nominal	Rango 1 [< 0.096] , rango 2 [0.096 - 0.160] , rango 3 [0.160 - 0.392] , rango 4 [> 0.392]
Ratio Consumo	nominal	Rango 1 [< 0.841] , rango 2 [0.841 - 0.994] , rango 3 [0.994 - 1.149] , rango 4 [> 1.149]
Zona	nominal	1,2,3,4,5,6
Sospecha de fraude (CL)*	entero	Cuenta de claves lectura agrupadas en "CL sospecha fraude"
Falla de medidor (CL)*	binominal	Cuenta de claves de lectura de falla_medidor7-12>0 = true
Consumo mínimo	nominal	Rango 1 [< 32.5],rango 2 [32.5 - 88.5] , rango 3 [88.5 - 148.5] ,rango 4 [> 148.5]

Tabla 10: Variables seleccionadas para modelo de regresión logística

Se debe recordar que el modelo de regresión logística no trabaja con variables nominales de múltiples categorías (n>2), por lo cual tanto el software SPSS como RapidMiner, efectúan de manera automática la transformación de cada categoría en variables ficticias o también denominadas variables *dummy*, de forma que una de las categorías tomará el rol de categoría de referencia, generando n−1 variables ficticias o *dummies*[20].

Construcción modelo regresión logística ponderado

Con motivo de realizar una comparación entre un modelo sin balanceo y uno balanceado se construye un modelo de regresión logística ponderado, que incorpora como variable dependiente (clase) la variable *fraude*, de tipo binaria, y para este modelo en particular, se utilizará el operador *"generador de pesos",* incluido en el software RapidMiner, el cual distribuye un peso dado sobre todos los ejemplos de la data, tal que la suma de los pesos por cada clase sea igual.

El conjunto de variables utilizadas coincide con el cual se realizó la regresión logística sin ponderar (tabla 10). Bajo este objetivo, para la etapa de validación se utiliza la misma técnica *Cross validation* incorporando en cada iteración un 90% de la muestra como data de entrenamiento y un 10% como data de prueba.

[19] * Variable correspondiente al segundo semestre de estudio (6 meses previos al descubrimiento del fraude).

[20] Para ver la codificación de las variables ficticias, dirigirse a anexo D

Construcción modelo árbol de decisión

Algoritmo ID3

Se construye un árbol de decisión tipo ID3 que incorpora como variable dependiente (clase) la variable de tipo binaria *fraude*. Las variables independientes son discretizadas previamente al igual que en el modelo regresión logística. La construcción del modelo se llevó a cabo utilizando la extensión weka del algoritmo implementada en el software Rapid Miner.

Algoritmo C4.5

Para corregir el problema de sobreajuste asociado al algortimo ID3 y verificar si efectivamente se obtienen mejores resultados, se prueba el modelo C4.5, el cual además de utilizar como criterio de corte el ratio de ganancia[21] , se diferencia del modelo ID3 incorporando posterior a la construcción del árbol un proceso de poda de este y permitiendo la utilización de variables nominales y continuas, por lo cual no es necesario realizar el proceso previo de discretización de las variables ratio consumo, consumo mínimo, coeficiente de variación y sospecha de fraude (CL).

Construcción modelo Random Forest

Se construye el modelo de Random Forest que incorpora como variable dependiente (clase) la variable tipo binaria *fraude*. La construcción del modelo se llevó a cabo utilizando la extensión weka del algoritmo implementada en RapidMiner.

Con el fin de comprobar si existen mejoras al utilizar una mayor cantidad de árboles para construir el modelo, se realizaron dos iteraciones del algoritmo, incorporando 10 y 100 árboles respectivamente.

El set completo de variables utilizado en la construcción del modelo, correspondió a un grupo de 31 variables (anexo G) junto a los siguientes parámetros:

Parámetro	Modelo 1	Modelo 2
Número de arboles	10	100
Número de variables	6	6
Máxima profundidad árbol	5	5

Tabla 11: Parámetros construcción modelo Radom Forest

Cabe destacar que de acuerdo a estudios realizados previamente [18], respecto al número de árboles y variables incorporadas, el modelo no presenta mejoras significativas al incorporar más de 100 árboles, mientras que para el caso del número de variables, el mismo estudio detalla que si bien no existe un consenso absoluto

[21] Ver sección 6.2.1 Marco conceptual - árbol de decisión

respecto a este parámetro, utilizar como parámetro M/5 (M: set completo de variables) entrega una buena aproximación a lo que podría considerarse una cantidad óptima.

7.5.3 Modelo de *clustering* (aprendizaje no supervisado)

A modo de complementar los resultados obtenidos mediante los modelos de clasificación, se desarrolla un modelo de aprendizaje no supervisado, el cual tiene como objetivo identificar clientes anómalos respecto a sus pares utilizando la data correspondiente a clientes No inspeccionados.

La base "No Inspeccionados" incorpora información de todo el período 2013 para 278144 clientes que no han sido inspeccionados a la fecha, donde un 96.3% corresponde a clientes residenciales, mientras que un 3.7% a comerciales.

Rubro	Número de casos	Porcentaje casos
Residencial	267817	96.29%
Comercial	10327	3.71%
TOTAL	**278144**	**100%**

Tabla 12: Distribución clientes por rubro en base de estudio *No inspeccionados*

Construcción modelo de clustering

Para la realización del modelo de clusterización se llevan a cabo las siguientes fases:

1. Selección variables de segmentación: Esto es, definir bajo que características se agruparan o diferenciarán las instancias. Estas variables deberán estar alineadas con el objetivo de la segmentación. En el caso del presente trabajo, se utilizarán cuatro variables que caracterizarán a cada clúster construido:

    - Ratio consumo
    - Consumo mínimo segundo semestre
    - Consumo promedio segundo semestre
    - Coeficiente variación segundo semestre

Para la construcción de este modelo, se trabajarán por separado los clientes residenciales de comerciales, de manera de evitar posibles errores en la construcción de los grupos producto de las diferencias de magnitud en los valores de algunas variables como por ejemplo *consumo promedio*.

2. Selección técnica a aplicar para la segmentación: El algoritmo escogido para construir los clusters en la base de estudio corresponde al algoritmo K-medias[22]

3. Selección medida de similitud: Correspondiente al criterio para establecer la similitud o "cercanía" de un par de observaciones. Para el presente proyecto se escoge la distancia euclidiana.

[22] Ver sección 6.2.2 Marco conceptual – aprendizaje no supervisado (algoritmo K-medias)

4. Decidir número de grupos (K): Para obtener el número óptimo de clústeres, se utilizará el índice Davies-Bouldin, definido en la sección 6.2.2.

5. Interpretación del perfil de los conglomerados: Una vez realizada la clusterización es necesario corroborar que los resultados obtenidos entregan la información que se espera obtener, caracterizando a cada grupo de acuerdo a las características (variables) escogidas.

Normalización

Previo a la construcción de los clúster, se normalizó la data a modo de generar distancias equitativas y aplicar como medida de similitud la distancia euclidiana. La normalización fue aplicando a los valores de cada variable la denominada *transformación Z*:

$$Z = \frac{x - \mu}{\sigma}$$

Donde μ y σ corresponden respectivamente al promedio y desviación estándar de la variable X.

Detección de casos anómalos

Luego de que se tiene cada cluster construido y caracterizado, se obtiene la distancia euclidiana de cada caso a su respectivo centroide, para luego ser ordenados de manera decreciente en función de esta medida.

Debido a que los clúster fueron construidos utilizando cuatro variables, la distancia entre el caso $P_1 = (a_1, b_1, c_1, d_1)$ $= \left(Ratio_{consumo_1}, Consumo_{promedio_1}, Consumo_{mínimo_1}, Coeficiente_{variación_1}\right)$ y el centroide $Q = \left(a_q, b_q, c_q, d_q\right)$ $= (Ratio\_consumo_q, Consumo\_promedio_q, Consumo\_mínimo_q, Coeficiente\_variación_q)$ está dada por:

$$D_E = \sqrt{(a_1 - a_q)^2 + (b_1 - b_q)^2 + (c_1 - c_q)^2 + (d_1 - d_q)^2}$$

A modo de ejemplo se muestra gráficamente (ilustraciones 8 y 9) el procedimiento realizado dentro de un clúster. Debido a la complejidad de trabajar con cuatro dimensiones, se expone de manera gráfica solo 3 y 2 dimensiones respectivamente.

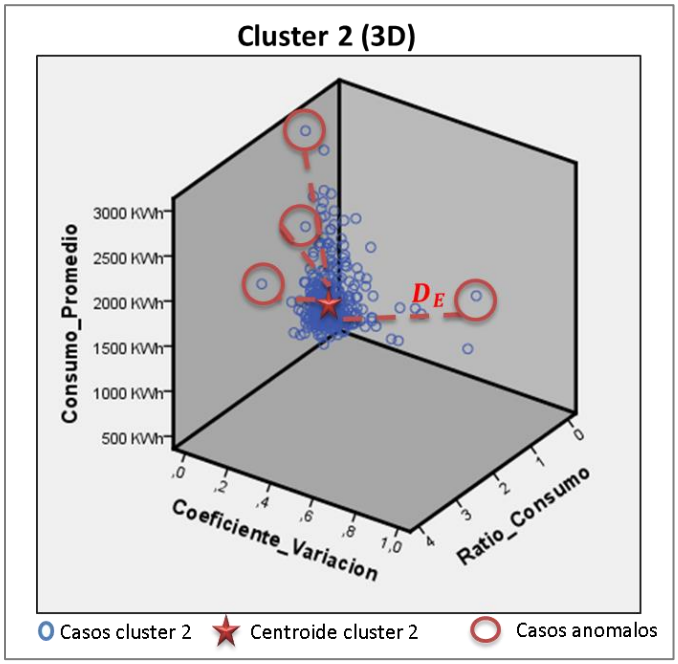

Ilustración 8: Distancia a centroide - Caso outlier(3D)

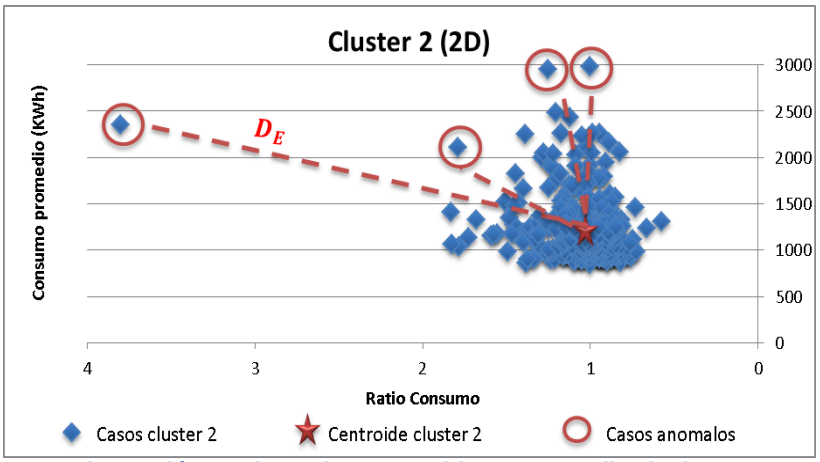

Ilustración 9: Distancia a centroide - Caso outlier (2D)

Para la definición de que casos serán considerados anómalos u *outliers* se consideran en primer lugar como *outliers*, aquellos clústeres que no resulten significativos en la distribución de casos en los grupos, es decir, aquellos clústeres que contengan menos de 100 casos. Éstos serán considerados como *"casos anómalos a nivel global"*, por ser grupos compuestos de una pequeña cantidad de casos y poseer un comportamiento muy distinto al resto.

Para la detección de *casos anómalos a nivel local (considerando solo los clústeres significativos)*, se extrae dentro de cada clúster el 1% de casos con las mayores distancias euclidianas, obteniendo un listado final de 2782 clientes.

7.6 Evaluación y comparación de resultados

7.6.1 Modelos de clasificación

Como es habitual en el desarrollo de modelos predictivos con clases previamente etiquetadas, la data a utilizar se divide en una muestra de entrenamiento y otra de prueba con motivo de evaluar qué porcentaje de los datos es clasificado de forma correcta. Recordando que se utiliza la metodología de validación cruzada, la cual divide la muestra en 10 partes iguales considerando en cada iteración un 90% de la data para entrenar el modelo y un 10% de la data para testearlo.

Para el problema actual, se debe tener en consideración un problema extra que la data posee, denominado *desbalanceo de clases*[23] el cual ha sido ampliamente estudiado en la literatura. El problema de una data desbalanceada consiste a grandes rasgos, en que una de las clases de estudio (en este caso fraude=1) posee pocos casos en comparación a la clase contraria (fraude =0). Como se mencionó en la fase de pre-procesamiento, la base a utilizar contiene tan solo un 2.2% de la clase fraude =1, versus un 97.8% de la clase fraude=0. Cuando esta probabilidad es muy baja se dificulta la creación de un modelo, ya que se hace difícil encontrar algún patrón en los datos que discrimine entre los datos positivos y negativos (fraude y no fraude). Como solución a esto se puede balancear la data otorgando más peso a la variable de estudio (fraude=1), sin embargo esto también puede conllevar a errores debido a que estamos induciendo un sesgo en el patrón natural de los datos que corresponde históricamente a una baja tasa de fraude. Por este motivo, se construirá el modelo con y sin balanceo y se compararán ambos resultados.

Dado que se cuenta con una muestra desbalanceada, se hace imposible que un modelo clasifique como fraude=1 aquel caso que tenga una probabilidad de fraude > 0.5, ya que si se extrae de forma aleatoria una muestra de *n* clientes, el porcentaje de fraude capturado sería el mismo que el relativo al tamaño de la muestra respecto al número total de datos, por lo cual exigir como criterio de corte al modelo p=0.5 no tiene sentido en el problema actual. Además, dado que la empresa posee una máxima capacidad de inspección, debe inspeccionar una cantidad fija de clientes sin importar la magnitud de la probabilidad de fraude obtenida.

Para evaluar la calidad de los modelos de clasificación construidos se utilizarán tres criterios descritos a continuación:

1. Curva de ganancia

Utilizando la probabilidad de fraude obtenida, se ordenan todos los clientes de manera decreciente para luego ser agrupados en deciles, donde el primer decil corresponderá al grupo con las mayores probabilidades de fraude obtenidas. De este modo se construye gráficamente la *curva de ganancia,* que incorpora en el eje horizontal el

[23] Ver sección 6.3 Marco conceptual - Desbalanceo de clases

porcentaje (decil) de clientes seleccionados, y en el eje vertical el porcentaje de fraude capturado al clasificar como fraude aquel porcentaje de clientes.

El mejor modelo será aquel que a menor porcentaje de clientes seleccionados, capture el mayor porcentaje de fraude. Para efectos del presente trabajo, se utilizará como medida de evaluación la ganancia que consiga el modelo en el primer decil de clientes.

2. Tasa fraude esperada

Utilizando los mismos deciles para la construcción de la curva de ganancia, se estudia la probabilidad promedio de fraude obtenida para cada grupo versus la proporción real de fraude existente. Al igual que el anterior criterio, el mejor modelo será aquel que entregue la tasa de fraude esperada más cercana a la real en el primer decil de clientes.

3. Tiempo ejecución modelo

El tiempo de ejecución del modelo corresponde al tiempo en que este tarda en completar su construcción en el software. A menor tiempo de ejecución, más eficiente se considerará un modelo.

7.6.2 Modelo de clustering

Uno de los principales inconvenientes de los modelos no supervisados es la fase de evaluación, ya que al no existir clases predefinidas resulta complejo evaluar la calidad de los modelos construidos. En particular para el caso del presente modelo, se trata de clientes que no han inspeccionados a la fecha por lo cual es imposible saber a priori que porcentaje de aquel grupo detectado como outliers corresponde efectivamente a un caso de fraude.

Para crear una métrica de comparación a los modelos supervisados, se aplicará el modelo supervisado de regresión logística a todos los casos calificados como outliers de modo de obtener la tasa promedio de fraude de este grupo. El modelo será evaluado positivamente en caso de existir un lift significativo entre la tasa real de fraude (2.2%) y la tasa esperada de fraude obtenida del listado de casos calificados como *outliers*.

8. Resultados

8.1 Modelos de clasificación

8.1.1 Modelo regresión logística

Coeficientes

Los coeficientes (betas) obtenidos de la regresión, junto a su significancia (global e individual) se adjuntan en la tabla a continuación:

Variable[24]	Tipo	Beta (fraude=1)	Significancia *p-value*
Cortes*	Binaria	0.590	.000
Consumo atípico (CL)*	Binaria	0.326	.002
Meses sin consumo*	Binaria	0.166	.044
Coeficiente variación consumo*	Binaria	0.000	.000
Coeficiente variación consumo* =rango 2 [0.096 - 0.160]	Binaria	0.101	.112
Coeficiente variación consumo* =rango 3 [0.160 - 0.392]	Binaria	0.633	.000
Coeficiente variación consumo* =rango 4 [> 0.392]	Binaria	0.675	.000
Consumo mínimo*			.000
Consumo mínimo* =rango 1 [< 32.5]	Binaria	0.535	.000
Consumo mínimo* =rango 2 [32.5 - 88.5]	Binaria	0.296	.001
Consumo mínimo* =rango 3 [88.5 - 148.5]	Binaria	-0.177	.064
Ratio consumo			.000
Ratio consumo =rango 1 [< 0.841]	Binaria	0.208	.002
Ratio consumo=rango 2 [0.841 - 0.994]	Binaria	-0.132	.099
Ratio consumo =rango 3 [0.994 - 1.149]	Binaria	-0.174	.037
zona			.000
zona=2	Binaria	0.022	.000
zona=3	Binaria	0.433	.000
zona=4	Binaria	0.314	.000
zona=5	Binaria	1.012	.000
zona=6	Binaria	0.681	.188
Sospecha de fraude(CL)*	Entero	0.121	.000
Falla de medidor(CL)*	Binaria	2.118	.000
Intercepto		-4.956	.000

Tabla 13: Coeficientes regresión logística

Cuando la variable es cualitativa con n categorías (siendo n>2), se analiza en el modelo la significancia individual de sus n−1 variables *dummies*, así como la significancia global de la variable que compara la presencia en bloque, de sus n−1 variables ficticias, frente a la ausencia de estas. Bajo este supuesto, se concluye que todas las variables resultan ser significativas en el modelo.

[24] * Variable correspondiente al periodo 2° semestre (6 meses previos al descubrimiento del fraude).

A modo de ejemplo la tabla 14 muestra la probabilidad de fraude que alcanzan distintos tipos de clientes según sus características, donde para obtener la probabilidad de que un cliente este cometiendo fraude, se reemplazan los coeficientes betas obtenidos en la fórmula de regresión:

$$\mathbb{P}(y = 1|x) = \frac{exp(b_0 + \sum_{i=1}^{n} b_i x_i)}{1 + exp(b_0 + \sum_{i=1}^{n} b_i x_i)}$$

Variable	Cliente id 173656	Cliente id 254356	Cliente id 324245
Cortes	1	0	0
Consumo atípico(CL)	0	1	1
Meses sin consumo	1	1	0
Coeficiente variación	rango 4 [> 0.392]	rango 4 [> 0.392]	rango 3 [0.160 - 0.392]
Consumo mínimo	rango 1 [< 32.5]	rango 1 [< 32.500]	rango 2 [32.500 - 88.500]
Ratio Consumo	rango 1 [< 0.841]	rango 1 [< 0.841]	rango 1 [< 0.841]
Zona	2	5	6
Sospecha fraude(CL)	0	0	0
Falla de medidor(CL)	0	0	1
Probabilidad (Fraude = 1)	**0.0595**	**0.1157**	**0.333**

Tabla 14: Ejemplificación de clientes con distintas probabilidades de fraude

Ejemplo:
Cliente id 173656: *"Cliente con al menos un corte en los últimos 6 meses, al menos un mes con consumo cero, coeficiente variación de consumo mayor a 0.392, ratio consumo menor a 0.841, perteneciente a la zona 2 y sin marcas de lectura asociadas a consumo atípico, sospecha fraude o fallas de medidor tiene un 5.95% de probabilidad de estar cometiendo fraude"*

Análisis de variables y coeficientes asociados

A modo de analizar si los resultados obtenidos con la regresión coinciden con la intuición, se analiza de manera descriptiva cada variable incorporada.

- Variable *Coeficiente variación consumo*

Los coeficientes (betas) obtenidos coinciden con la intuición y con el análisis descriptivo de esta variable, a modo de ejemplo, es de esperar que un cliente con un coeficiente de variación bajo, es decir, una desviación estándar pequeña, entregue el coeficiente beta más pequeño, ya que se espera que un cliente normal posea poca variación en su consumo mensual. De la misma manera, es de esperar que la tasa de fraude aumente cuando el coeficiente de variación aumenta, lo que puede verse claramente representado en la figura 15.

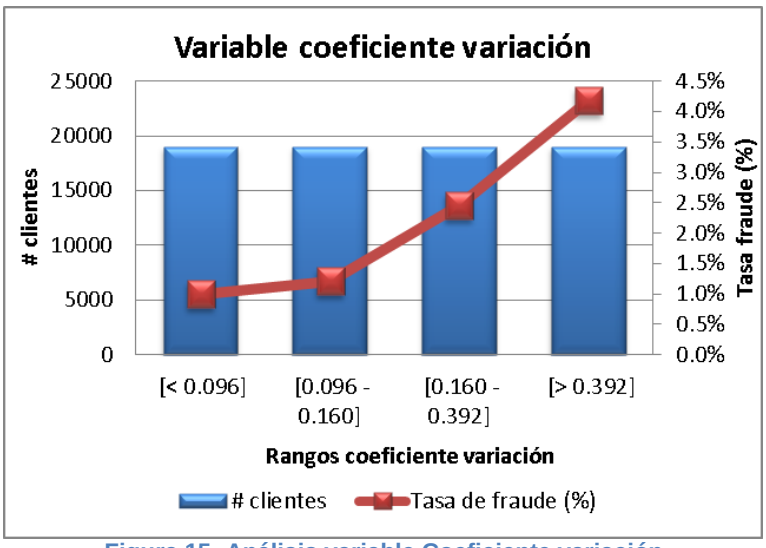

Figura 15: Análisis variable Coeficiente variación

- Variable *Cortes de suministro*

La variable cortes de suministro en los últimos 6 meses, también coincide con lo que se esperaría obtener, ya que entrega como coeficiente beta 0.59 lo cual indica que la presencia de más de un corte en un periodo de seis meses, incrementa la propensión al fraude. Según se aprecia en la figura 16, la presencia de la variable se traduce en una tasa dos veces mayor al caso en que no exista.

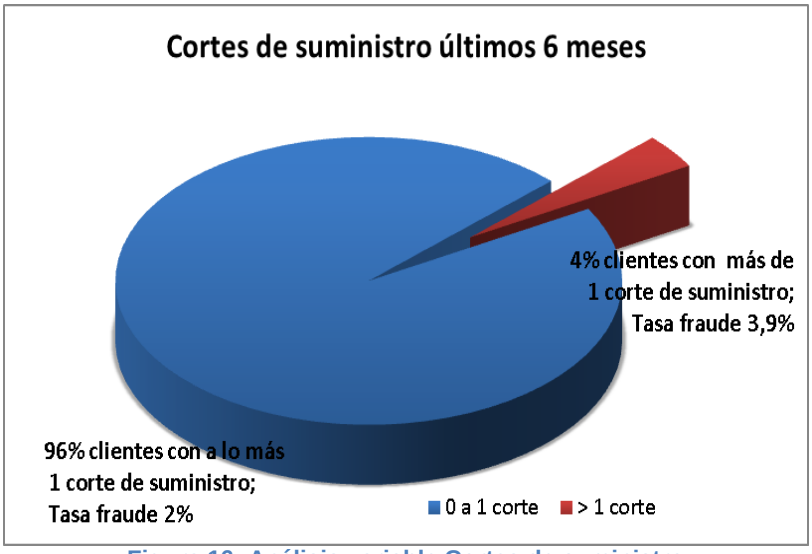

Figura 16: Análisis variable Cortes de suministro

- Variable *Falla de medidor (CL)*

La variable asociada a una clave de lectura correspondiente a "falla de medidor", que indica si el cliente tuvo alguna clave de lectura asociada a fallas en su medidor en los últimos 6 meses, corresponde a la variable más relevante del modelo entregando el mayor coeficiente beta igual a 2.1177. Esto quiere decir que para un cliente que tuvo alguna clave de lectura asociada a este fenómeno, aumenta considerablemente su probabilidad de estar cometiendo fraude energético.

Fallas_medidor_7-12 (CL)	Fraude = 0	Fraude = 1	Total	% de Total general
0	74231	1620	**75851**	**99.75%**
≥ 1	133	55	**188**	**0.25%**
Total general	74364	1675	**76039**	**100%**

Tabla 15: Frecuencia clave de lectura fallas medidor según clase

Si bien esta variable tiene la mayor importancia dentro del modelo, se aprecia en el análisis descriptivo de la tabla 15 que esta resulta tener pocos casos para el segmento positivo (Cuenta de fallas_medidor_7-12 ≥ 1) con solo un 0.25% de los casos, por lo cual la probabilidad de que se cuente con esta información en un cliente es muy baja, sin embargo cuando se encuentra, la tasa de fraude se incrementa considerablemente, alcanzando un 29.3% en este segmento.

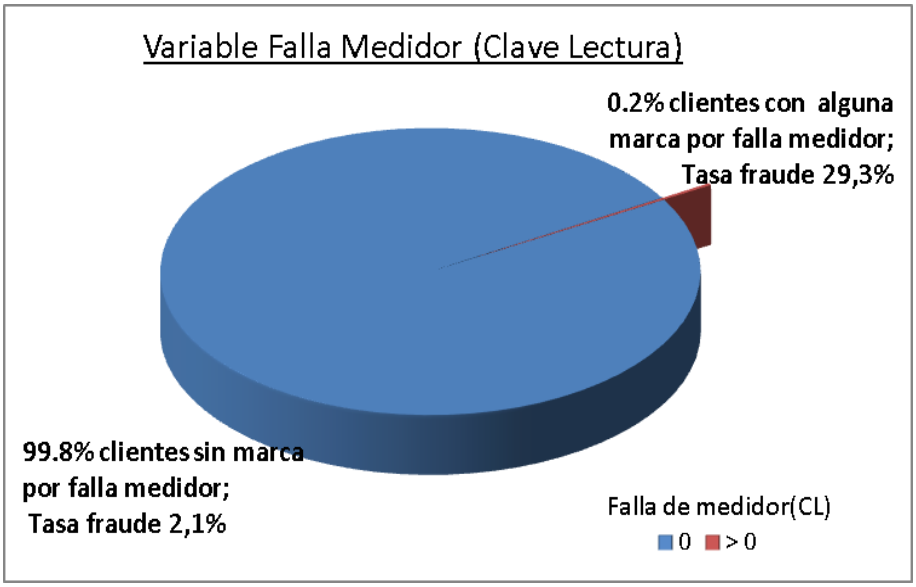

Figura 17: Análisis variable falla de medidor (CL)

- Variables *Consumo atípico (CL)* y *Sospecha fraude (CL)*

En el caso de la primera, un 11.5% de los clientes posee 1 o más clave de lectura asociada a consumo atípico, y se aprecia en la figura 18 que dentro de este grupo la tasa de fraude es mayor que en el grupo sin consumo atípico.

consumo_atipico_7-12 (CL)	Fraude = 0	Fraude = 1	Total	% de Total general
0	65892	1388	**67280**	**88.5%**
≥ 1	8472	287	**8759**	**11.5%**
Total general	74364	1675	**76039**	**100%**

Tabla 16: Frecuencia clave de lectura consumo atípico según clase

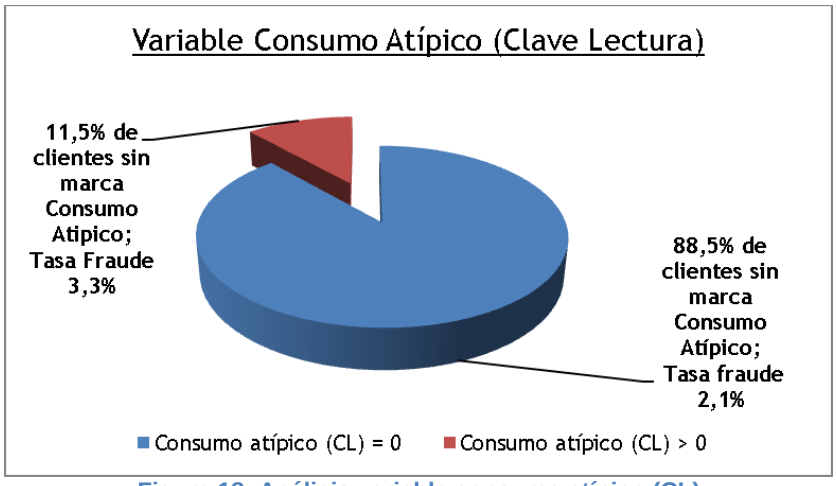

Figura 18: Análisis variable consumo atípico (CL)

Para la variable *Sospecha fraude (CL)* coincidente con el coeficiente beta entregado por la regresión (0.121) la tasa de fraude aumenta al incrementarse las marcas de esta clave de lectura. Según demuestra la figura 19 un 6.6% de los clientes posee una marca asociada a sospecha fraude alcanzando una tasa de fraude de un 3.4%, mientras que el grupo sin marcas, correspondiente a un 88.2% de los casos, alcanza una tasa de un 1.9%.

Sospecha_fraude_7-12 (CL)	Fraude = 0	Fraude = 1	Total general
0	65783	1305	**67088**
≥ 1	8581	370	**8951**
Total general	74364	1675	**76039**

Tabla 17: Frecuencia clave de lectura sospecha de fraude

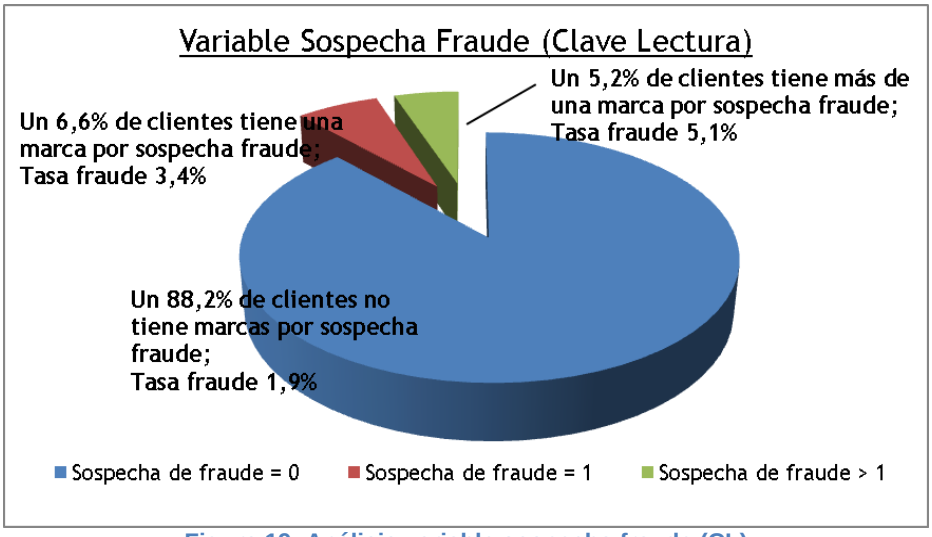

Figura 19: Análisis variable sospecha fraude (CL)

· Variable *Ratio consumo y consumo mínimo*

Se observa que los coeficientes obtenidos para la variable ratio consumo resultan esperables en los primeros tres rangos (coeficientes betas decrecientes), sin embargo se destaca una baja en el rango 4 (categoría de referencia, beta=0) asociado a los ratios más altos. La figura 20 grafica este fenómeno, donde podría plantearse como hipótesis que por solo tener un comportamiento de consumo anormal (ya sea por alzas o bajas de consumo) es más propenso al hurto de energía.

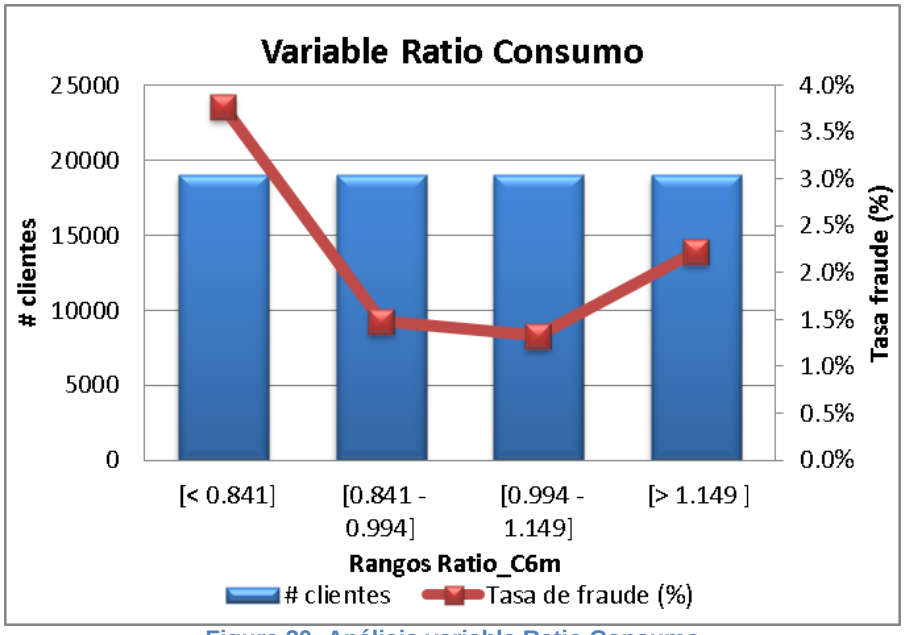

Figura 20: Análisis variable Ratio Consumo

Para la variable consumo mínimo ocurre algo similar, según se aprecia en la figura 21 esto podría seguir la misma lógica de la hipótesis anteriormente planteada, sin embargo

como es de esperar, las mayores tasas de fraudes se encuentran en los primeros cuartiles, donde el consumo mínimo no supera los 88.5 KWh[25].

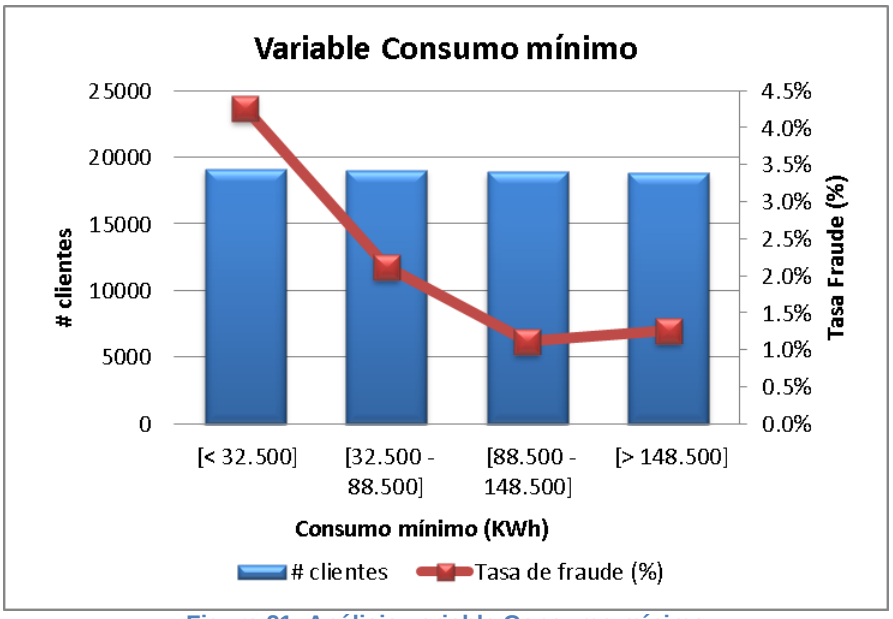

Figura 21: Análisis variable Consumo mínimo

- Variable *zona*

De acuerdo a los resultados del modelo, la variable zona resulta significativa en la predicción de un fraude, encontrándose mayores tasas en las zonas 5 y 6, aun cuando el número de inspecciones de estas zonas es menor en comparación a las zonas 1 y 2.

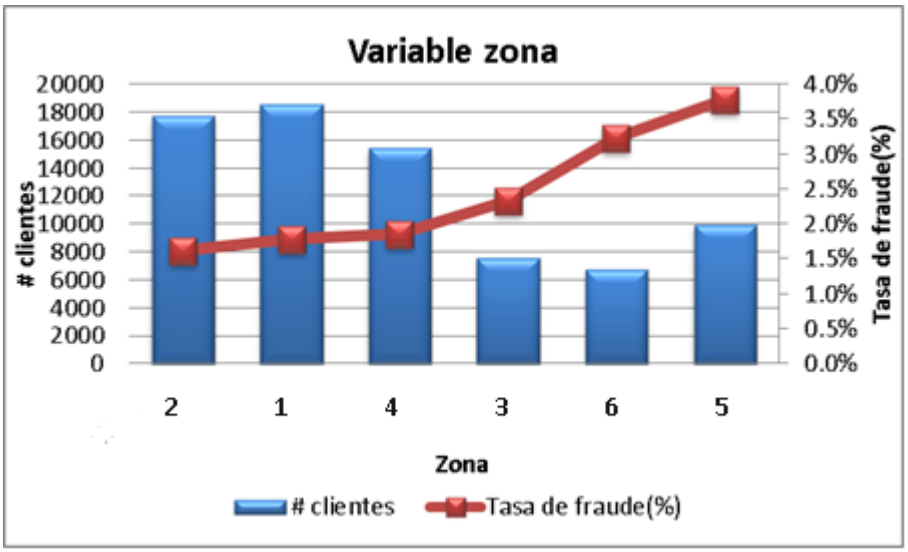

Figura 22: Análisis variable zona

[25] Un consumidor residencial en promedio consume 150 KWh mensual.

La principal hipótesis de este fenómeno recae en que en las zonas 1, 2 y 4, periódicamente se realizan las denominadas inspecciones de "barrido" donde se selecciona algún sector de la zona y se inspecciona de forma masiva. Al contrario, en las zonas 3, 5 y 6 se realizan inspecciones más minuciosas basadas netamente en sospechas de fraudes previas de los clientes, lo cual implicaría, según las cifras encontradas, inspecciones de mayor efectividad.

El principal inconveniente de incorporar la variable zona al modelo, es que al obtener un listado de clientes con las probabilidades más altas de fraude, es probable que este se concentre solo en una zona, indicada como la más fraudulenta y con el mayor coeficiente dentro de la regresión.

Evaluación de resultados

Clasificación

Cabe destacar que el criterio de clasificación por default del algoritmo en todos los programas donde está implementado corresponde a Probabilidad (fraude=1) = 0.5, es decir, un cliente con una probabilidad menor a 0.5 no será clasificado como fraude.
La figura 23 muestra un gráfico de frecuencias para cada valor de Probabilidad (fraude=1) obtenido en la data de prueba, donde se aprecia una clara concentración de casos en valores pequeños de p.

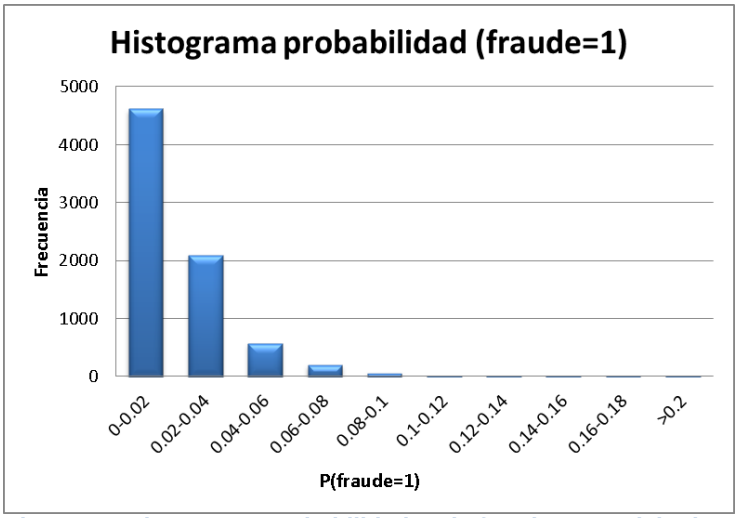

Figura 23: Histograma probabilidades de fraude en modelo de regresión

De manera natural surge la pregunta de si ¿es correcto clasificar como fraude un caso con Probabilidad de fraude >0.5? La respuesta a esta interrogante incorpora el que hecho que exista una máxima capacidad de inspección, que implicará inspeccionar mensualmente aquellos clientes que alcancen las mayores probabilidades de fraude, independiente de cual sea su valor. Por este motivo es que se utiliza como criterio de evaluación la *curva de ganancia*, la cual grafica el porcentaje de clientes declarados como fraude versus el verdadero porcentaje de fraude que se detecta en cada grupo.

La tabla 18 adjunta detalla las probabilidades promedio de fraude obtenidas para cada decil. Es posible apreciar que el grupo más propenso a fraude (decil n°1) congrega todos aquellos clientes cuya probabilidad de fraude obtenida sea mayor a 0.043, es decir, a un 4.3%.

% clientes	Probabilidad de corte	Probabilidad promedio fraude	Total acumulado	Acumulado Fraude = 1	Porcentaje fraude capturado	Tasa (Probabilidad) real de fraude
10%	≥ 0.043	**6.74%**	747	59	35.12%	**7.90%**
20%	≥ 0.031	**5.23%**	1455	83	49.40%	**5.70%**
30%	≥ 0.025	**4.36%**	2277	99	58.93%	**4.35%**
40%	≥ 0.020	**3.82%**	3037	112	66.67%	**3.69%**
50%	≥ 0.016	**3.41%**	3793	128	76.19%	**3.37%**
60%	≥ 0.012	**3.08%**	4557	139	82.74%	**3.05%**
70%	≥ 0.010	**2.79%**	5320	148	88.10%	**2.78%**
80%	≥ 0.008	**2.56%**	6055	157	93.45%	**2.59%**
90%	≥ 0.006	**2.35%**	6843	162	96.43%	**2.37%**
100%	≥ 0	**2.17%**	7604	168	100.00%	**2.21%**

Tabla 18: Resultados de clasificación - Modelo Regresión logística

Complementando el análisis anterior, según se aprecia en la gráfica de la figura 24, definiendo como clientes fraudulentos al primer decil de clientes (evaluado en la data prueba), se logra capturar un 35.12% del fraude, obteniendo una probabilidad esperada de un 6.74%.

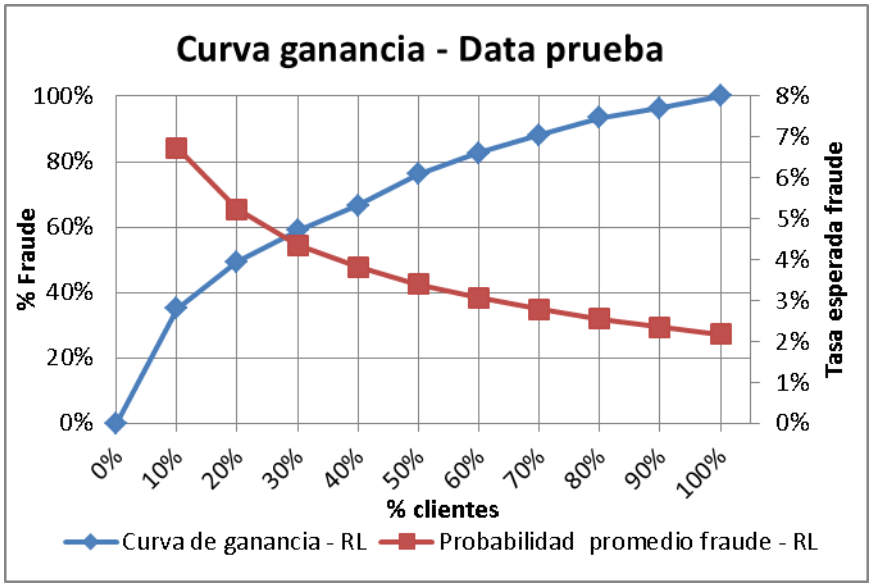

Figura 24: Curva de ganancia - Modelo Regresión logística

Cabe destacar que la probabilidad real de fraude versus la probabilidad esperada de cada grupo, coincide en casi la totalidad de los casos, a excepción de los dos primeros deciles donde se tiene un pequeño sesgo en la probabilidad obtenida.

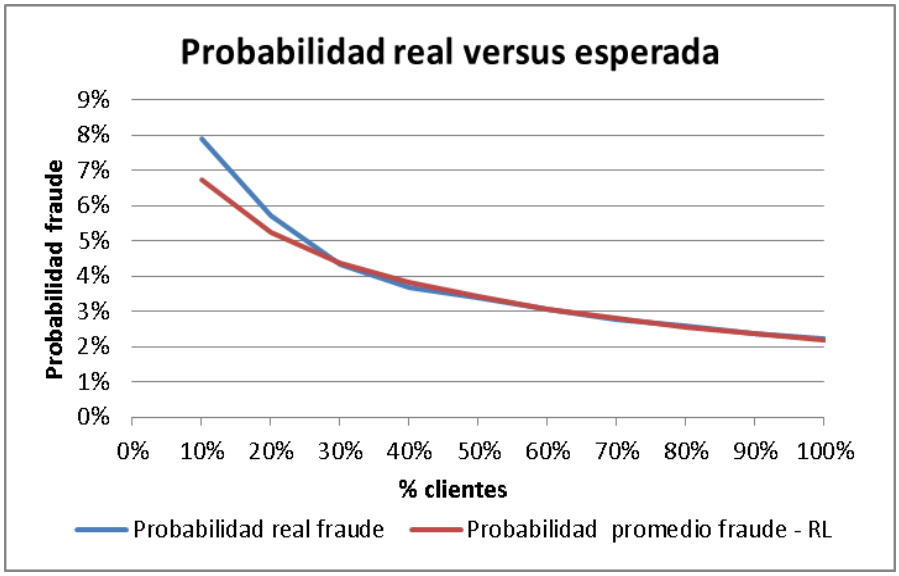

Figura 25: Gráfico de probabilidad real versus esperada - Modelo Regresión logística

Eficiencia operacional

Respecto al tiempo de ejecución del modelo de regresión logística este tarda 1.5 minutos en completar su ejecución, incluyendo como parte del proceso la transformación de las variables independientes desde variables continuas a nominales.

8.1.2 Modelo regresión logística (ponderado)

Coeficientes

El modelo de regresión logística ponderada entrega resultados similares en sus coeficientes, donde según se aprecia en la tabla 19, el único cambio importante recae en el valor del intercepto:

		Modelo sin ponderar		Modelo ponderado	
Variable	Tipo	*B*	Significancia *p-value*	*B*	Significancia *p-value*
Cortes*	Binaria	0.590	.000	0.574	.000
Consumo atípico (CL)*	Binaria	0.326	.002	0.347	.000
Meses sin consumo*	Binaria	0.166	.044	0.152	.000
Coeficiente variación consumo*	Binaria	0.000	.000		.000
Coeficiente variación consumo* =rango 2 [0.096 - 0.160]	Binaria	0.101	.112	0.103	.000
Coeficiente variación consumo* =rango 3 [0.160 - 0.392]	Binaria	0.633	.000	0.618	.000
Coeficiente variación consumo* =rango 4 [> 0.392]	Binaria	0.675	.000	0.670	.000
Consumo mínimo*			.000		.000
Consumo mínimo* =rango 1 [< 32.5]	Binaria	0.535	.000	0.505	.000
Consumo mínimo* =rango 2 [32.5 - 88.5]	Binaria	0.296	.001	0.311	.000
Consumo mínimo* =rango 3 [88.5 - 148.5]	Binaria	-0.177	.064	-0.169	.000
Ratio consumo			.000		.000
Ratio consumo =rango 1 [< 0.841]	Binaria	0.208	.002	0.199	.000
Ratio consumo=rango 2 [0.841 - 0.994]	Binaria	-0.132	.099	-0.120	.000
Ratio consumo =rango 3 [0.994 - 1.149]	Binaria	-0.174	.037	-0.201	.000
zona			.000		.000
zona=2	Binaria	0.022	.000	-0.147	.000
zona=3	Binaria	0.433	.000	0.238	.000
zona=4	Binaria	0.314	.000	0.213	.000
zona=5	Binaria	1.012	.000	0.835	.000
zona=6	Binaria	0.681	.188	0.511	.000
Sospecha de fraude(CL)*	Entero	0.121	.000	0.154	.000
Falla de medidor(CL)*	Binaria	2.118	.000	2.212	.000
Intercepto		-4.956	.000	-1.034	.000

Tabla 19: Coeficientes regresión logística ponderada

Evaluación y comparación de resultados

Como es de esperar en un modelo de clases balanceadas, las probabilidades asociadas a la presencia de fraude alcanzan valores más altos en comparación a un modelo desbalanceado como el caso anterior. El modelo ponderado sitúa en el primer decil a todos los casos con probabilidad mayor a 0.656:

% clientes	Probabilidad de corte	Probabilidad promedio fraude	Total acumulado	Acumulado Fraude = 1	Porcentaje fraude capturado
10%	≥ 0.656	**73.35%**	747	57	33.93%
20%	≥ 0.588	**67.93%**	1455	81	48.21%
30%	≥ 0.535	**63.67%**	2282	99	58.93%
40%	≥ 0.468	**60.27%**	3041	113	67.26%
50%	≥ 0.414	**57.04%**	3803	129	76.79%
60%	≥ 0.360	**54.03%**	4557	139	82.74%
70%	≥ 0.312	**51.17%**	5300	147	87.50%
80%	≥ 0.274	**48.50%**	6036	159	94.64%
90%	≥ 0.233	**45.82%**	6827	163	97.02%
100%	≥ 0	**43.24%**	7604	168	100.00%

Tabla 20: Resultados de clasificación - Modelo Regresión logística ponderado

La curva de ganancia situada a la derecha de la figura 26, construida con la data de prueba, muestra que con el primer decil de clientes se logra capturar un 33.93% del fraude mientras que en el modelo de regresión sin ponderar este valor asciende a 35.12%.

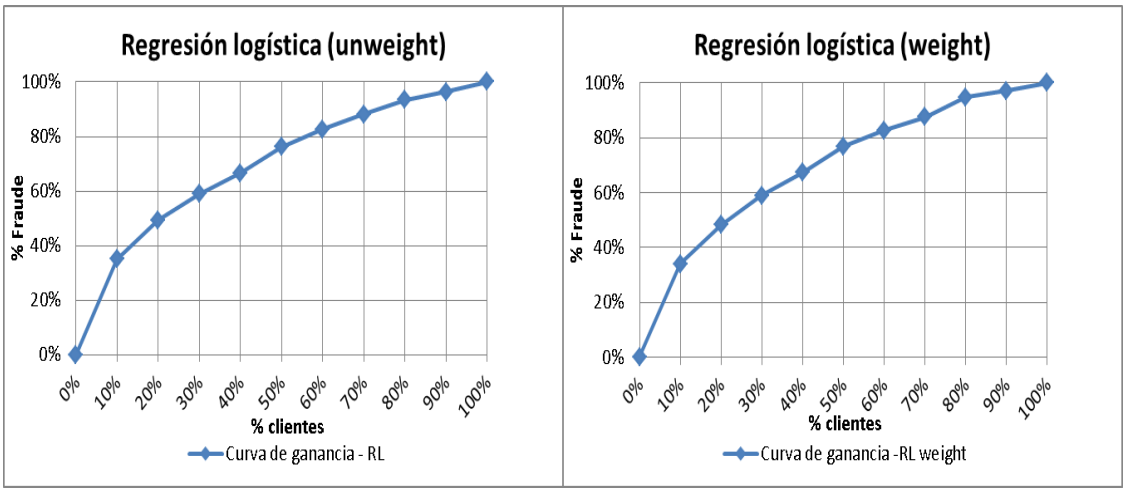

Figura 26: Comparación curvas de ganancia modelo regresión con y sin ponderación

Comparando ambos modelos, es posible concluir que para este problema en particular el desbalanceo de clases no afecta los resultados que se esperan obtener, ya que al existir una máxima capacidad de inspección, se requiere para ambos casos el ordenamiento en forma decreciente de los clientes, en función de su probabilidad de fraude, obteniéndose similares resultados según se observa en las curvas de ganancia de ambos modelos (figura 26).

8.1.3 Modelo árbol de decisión

Resultados modelo ID3

Como es de esperar, el modelo ID3, al no existir el proceso de poda, entrega un extenso árbol (extracto en anexo E).

Cabe destacar que todo árbol de decisión puede ser leído como un conjunto de reglas del tipo *Sí* {condición}, *Entonces* {clase predicha por la regla}. Cada regla se asocia a un camino desde un nodo raíz hasta un nodo terminal. A modo de ejemplo se detallan algunas reglas obtenidas desde el árbol ID3:

Ejemplo 1:

$$SÍ\ coeficiente_{variación_{7-12}} < 0.096 \cap zona: 2$$
$$\cap\ sospecha_{fraude_{CL_{7-12}}}: true \cap falla_{medidor_{CL_{7-12}}}: true,$$
$$ENTONCES\ fraude = 1$$

Ejemplo 2:

$$SÍ\ zona: 4 \cap meses_{cero_{consumo_{7-12}}}: true$$
$$\cap\ consumo_{minimo_{7-12}} < 32.5, ENTONCES\ fraude = 1$$

Las reglas obtenidas coinciden con lo que se espera obtener, entregando por ejemplo, una mayor probabilidad de fraude a las zonas 4 o 5, o a aquellos clientes que poseen al menos un mes sin registrar consumo durante el último semestre.

Respecto a los resultados de clasificación, de acuerdo a la curva de ganancia de la data de prueba (figura 28), los resultados no coinciden con los de la data de calibración (figura 27), obteniéndose una ganancia de solo un 27.9% en el primer decil de clientes a diferencia del 43.9 % obtenido en la calibración del modelo. Lo anterior es parte de uno de los inconvenientes que presenta el modelo ID3 denominado sobreajuste, donde el árbol construido se ajusta muy bien a la data de calibración generando reglas (caminos del árbol) demasiado específicas, que al ser aplicadas a la data de prueba no logran clasificar correctamente los casos.

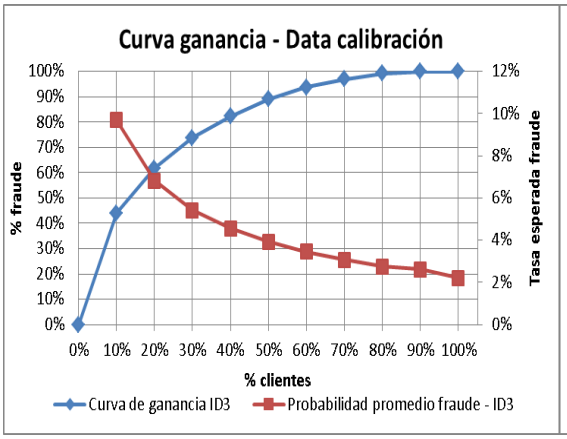

Figura 27: Curva ganancia ID3 - calibración

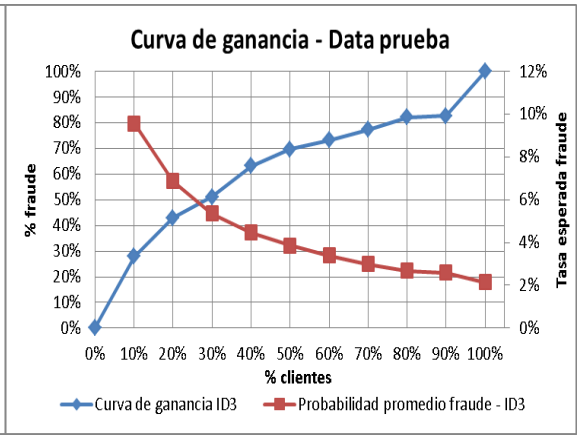

Figura 28: Curva ganancia ID3 - prueba

Resultados modelo C4.5

Como se observa en ambas curvas de ganancia (figuras 29 y 30), los resultados coinciden tanto para la data de calibración como la de prueba, logrando capturar en el primer decil un 32.7% y 32.1% de fraude respectivamente, comprobando que el modelo no incorpora sobreajuste y logra clasificar correctamente los ejemplos no utilizados en la construcción del modelo.

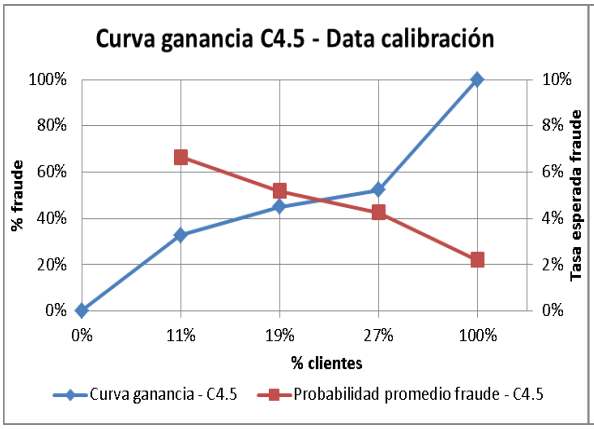

Figura 29: Curva ganancia C4.5 - calibración

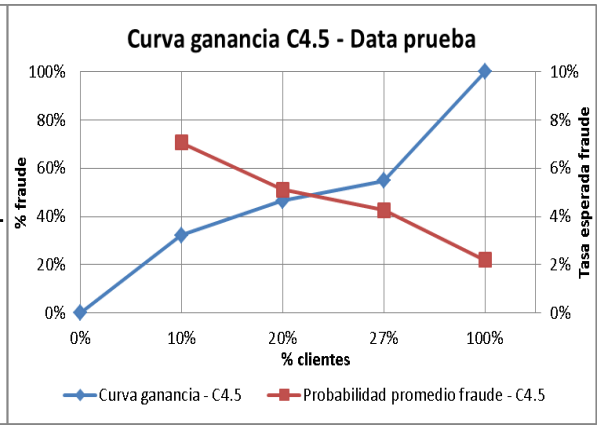

Figura 30: Curva ganancia C4.5 – prueba

Cabe destacar del modelo C4.5, que al ser sometido a un proceso de poda posterior a la construcción del árbol, la cantidad de nodos terminales (árbol completo en anexo F) disminuye considerablemente respecto al árbol ID3. Recordando que cada nodo terminal se asocia a una probabilidad de fraude es esperable que la mayoría de los casos (más de un 70%) sean etiquetados con la misma probabilidad, sin embargo, al igual que en el modelo de regresión logística el primer decil de la data de prueba corresponde al grupo de clientes con una probabilidad de fraude mayor o igual a 0.043, y alcanza una probabilidad de fraude esperada de 7.07% versus el 6.87% existente en la realidad.

% clientes	Probabilidad de corte	Probabilidad promedio fraude	Total acumulado	Acumulado Fraude = 1	Porcentaje fraude capurado	Tasa (probabilidad) real de fraude
10%	≥ 0.043	**7.07%**	786	54	32.14%	**6.87%**
20%	≥ 0.025	**5.11%**	1504	78	46.43%	**5.19%**
27%	≥ 0.020	**4.26%**	2084	92	54.76%	**4.41%**
100%	≥ 0	**2.19%**	7604	168	100.00%	**2.21%**

Tabla 21: Resultados de clasificación - Modelo Árbol decisión C4.5

Eficiencia operacional

Respecto al tiempo de ejecución de ambos modelos, para el caso del modelo ID3 su ejecución en el software Rapid Miner tarda 1.2 minutos mientras que el modelo C4.5 toma 2.8 minutos en completarse.

8.1.4 Modelo random forest

El siguiente cuadro resume los resultados para ambos modelos construidos, donde es posible apreciar que en cuanto a ganancia de información en el primer decil de clientes (asociados a las mayores probabilidades de fraude), el modelo construido con 100 árboles entrega mejores resultados que el modelo construido con solo 10, sin embargo operacionalmente el modelo 2 resulta ser poco eficiente tardando más de un día en lograr su ejecución.

Criterio	Modelo 1 (RF 10)	Modelo 2 (RF 100)
Ganancia decil n°1	34.52%	**39.88%**
Tasa fraude esperada decil n°1	7.42%	6.84%
Tiempo de ejecución	3.5 horas	**25.1 horas**

Tabla 22: Resultados según parámetros - Modelo Random Forest

Debido a que con el modelo n° 2 se obtuvieron mejores resultados de clasificación, no se entrará en más detalles respecto al modelo n°1.

Modelo 2: RF 100

Análogo al modelo de regresión logística sin ponderar y al árbol de decisión, el modelo de Random Forest también entrega bajas probabilidades de fraude. Según es posible apreciar en la tabla 23, el rango más alto comienza en 0.043, sin embargo como se comprobó en el modelo de regresión logística ponderada, la magnitud de esta probabilidad no afecta los resultados que se espera obtener ya que se debe solo al problema de desbalanceo.

Decil	Probabilidad de corte	Probabilidad promedio fraude	Total acumulado	Acumulado Fraude = 1	Porcentaje fraude capturado	Tasa (probabilidad) real de fraude
10%	≥ 0.043	6.84%	761	67	**39.88%**	8.80%
20%	≥ 0.030	5.21%	1521	90	**53.57%**	5.92%
30%	≥ 0.023	4.34%	2282	107	**63.69%**	4.69%
40%	≥ 0.018	3.77%	3042	123	**73.21%**	4.04%
50%	≥ 0.015	3.35%	3803	132	**78.57%**	3.47%
60%	≥ 0.013	3.02%	4563	145	**86.31%**	3.18%
70%	≥ 0.010	2.76%	5323	151	**89.88%**	2.84%
80%	≥ 0.009	2.53%	6084	157	**93.45%**	2.58%
90%	≥ 0.007	2.34%	6844	164	**97.62%**	2.40%
100%	≥ 0	2.17%	7604	168	**100.00%**	2.21%

Tabla 23: Resultados de clasificación - Modelo Random Forest (100 árboles)

La curva de ganancia del modelo (figura 31) indica que, bajo este criterio, el modelo de Random Forest construido con 100 árboles de decisión, entrega los mejores resultados en comparación a todos los modelos previamente construidos sin embargo, subestima la probabilidad de fraude esperada para todos los casos (figura 32).

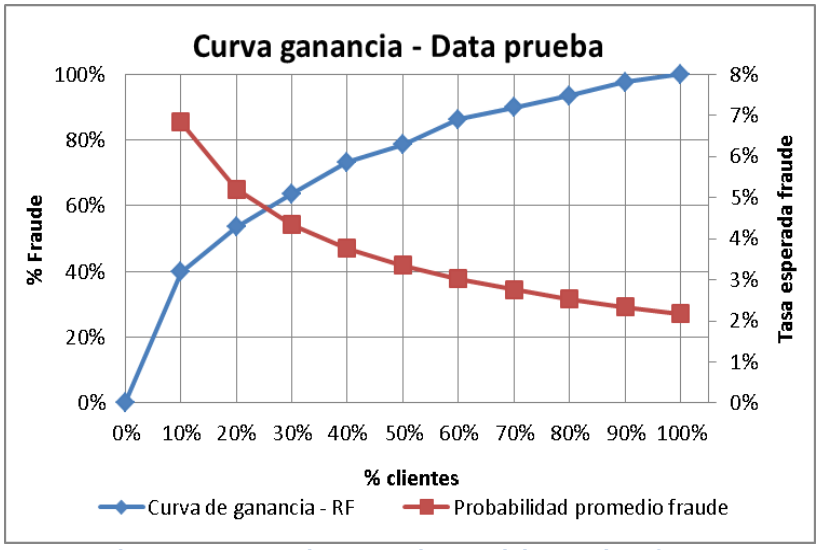

Figura 31: Curva de ganancia - Modelo Random forest

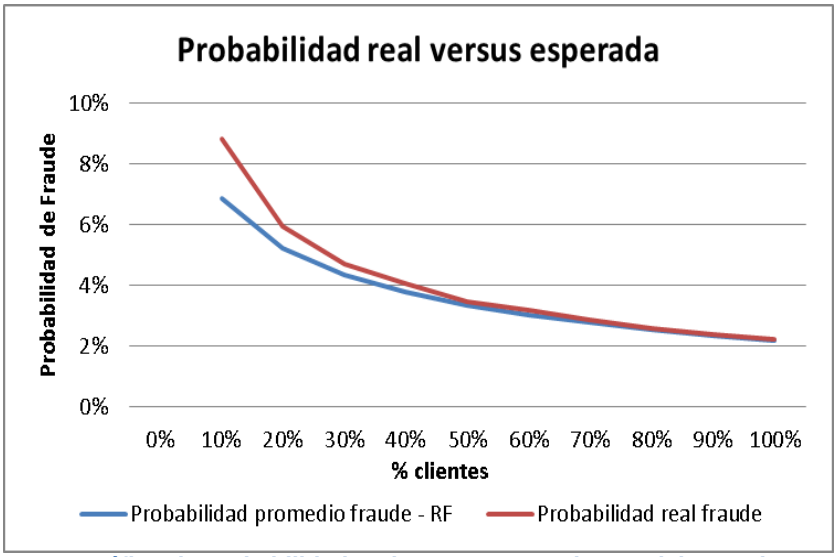

Figura 32: Gráfico de probabilidad real versus esperada - Modelo Random Forest

8.1.5 Comparación y elección modelo

Para comparar los modelos construidos se utilizaron tres criterios considerados los más importantes:

Criterio	Modelo Regresión logística	Modelo Árbol de decisión	Modelo Random Forest
Ganancia (decil n°1)	35.12%	32.1%	39.29%
Tasa fraude esperada (decil n°1)	6.74%	7.07%	6.04%
Tiempo de ejecución	1.5 minutos	2.8 minutos	25.1 horas

Tabla 24: Cuadro comparativo de modelos

Según es posible apreciar en la tabla 24, si bien el modelo random forest entrega los mejores resultados en cuanto a clasificación, su tiempo de ejecución es mayor a los otros dos modelos. Sumado a esto, de acuerdo a los objetivos del trabajo, se espera poder analizar cómo y cuánto influyen las variables relacionadas al hurto de energía, por lo cual en este sentido el modelo de regresión logística es el que entrega la mejor solución.

Como conclusión se escoge el modelo de regresión logística por sobre los otros dos modelos, debido a la facilidad en su interpretación y breve tiempo de ejecución.

Aplicación modelo escogido: regresión logística

A modo de aplicar el modelo en instancias no clasificadas, se aplica el modelo de regresión logística construido sobre la data de clientes no inspeccionados, efectuando previamente las transformaciones necesarias[26] de las variables utilizadas en el modelo.

Aplicando el modelo, se ordenan los clientes de manera decreciente según su probabilidad de fraude obtenida y se seleccionan los primeros 7800 (de acuerdo a la máxima capacidad de inspección impuesta por la empresa). Como se observa en la figura 33, los consumidores más propensos a estar hurtando energía se concentran en la zona 5, resultado esperable debido a que el modelo de regresión entrega una mayor puntuación a casos pertenecientes a esta zona, seguido de los casos pertenecientes a la zona 6.

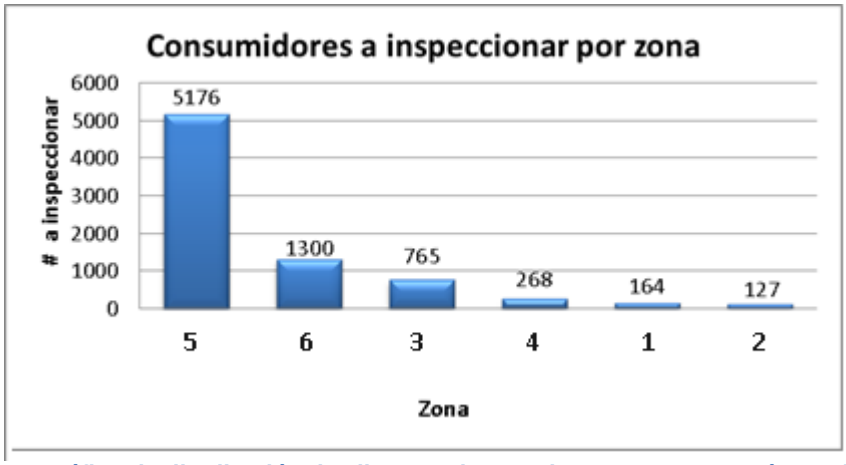

Figura 33: Gráfico de distribución de clientes a inspeccionar por zona según modelo de regresión

Bajo la mirada del negocio, es poco factible focalizar un 66% de las inspecciones solo en una zona, o asignar a la zona 2 solo un 1.6% de las inspecciones donde históricamente se ha registrado la mayor cantidad de estas. Como solución a este inconveniente se abordan dos alternativas: (1) restringir cantidad de inspecciones por zona, según la proporción histórica de inspecciones en estas[27], (2) Realizar modelo de regresión intra-zona (local) incorporando información más detallada, como por ejemplo, la comuna a la cual pertenece el cliente, información que a nivel global no es considerada debido a la complejidad de su obtención en la data disponible.

Alternativa (1): Restricción capacidad máxima de inspección por zona

Luego de aplicar el modelo de regresión a cada cliente, se agrupan según la zona a la cual pertenecen y se ordenan de manera decreciente según su probabilidad de fraude. Posteriormente se escogen los más propensos dentro de cada zona, considerando la proporción de inspecciones detallada a continuación:

[26] Se discretizan variables continuas utilizando los mismos rangos y criterios utilizados al construir la regresión.

[27] Ver sección 5 Análisis descriptivo – inspecciones por zona

Zona	Proporción de inspecciones	Total de inspecciones a realizar
1	23%	1768
2	24%	1872
3	11%	832
4	21%	1629
5	12%	971
6	9%	728
Total	**100%**	**7800**

Tabla 25: Proporción de clientes a inspeccionar por zona

Alternativa (2): Modelo de regresión logística intra-zona

A modo de evaluar la alternativa de un análisis local, se construye un modelo de regresión logística a nivel intra-zona. Este modelo incorpora dos variables extras que a nivel global no fueron consideradas, por el hecho de no aportar más información al modelo solo complejizándolo (variable consumo promedio), o por no contar con la información de manera inmediata. Este último caso corresponde a la variable comuna, la cual es información que no se tiene disponible en las fuentes oficiales, si no que debió ser extraída de forma manual desde los archivos iniciales puestos a disposición por la empresa, información que no se encuentra cargada de manera oficial en los servidores destinados al proyecto.

Realizando el modelo de regresión logística a nivel local, utilizando la zona 5, se obtuvieron los siguientes resultados para sus coeficientes:

Variable	Tipo	Beta (fraude=1)	Significancia *p-value*
Cortes*	Binaria	0.756	0.000
Consumo atípico (CL)*	Binaria	0.052	0.850
Meses sin consumo*	Binaria	0.249	0.190
Coeficiente variación consumo*	Binaria		0.003
Coeficiente variación consumo* =rango 2 [0.095 - 0.150]	Binaria	-0.014	0.950
Coeficiente variación consumo* =rango 3 [0.150 - 0.300]	Binaria	0.499	0.013
Coeficiente variación consumo* =rango 4 [>0.300]	Binaria	0.641	0.005
Consumo mínimo (KWh)*			0.032
Consumo mínimo* =rango 1 [< 47.5]	Binaria	0.445	0.174
Consumo mínimo* =rango 2 [47.5 - 97.5]	Binaria	-0.147	0.630
Consumo mínimo* =rango 3 [97.5 - 152.5]	Binaria	-0.048	0.859
Ratio consumo			0.002
Ratio consumo =rango 1 [< 0.857]	Binaria	-0.610	0.001
Ratio consumo=rango 2 [0.857 - 1.004]	Binaria	-0.493	0.007
Ratio consumo =rango 3 [1.004 - 1.165]	Binaria	-0.131	0.360
Consumo promedio (KWh)*			0.002
Consumo promedio * =rango 1 [< 82.583]	Binaria	0.447	0.103
Consumo promedio* =rango 2 [82.583 - 131.417]	Binaria	-0.127	0.649
Consumo promedio* =rango 3 [131.417 - 197.250]	Binaria	-0.284	0.262
Comuna			0.000
Comuna=5.a	Binaria	1.024	0.001
Comuna=5.b	Binaria	-0.324	0.455
Comuna=5.c	Binaria	0.484	0.060
Comuna=5.d	Binaria	-0.690	0.237
Comuna=5.e	Binaria	-0.248	0.474
Comuna=5.f	Binaria	1.037	0.001
Comuna=5.g	Binaria	-0.202	0.589
Comuna=5.h	Binaria	-0.062	0.810
Falla de medidor(CL)*	Binaria	2.510	0.000
Intercepto		-3.881	0.000

Tabla 26: Coeficientes modelo regresión logística – Zona 5

Según es posible observar en la tabla 26, la significancia de los coeficientes obtenidos, no cumplen los parámetros estipulados para la significancia del modelo (*p-value*<0.1) en la mayoría de las variables introducidas, por lo cual se puede concluir que el modelo no resulta significativo. Este resultado es esperado debido al bajo número de fraudes que existen a nivel local, en particular para la zona 5, se cuenta solo con 375 casos de fraude, lo cual dificulta la creación de un modelo representativo.

8.2 Modelo de clustering

Clientes residenciales

Utilizando el algoritmo K-medias, se realizaron varias iteraciones cambiando el valor del parámetro K, escogiendo aquel que minimiza el índice Davies-Bouldin. La figura 34 muestra la variación del índice DB en función del parámetro K, donde es posible apreciar que la mejor partición de los datos se alcanza bajo la construcción de **K=15** clústeres.

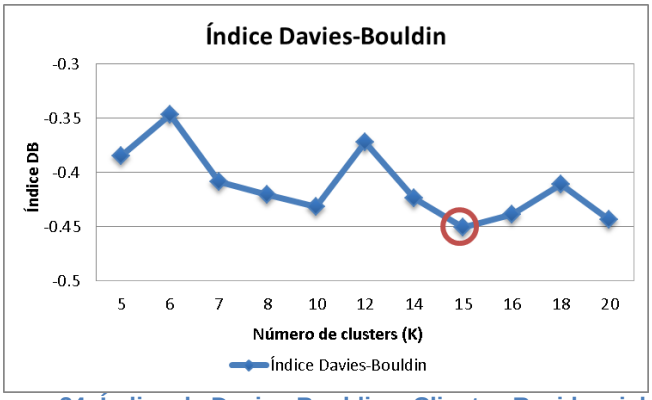

Figura 34: Índice de Davies-Bouldin – Clientes Residenciales

Para el parámetro K obtenido, la distribución de casos en los K=15 clústeres construidos se detalla en la tabla 27 donde se observa que solo los clústeres 1,7,9,11,13,14 y 15 poseen un tamaño significativo (más de 100 casos).

Grupo	Número de casos	Centroide ratio consumo	Centroide consumo promedio 2° semestre (KWh)	Centroide consumo mínimo 2° semestre (KWh)	Centroide coeficiente variación 2° semestre
clúster 1	**115247**	**1.05**	**92.62**	**73.28**	**0.16**
clúster 2	1	2069.00	344.83	231.00	0.41
clúster 3	2	979.00	163.17	0.00	0.70
clúster 4	1	89.87	11038.50	111.00	2.42
clúster 5	32	384.91	129.99	17.84	0.79
clúster 6	1	1654.00	275.67	0.00	1.83
clúster 7	**3086**	**1.11**	**755.03**	**605.63**	**0.16**
clúster 8	1	59.03	7762.83	83.00	2.42
clúster 9	**115**	**142.46**	**73.79**	**15.59**	**0.97**
clúster 10	18	642.50	113.39	12.06	0.87
clúster 11	**30236**	**1.37**	**89.42**	**17.48**	**0.68**
clúster 12	1	1420.00	236.67	71.00	0.67
clúster 13	**23634**	**1.12**	**358.01**	**286.10**	**0.16**
clúster 14	**88147**	**1.09**	**191.87**	**155.75**	**0.15**
clúster 15	**7295**	**1.52**	**29.27**	**2.24**	**1.81**
Total	**267817**				

Tabla 27: Número de casos y centroides por clúster (Residencial)

Los centroides de cada grupo (tabla 27) ayudan a identificar a qué tipo de clientes corresponde cada clúster, por lo cual se efectúa una breve descripción de cada grupo calificado como significativo:

- **Clúster 1**: Grupo *"clientes residenciales con consumo promedio bajo".* Su ratio consumo es cercano a 1 por lo cual no poseen grandes variaciones entre un semestre y otro. El centro del grupo en la variable consumo promedio alcanza los 92 KWh mientras que el consumo mínimo llega a los 73 KWh lo que demuestra un consumo parejo en el tiempo pero considerado bajo en un cliente residencial.

- **Clúster 7**: Grupo *"clientes residenciales con consumo promedio alto"*. Poseen un alto consumo promedio, por lo cual es probable que en la realidad no correspondan a clientes residenciales si no a comerciales o industriales, sin embargo, se caracterizan en que a pesar de presentar un alto consumo este se mantiene estable en el tiempo.

- **Clúster 9**: Grupo *"clientes residenciales con consumo promedio bajo y altas variaciones"*. Poseen un consumo promedio bajo, un gran ratio consumo (142) y un coeficiente de variación lejano a cero, lo cual indica grandes variaciones de consumo en el tiempo.

- **Clúster 11**: Grupo *"clientes residenciales con consumo promedio bajo y medianas variaciones"*. Poseen un consumo promedio considerado bajo para un cliente residencial, sumado a un coeficiente de variación y un ratio consumo alejados de los valores considerados como ideales para un cliente (0 y 1 respectivamente).

- **Clúster 13**: Grupo *"clientes residenciales con consumo promedio alto"*. Poseen un alto consumo promedio, sin embargo no presentan variaciones importantes de este, según se observa en los valores de las variables ratio consumo y coeficiente de variación.

- **Clúster 14**: Grupo *"clientes residenciales con consumo promedio normal".* Su consumo promedio se encuentra en 191 KWh, mientras que su consumo mínimo alcanza los 155 KWh. Para la variable ratio consumo el centro del grupo es cercano a 1 y su coeficiente de variación se encuentra entre los menores valores obtenidos, lo cual es muestra de un grupo sin variaciones de consumo en el tiempo.

- **Clúster 15**: Grupo *"clientes residenciales con consumo promedio bajo y quiebres de consumo"*. Dentro de este grupo, tanto la variable consumo promedio como consumo mínimo toman valores considerados pequeños en magnitud, en particular la variable consumo mínimo promedio del grupo alcanza solo los 2.2KWh indicando cambios abruptos de consumo o también denominados "quiebres". Lo anterior es acompañado de un alto ratio consumo (1.5) por lo cual dentro de este grupo existirán una gran cantidad de clientes con altas fluctuaciones y puntos de quiebres en el consumo.

Detección casos outliers o anómalos

Como se mencionó en la sección anterior, no todos los grupos obtenidos contienen una cantidad significativas de casos. Aquellos grupos que posean menos de 100 casos serán considerados como grupos anómalos a nivel global, ingresando en primer lugar al listado de casos *outliers.* Esta primera selección suma en total 57 casos residenciales. Para la detección de casos anómalos a nivel local (clúster), utilizando la metodología descrita en la sección 7.6.2, se obtiene un total de 2676 casos que dentro de cada clúster significativo, poseen las mayores distancias euclidianas a sus respectivos centroides. Agregando ambos criterios se obtiene un total de 2733 casos *outliers* de tipo residencial.

Casos ejemplo

- *Clúster 13: Grupo "Clientes residenciales con consumo promedio alto":*

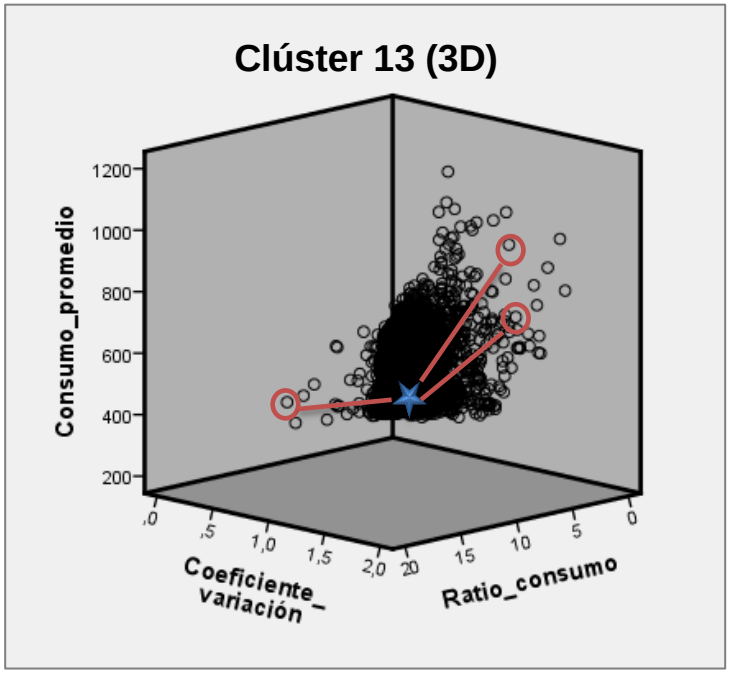

Figura 35: *Outliers* clúster 13 – Residencial (3D)

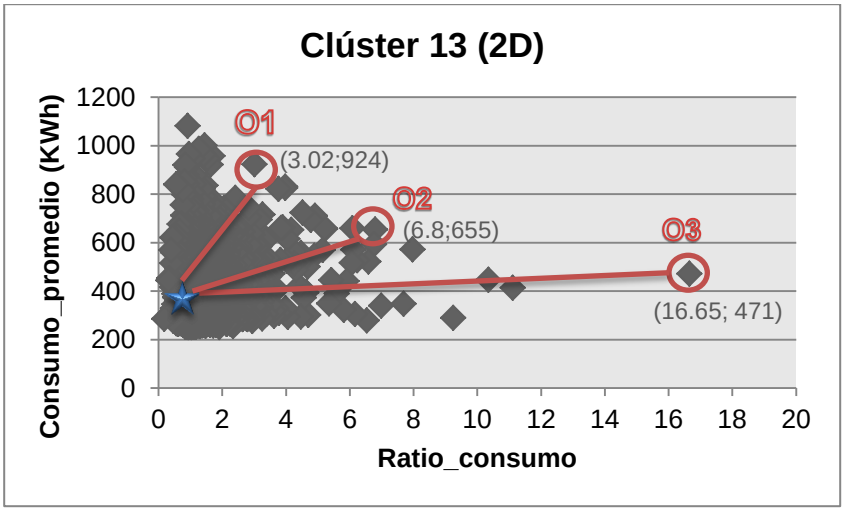

Figura 36: *Outliers* clúster 13 - Residencial (2D)

Como se definió anteriormente, el clúster 13 corresponde a aquellos clientes residenciales con un consumo promedio alto, con un consumo promedio de 358 KW, consumo mínimo promedio 286 KWh, coeficiente de variación de 0.16 y un ratio consumo cercano a 1 lo cual habla de un grupo que si bien posee un alto consumo (respecto al normal de un cliente residencial), no presenta fluctuaciones importantes de éste en el tiempo. Las figuras 35 y 36, en 3 y 2 dimensiones respectivamente, muestran ejemplos de casos *outliers* detectados dentro de este grupo. Los casos ejemplo de *outliers* denominados "O1", "O2" y "O3" corresponden a clientes residenciales con las siguientes características:

Outlier ejemplo	Ratio consumo	Consumo promedio 2° semestre (KWh)	Consumo mínimo 2° semestre (KWh)	Coeficiente variación 2° semestre
O1	3.02	924	0	1.34
O2	6.8	655	88	1.45
O3	16.65	471	58	0.72

Tabla 28: Casos ejemplo outliers - Clúster 13 residencial

Aplicación modelo regresión logística

A modo de evaluar y establecer métricas de comparación entre los resultados obtenidos, se aplica el modelo de regresión logística para obtener las probabilidades de fraude esperadas de cada clúster y las de cada cliente incluido en el listado total de *outliers.* La tabla 29 contiene el detalle de la tasa esperada de fraude para cada clúster de clientes residenciales junto a la tasa de fraude de sus respectivos *outliers* detectados tanto a nivel local como global[28].

[28] *Outliers* tipo global: Clústeres completos definidos como *outliers* dado que no alcanzan a incluir un tamaño significativo de casos (<100)
Outliers tipo local: Grupo de casos que dentro de aquellos clústeres significativos son identificados como outliers debido a que representan el 1% de casos más alejados de su respectivo centroide.

Clúster	N° casos clúster	Tasa fraude esperada clúster	N° casos *outliers*	Tasa fraude esperada *outliers*
clúster 1	115247	1.87%	1152	3.24%
clúster 2	1	2.09%	1	2.09%
clúster 3	2	2.51%	2	2.51%
clúster 4	1	2.51%	1	2.51%
clúster 5	32	3.72%	32	3.72%
clúster 6	1	5.62%	1	5.62%
clúster 7	3086	1.46%	30	2.04%
clúster 8	1	3.52%	1	3.52%
clúster 9	115	3.76%	2	3.42%
clúster 10	18	3.28%	18	3.28%
clúster 11	30236	3.93%	302	3.91%
clúster 12	1	4.87%	1	4.87%
clúster 13	23634	1.36%	236	3.74%
clúster 14	88147	1.25%	881	2.54%
clúster 15	7295	4.96%	73	3.42%
TOTAL	**267817**	**1.93%**	**2733**	**3.13%**

Tabla 29: Tasas esperadas de fraude clústeres y *outliers* – Residencial

Aun cuando el modelo de clustering y el de regresión no son comparables, es de esperar que la tasa de fraude en los casos outliers sea mayor o igual a la de su clúster de pertenencia. Este suceso se cumple en todos los casos a excepción de los clústeres 9 (el cual solo contiene dos casos outliers) y 15, donde los casos más alejados del grupo no coinciden con los de mayor probabilidad de fraude. Sin embargo, cabe destacar que el clúster 15 alcanza la tasa de fraude más alta entre los grupos lo cual es respaldado por los resultados obtenidos desde el modelo de clustering que señala a este grupo como: clientes residenciales con consumo promedio bajo y quiebres de consumo, que alcanzan un consumo promedio de solo 29 KWh junto a un coeficiente de variación y ratio consumo de 1.8 y 1.5 respectivamente.

Además, como es de esperar se observa que el clúster con la menor tasa de fraude esperada corresponde al clúster 14, el cual corresponde a clientes residenciales con consumo promedio normal (191KWh), un coeficiente de variación de 0.15 y un ratio de 1.09, evidencia de un consumo estable en el tiempo.

Clientes comerciales

Análogo al caso de los clientes residenciales, utilizando el algoritmo K-medias, se realizaron varias iteraciones cambiando el valor del parámetro K escogiendo aquel que minimiza el índice Davies-Bouldin. La figura 35 muestra la variación del índice DB en función del parámetro K, donde es posible apreciar que la mejor partición de los datos se alcanza bajo la construcción de **K=12** clústeres.

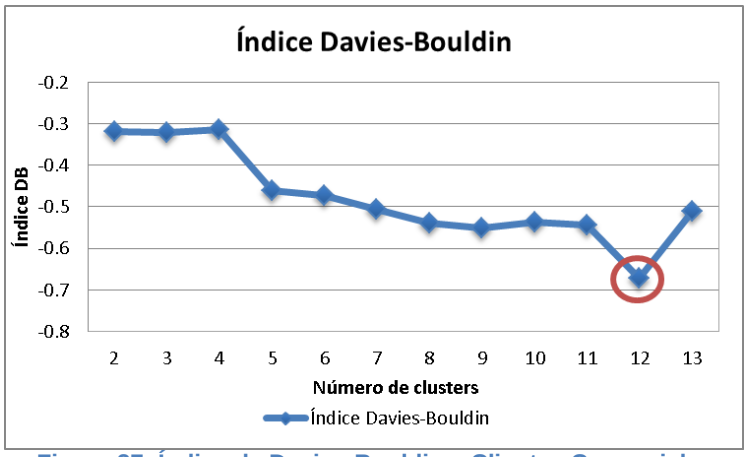

Figura 37: Índice de Davies-Bouldin – Clientes Comerciales

Para el parámetro K obtenido, la distribución de casos en los K=12 clústeres construidos se detalla en la tabla 30, donde se observa que solo los clústeres 1, 5, 7, 8 y 9 poseen un tamaño significativo (definido en 100 casos).

Grupo	Número de casos	Centroide Ratio consumo	Centroide consumo promedio 2° semestre (KWh)	Centroide consumo mínimo 2° semestre (KWh)	Centroide coeficiente variación 2° semestre
clúster 1	**5623**	**1.08**	**147.86**	**116.09**	**0.17**
clúster 2	1	970.00	161.67	138.00	0.08
clúster 3	2	509.00	374.75	0.00	1.86
clúster 4	2	255.33	301.42	15.00	1.22
clúster 5	**585**	**1.10**	**1080.39**	**899.84**	**0.13**
clúster 6	5	68.51	70.80	6.40	1.30
clúster 7	**1477**	**1.43**	**130.60**	**22.69**	**0.70**
clúster 8	**2336**	**1.05**	**498.98**	**394.91**	**0.16**
clúster 9	**288**	**1.89**	**55.68**	**1.40**	**1.82**
clúster 10	2	441.29	256.17	45.00	0.56
clúster 11	3	111.52	180.89	53.33	0.51
clúster 12	3	200.67	66.89	0.67	1.48
Total	**10327**				

Tabla 30: Número de casos y centroides por clúster (Comercial)

Los centroides de cada grupo (tabla 30) ayudan a identificar a qué tipo de clientes corresponde cada clúster, por lo cual se efectúa una breve descripción de cada grupo calificado como significativo:

- **Clúster 1**: Grupo *"clientes comerciales con bajo consumo promedio"*. Poseen un consumo promedio y mínimo considerado bajo para un cliente comercial[29], por lo cual es probable que en la realidad correspondan a un cliente residencial. Además, poseen un ratio consumo cercano a 1 y un bajo coeficiente de variación, lo cual indicaría que no presentan fluctuaciones importantes de consumo.

- **Clúster 5**: Grupo *"clientes comerciales con consumo promedio alto"*. Poseen un alto consumo promedio y mínimo en comparación al consumo normal de un cliente comercial (350KW), sin embargo las variables ratio consumo y coeficiente variación alcanzan valores que indican un consumo estable en el tiempo.

- **Clúster 7**: Grupo *"clientes comerciales con consumo promedio bajo"*. Poseen un consumo promedio considerado bajo para un cliente comercial sumado a un consumo mínimo bajo para cualquier tipo de cliente. En el caso de las variables ratio consumo y coeficiente de variación, ambas se alejan de los valores ideales en dichos parámetros (cercanos a 1 y 0 respectivamente), por lo cual se consideran clientes con importantes fluctuaciones de consumo en el tiempo.

- **Clúster 8**: Grupo *"clientes comerciales con consumo promedio normal"*. Poseen un consumo promedio y mínimo considerado normal para un cliente comercial, sumado a los pequeños valores en las variables ratio consumo y coeficiente de variación, que indican un consumo estable en el tiempo.

- **Clúster 9**: Grupo *"clientes comerciales con bajo consumo promedio y altas variaciones"*. Poseen un consumo promedio y mínimo bajo para cualquier tipo de cliente, esto sumado a un alto ratio consumo y coeficiente de variación, lo cual indicaría que dentro de este grupo existen clientes con un consumo bajo y muy variable en el tiempo.

Detección casos outliers o anómalos

Como ya fue mencionado, aquellos grupos que posean menos de 100 casos serán considerados como grupos anómalos a nivel global, ingresando en primer lugar al listado de casos *outliers.* Esta primera selección en el caso de los clientes comerciales, suma en total 18 casos. Para la detección de casos anómalos a nivel local (clúster), utilizando la metodología descrita en la sección 7.6.2, se obtiene un total de 103 casos, conformando una lista final de 121 casos *outliers* de tipo comercial.

[29] Cliente comercial consume mensualmente en promedio 350 KWh

Aplicación modelo regresión logística

Análogo al caso de los clientes residenciales, se aplica el modelo de regresión logística para obtener las probabilidades de fraude esperadas de cada clúster y las de cada cliente incluido en el listado total de *outliers.* La tabla 31 contiene el detalle de la tasa esperada de fraude para cada clúster de clientes comerciales junto a la tasa de fraude de sus respectivos *outliers* detectados tanto a nivel local como global.

Clúster	N° casos clúster	Tasa fraude esperada clúster	N° casos *outliers*	Tasa fraude esperada *outliers*
clúster 1	5623	1.79%	57	2.09%
clúster 2	1	0.90%	1	0.90%
clúster 3	2	5.16%	2	5.16%
clúster 4	2	4.91%	2	4.91%
clúster 5	585	1.44%	6	1.29%
clúster 6	5	4.32%	5	4.32%
clúster 7	1477	4.18%	14	4.80%
clúster 8	2336	1.50%	23	4.09%
clúster 9	288	5.56%	3	4.78%
clúster 10	2	1.85%	2	1.85%
clúster 11	3	4.14%	3	4.14%
clúster 12	3	5.33%	3	5.33%
TOTAL	**10327**	**2.16%**	**121**	**3.12%**

Tabla 31: Tasas esperadas de fraude clústeres y *outliers* – Comercial

Se observa en la tabla 31 que las tasas de fraude obtenidas desde el modelo de regresión coinciden con lo esperado indicando a los clústeres 7 y 9 como los de mayor probabilidad de fraude. El clúster 7 corresponde a clientes comerciales con consumo promedio bajo (130KWh) y un ratio consumo de 1.43 que indica variaciones en el tiempo. El clúster 9 incorpora aquellos clientes con un bajo consumo promedio y altas variaciones en el tiempo alcanzando un coeficiente de variación promedio de 1.8 y consumo promedio de apenas 50KWh. Con los antecedentes mencionados, es de esperar que ambos grupos posean una tasa de fraude esperada más alta que el resto ya que su comportamiento dista bastante de consumidor comercial catalogado como normal.

Evaluación de resultados

El resultado final, que incorpora los resultados para ambos tipos de clientes, consiste en un listado de 2854 casos *outliers,* entregado de manera complementaria al listado de inspección obtenido mediante el modelo de regresión logística.

Como medida de evaluación, se utilizó el modelo de regresión logística para obtener las probabilidades de fraude esperadas de cada clúster y las de cada cliente incluido en el listado total de *outliers.* Dentro de este último grupo se obtiene una tasa de fraude esperada de 3.12%, la cual supera en un 0.92% a la tasa promedio actual (2.2%), es decir, se esperaría un incremento de un 41% respecto a la situación actual.

De los resultados obtenidos es posible concluir que aun cuando el modelo realizado no es comparable a ningún modelo de aprendizaje supervisado, los resultados del modelo de regresión, para los 2854 casos (mismo número de casos del listado *outliers*) de mayor probabilidad de fraude alcanzan una tasa promedio de fraude de 13.69%, versus el 3.12% que se esperaría obtener inspeccionando el listado de casos *outliers*.

9. Evaluación económica

Antecedentes

La normativa vigente [1] por parte de la superintendencia de electricidad y combustibles (en adelante SEC) establece que una vez que la empresa eléctrica haya comprobado el hurto de energía por parte de un cliente queda autorizada a cobrar el CNR retroactivamente, previo autorización de la SEC, el tiempo por el cual haya existido la irregularidad con un periodo máximo cobro de 12 meses, debido a que se considera obligación de la empresa mantener un control permanente del consumo de energía de sus clientes.

Ingresos por cobro retroactivo en caso fraude (consumo no registrado)

Como se mencionó en el párrafo anterior, la empresa está autorizada a cobrar retroactivamente un periodo máximo de 12 meses de consumos no registrados, por lo cual para el presente proyecto y dado que se consideran variables semestrales, se considerará como periodo promedio de recuperación un plazo de 6 meses pero se realizará también un análisis de sensibilidad al fijar otros plazos.

Para efectos del presente trabajo se calculará la diferencia entre el consumo promedio del semestre previo al descubrimiento del fraude del cliente, debido a que se tomara como plazo promedio un periodo de 6 meses, versus el consumo promedio del trimestre posterior a la regularización del suministro, plazo en el cual el suministro debe encontrarse totalmente regularizado.

De acuerdo a información oficial de la empresa en estudio, la tarifa del suministro eléctrico se descompone en dos items: cargo fijo ($/mes) y energía base ($/KWh). De acuerdo a las tarifas de estos ítems 1 KWh de energía base, tanto en el sector residencial como comercial, equivale en promedio a $118.

La tabla 32 muestra el consumo promedio mensual de los clientes que fueron descubiertos como fraude durante el período enero-13/ diciembre-13. Se aprecia que la diferencia entre el consumo promedio del semestre previo al descubrimiento de fraude, versus el trimestre posterior alcanza 116 KWh lo cual se traduce en que la recuperación promedio **mensual** por descubrir un consumo no registrado alcanzaría los **$13069**.

Rubro	Consumopromedio 6m_pre (KWh)	Consumopromedio 3m_post (KWh)	Delta Consumo	% de total fraude
Comercial	450.97	742.08	291.11	5%
Residencial	110.16	217.16	107.00	95%
Total general	**127.20**	**243.41**	**116.20**	100.0%

Tabla 32: Consumo previo y posterior a descubrimiento fraude

Capacidad y costos de Inspección

Debido a la normativa vigente y a que el hurto de energía incurre en pérdidas para la empresa, esta inspecciona mensualmente[30] a un grupo de clientes con el objetivo de verificar la presencia de consumos no registrados de energía.

La capacidad operacional de inspección en toda la región alcanza las 7800 inspecciones mensuales, según información correspondiente al año 2013 como se detalla en la siguiente tabla:

Personal operativo	Actividad	Personal 2013	Cuadrillas 2013	Actividad/día	#/día	#/mes
Inspectores perdidas	Inspección	26	26	15	390	**7800**

Tabla 33: Capacidad de inspección mensual

El costo de inspección, según fuentes de la propia empresa, asciende a **$5,820 por cada inspección.**

Cálculo de beneficios

El objetivo del cálculo de beneficios es obtener la tasa de fraude esperada que, dada una cierta cantidad de inspecciones, maximiza las ganancias para la empresa, obteniendo como resultado un *criterio de inspección,* tal como se especifica en la siguiente frase: *"inspeccionar todos aquellos clientes con una probabilidad de fraude mayor o igual a X%".*

Los beneficios obtenidos se calculan como ingresos – costos, donde los ingresos corresponden a la cantidad esperada de fraudes capturados (tasa fraude) por el monto promedio de recuperación. Al mismo tiempo, los costos se componen como la cantidad de inspecciones realizadas por el costo unitario de estas. Esto puede verse resumido en la formula a continuación:

$$\boldsymbol{Beneficio}(\$) = \left(\#\boldsymbol{inspeccionados} * \boldsymbol{tasa\ fraude\ esperada} * \$\frac{\boldsymbol{fraude}}{\boldsymbol{mes}} * \#\boldsymbol{meses\ cobro\ retroactivo}\right) - (\#\boldsymbol{inspeccionados} * \boldsymbol{costo\ unitario\ inspección})$$

De acuerdo a los resultados obtenidos en el capítulo anterior, el análisis económico se realiza utilizando las probabilidades esperadas de fraude obtenidas desde el modelo de regresión logística.

[30] Ver sección 5 Análisis descriptivo - inspecciones

Situación actual y punto óptimo de inspección

Actualmente la empresa declara tener una capacidad máxima de inspección de 7800 inspecciones mensuales. La tasa de fraude real existente corresponde a un 2.2% por lo cual, bajo el supuesto que se realizasen 7800 inspecciones se estarían generando pérdidas mensuales por más de $27 millones de pesos[31].

Al aplicar el modelo de regresión logística a la base de datos de clientes no inspeccionados se obtiene la probabilidad de fraude de cada cliente, obteniendo hasta qué punto se debe inspeccionar (número de clientes) de modo tal de maximizar el beneficio.

Bajo el supuesto de que el plazo promedio de cobro retroactivo es un periodo de 6 meses, por cada fraude descubierto se cobraría a cada cliente la suma de $78414($13,069x6) y el punto de máximo beneficio (figura 38) se alcanzaría inspeccionando los primeros 4938 clientes con mayor probabilidad de fraude, o análogamente, todos aquellos clientes con una probabilidad de fraude mayor o igual a 0.0674 (6.74%). La tasa esperada de fraude asociada a este punto es de un 8.67%, es decir, al inspeccionar estos 4.938 clientes en promedio se obtiene una tasa de fraude de 8.67%. Los beneficios en este punto ascienden a $7,484,416, mientras que si se quisiera utilizar la capacidad de inspección hasta su máximo (7800 inspecciones), se esperaría obtener beneficios por $6,406,292. Si bien al realizar la máxima cantidad de inspecciones no se generan pérdidas, se recomienda detenerse en el punto óptimo antes mencionado.

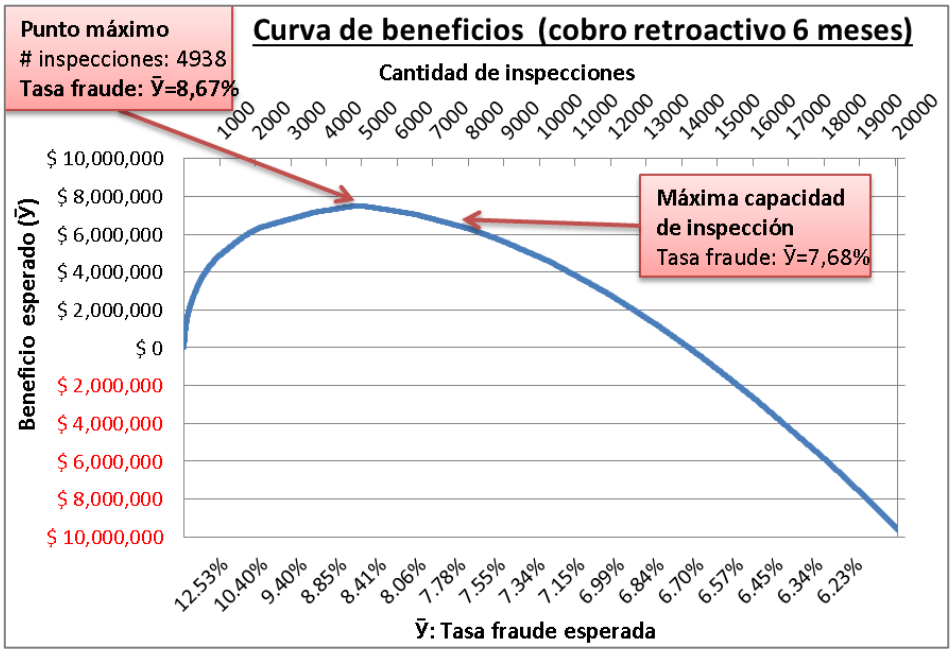

Figura 38: Gráfica curva de beneficio según cobro retroactivo de 6 meses

[31] Cálculo realizado en base a la fórmula de beneficios

Análisis de sensibilidad

En el análisis realizado previamente se estableció como tiempo promedio de cobro retroactivo un periodo de 6 meses por lo que para efectos de evaluar hasta que plazo aún se obtendrían beneficios se realiza un análisis de sensibilidad sujeto a la variable *tiempo de cobro retroactivo.* Como es de esperar, según se observa en la figura 39, mientras menor es el tiempo de cobro menor es el beneficio obtenido.

Se observa que para 1 mes de cobro ya se obtienen resultados positivos sin embargo el máximo beneficio alcanza los $7,095 al realizar solo 12 inspecciones lo cual carece de sentido.

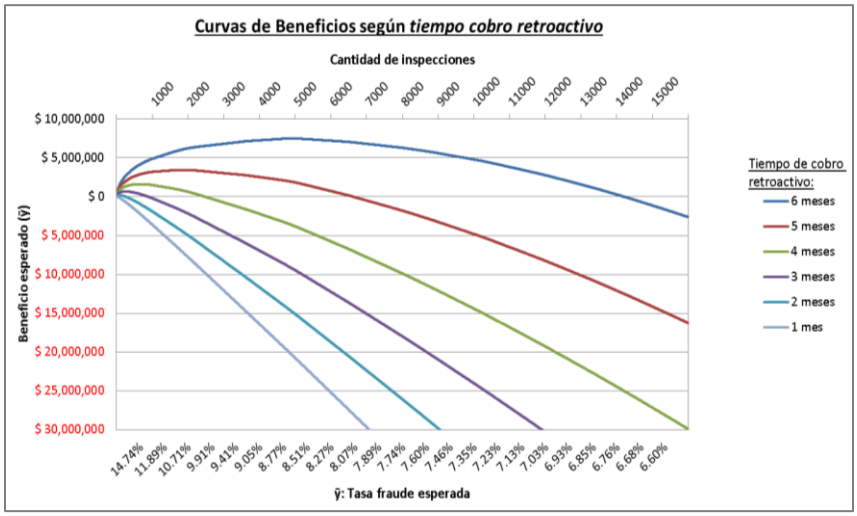

Figura 39: Curvas de beneficio según tiempo de cobro retroactivo (1-6 meses)

De forma análoga, al aumentar el plazo de recuperación los beneficios (figura 40) comienzan a incrementarse, pero también la cantidad de inspecciones a realizar para alcanzar dicho punto. En el mejor de los casos, asociado al cobro retroactivo por un periodo de 12 meses, es posible alcanzar un beneficio máximo de $ 99,376,805 realizando 37573 inspecciones con una tasa esperada de fraude de 5.05%. La tabla adjunta en el anexo H detalla los puntos de máximo beneficio asociados a cada plazo de recuperación considerado.

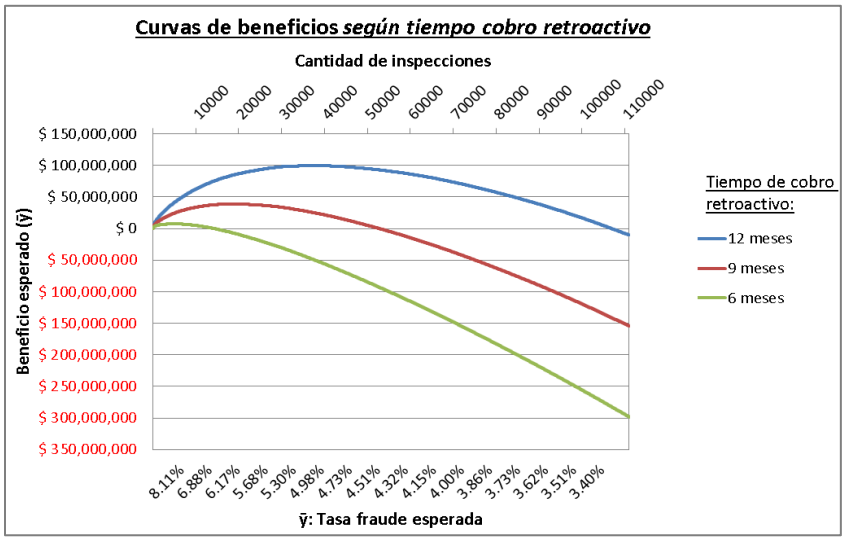

Figura 40: Curvas de beneficio según tiempo de cobro retroactivo (6,9,12 meses)

De los análisis realizados se concluye que utilizando la máxima capacidad de inspección mensual solo se obtendrían beneficios al considerar un periodo de cobro retroactivo desde 6 meses, ya que para un plazo menor a este la empresa estaría incurriendo en pérdidas, es decir, la recuperación monetaria del fraude descubierto no estaría cubriendo los costos de inspección.

Además, como es de esperar, un mayor beneficio se encuentra asociado a una mayor tasa de fraude obtenida, por lo cual si la empresa quisiera estar protegida frente a posibles pérdidas por concepto de inspección, es preferible realizar una menor cantidad de estas, no alcanzando la máxima capacidad, ya que el solo hecho de realizar 7800 inspecciones implica un costo fijo mensual de $45 millones de pesos, es decir, para lograr cubrir estos gastos, considerando un periodo de recuperación de seis meses, al menos debiesen ser descubiertos 579 fraudes lo que es equivalente a obtener una tasa de un 7.4%.

10. Conclusiones

10.1 Conclusiones del trabajo realizado

Las conclusiones generales del trabajo se enmarcan en los resultados obtenidos en los modelos predictivos construidos. El modelo de regresión logística logró capturar un 35.12% del fraude en el primer decil de los clientes de la data de prueba, lo cual es un resultado positivo si se compara a un modelo aleatorio en que al extraer un 10% de los clientes, en promedio se debiese capturar la misma proporción de fraude, es decir, un 10%. En cuanto a la interpretación de los resultados el modelo mostró ser el más adecuado, lo cual está directamente relacionado a uno de los objetivos impuestos en el trabajo. De dicha interpretación se desprende que aquellos clientes que no mantienen un historial de consumo parejo en el tiempo son más propensos a estar hurtando energía, mientras que aquellos clientes con más de un corte de suministro en un periodo de seis meses, incrementan su propensión al fraude, siendo esperables ambos resultados. Además cabe destacar la significancia de la variable zona, lo cual se explicaría porque las inspecciones se realizan en distinta cantidad y metodología, existiendo, por ejemplo, en las zonas 1 y 2 inspecciones masivas o “de barrido” que implicarían menores tasas de fraudes. Además es importante mencionar que las variables generadas a partir de claves de lectura aportan información, por lo cual se recomendaría a la empresa enfatizar con sus trabajadores la tarea de registro de estas marcas.

Del segundo modelo correspondiente a una regresión logística ponderada, se pudo comprobar que la data desbalanceada del problema no afecta los resultados obtenidos, ya que para el primer decil de clientes el modelo que incorpora un peso mayor a la variable fraude captura un 33.93% de este, versus un 35.12% capturado por el modelo sin ponderaciones. Además se aprecia la similitud en los coeficientes de ambas regresiones a excepción del intercepto, el cual marca la diferencia de magnitud en las probabilidades obtenidas para ambos modelos.

Respecto al modelo Random Forest este presenta los mejores resultados en cuanto a clasificación, logrando capturar para el primer decil de clientes un 39.88% del fraude, superando al modelo de regresión en 4.8%, sin embargo el principal inconveniente asociado al modelo radica en un extenso tiempo de ejecución (más de 25 horas) el cual no compensa la ganancia antes mencionada, por lo que el modelo de regresión logística es escogido como el mejor candidato.

Desde el análisis económico se desprende que actualmente la empresa no se encuentra en su óptimo de inspección, ya que con la tasa actual de fraude (2.2%) estarían generando, en promedio, pérdidas por más de 27 millones de pesos mensuales. De acuerdo al modelo de regresión construido y bajo el supuesto de que se recuperasen en promedio 6 meses de consumos no registrados, debiesen inspeccionarse los 4938 clientes con mayores probabilidades de fraude obteniendo en promedio una tasa de un 8.67%, superior al 2.2% actual.

Además se comprobó empíricamente que el modelo Random Forest entrega mejores resultados de clasificación, mediante la incorporación de aleatoriedad en la construcción de cada árbol individual, entregando los mejores resultados de clasificación entre los modelos estudiados, aun cuando operacionalmente resulta ser el más deficiente. Si esto no fuera un problema para una compañía sería recomendable su utilización en problemas de similares características.

Respecto al modelo complementario de clustering cabe destacar que si bien el modelo no es comparable a los resultados obtenidos en los modelos de clasificación, al aplicar el modelo de regresión al listado obtenido, se obtiene una tasa promedio de fraude de un 3.12%, la cual si bien es mayor a la tasa promedio existente, es ampliamente superada por la tasa esperada de fraude obtenida en el modelo de regresión, que alcanza un 13.69% en los 2857 casos (número de casos del listado *outliers*) de mayor probabilidad de fraude. Por lo anterior, se concluye que el modelo supervisado entregaría mejores resultados y el modelo de *clustering* solo debe considerarse de manera opcional y complementaria.

A modo de resumen, se concluye que los objetivos planteados al comienzo del trabajo fueron realizados en su totalidad, obteniendo modelos predictivos con resultados satisfactorios junto al nivel de importancia y el procedimiento de construcción de las variables relevantes en la identificación de un fraude.

10.2 Limitaciones del trabajo

Dentro de las limitaciones del trabajo se destaca la dificultad de la extracción de información por parte de la base de datos de SERVEL, lo cual tenía como objetivo extraer el número de personas asociadas a una misma dirección a fin de obtener una aproximación de la densidad poblacional por hogar. Esta labor no fue concluida debido al gran tamaño y formato del archivo (3gb, formato XML), y que posteriormente fue imposible de cruzar dicha información con la base de clientes puesta a disposición por parte de la empresa. Un problema similar se tuvo con la base de datos proveniente de la encuesta CASEN la cual tenía como objetivo inicial obtener información sociodemográfica de los clientes, sin embargo no se tuvo éxito debido a que los datos disponibles de esta fuente se agrupan por muestras de población, información que no pudo ser asociada a los clientes estudiados.

Una segunda limitación del trabajo recae en la pequeña cantidad de fraudes disponibles para el modelamiento, lo cual no tuvo complicaciones al construir un modelo a nivel global, pero si a nivel local, ya que al intentar realizar un modelo a nivel intra-zona se obtuvo un modelo no significativo debido a la baja cantidad de fraudes a nivel desagregado.

10.3 Recomendaciones y trabajos futuros

Uno de los objetivos de futuros trabajos debiese apuntar a indagar más en información geográfica de los clientes. Esto se ve avalado en los resultados los cuales entregan como variable relevante la zona a la cual un cliente pertenece, por lo que sería interesante esclarecer las verdaderas hipótesis de que esto ocurra, incorporando por ejemplo, a nivel más desagregado información demográfica, como densidad de población del sector o información socioeconómica como tipos de barrios (vulnerables, con mayor poder adquisitivo, etc.) que podrían explicar parte del efecto zona identificado.

Además, como se mencionó en la sección previa de limitaciones del trabajo, no fue posible estudiar el comportamiento de los clientes a nivel intra-zona, debido al bajo número de fraudes existentes a este nivel más desagregado resultando imposible establecer patrones verdaderamente significativos. Como futura línea de investigación se plantea un estudio de manera particular a cada zona debido a que a nivel global pueden existir factores que no están considerados en la actualidad, por ejemplo, clientes pertenecientes a zona 1, que posee una población flotante mucho mayor en época estival, pueden tener un comportamiento distinto a clientes situados en la zona 5 donde no ocurre este fenómeno.

Otra información que podría haber enriquecido el trabajo hubiese sido tener una mayor cantidad de información de los clientes, a modo de ejemplo si se tuviese información del número de personas que habitan una vivienda o información de otros servicios básicos como el agua, sería posible realizar estimaciones de consumo mensual.

Respecto al análisis económico realizado se observa que el beneficio monetario, y por ende que tan rentable es para la empresa inspeccionar a sus clientes, es muy sensible al monto y a la cantidad de meses de cobro retroactivo. Por este motivo un aporte valioso al trabajo realizado sería perfeccionar la estimación del monto puede cobrarse a cada cliente en caso de descubrir fraude y en la misma línea la cantidad de meses que el cliente ha estado hurtando energía.

Finalmente, como se mencionó en la primera parte de las conclusiones, la información derivada de las claves de lectura, recabada en terreno por personal de la empresa, es de vital importancia en el modelo por lo cual, se recomienda a la empresa enfatizar en esta tarea incorporando por ejemplo, incentivos al personal de modo de registrar en mayor número las denominadas marcas o claves de lectura.

11. Bibliografía

[1]. Superintendencia de electricidad y combustibles «Legislación aplicada a instalaciones eléctricas fraudulentas, por la superintendencia de electricidad y combustibles,»http://www.sec.cl/pls/portal/docs/PAGE/SECNORMATIVA/ELECTRICIDAD_OFICIOS/OC_7278_03.PDF

[2]. Superintendencia de electricidad y combustibles, «FIJA REGLAMENTO DE LA LEY GENERAL DE SERVICIOS ELECTRICOS,» http://www.sec.cl/pls/portal/docs/PAGE/SECNORMATIVA/ELECTRICIDAD_DECRETOS/DECRETO327.PDF

[3]. C. Phua, V. Lee, K. Smith, R. Gayler, «A Comprehensive Survey of Data Mining-based Fraud Detection Research»

[4]. J. Han, M. Kamber: Data Mining: Concepts and Techniques. Morgan Kaufmann, 2a edition, 2006.

[5]. J.Akthar, C. Hahne: RapidMiner 5, Operator Reference, 2012

[6]. F. Morales, «SISTEMA PREDICTIVO PARA LA ASIGNACIÓN DE BÚSQUEDA DE HURTOS Y FRAUDES EN UNA EMPRESA ELÉCTRICA,» Memoria para optar al título de ingeniero civil industrial, Santiago, 2013.

[7]. D. Morales, «DISEÑO DE UN DATA MART Y APLICACIÓN DE DATA MINING EN LA DETECCION DE FRAUDES PARA LA COMPAÑÍA GENERAL DE ELECTRICIDAD S.A.,» Memoria para optar al título de ingeniero civil industrial,Santiago, 2002.

[8]. Dr. J. Jebamalar, «Assessment of Fraud Pretentious Business Region Research Articles Using Data Mining Approaches», 2013

[9]. A. Kokkinaki, «On Atypical Database Transactions: Identification of Probable Frauds using Machine Learning for User Profiling» Proc. Of IEEE Knowledge and Data Engineering Exchange Workshop, 107-112, (1997).

[10]. J. Burez, D. Van den Poel «Handing class inbalance in customer churn prediction» Expert systems with applications 36, 4626-4636, 2009

[11]. S.Maldonado, «Apuntes DIP78J: Diplomado Business Intelligence (diapositivas), » FCFM. Departamento de Ingeniería Civil Industrial, Universidad de Chile, 2012

[12]. F. Barrientos, «DISEÑO E IMPLEMENTACIÓN DE UNA METODOLOGÍA DE PREDICCIÓN DE FUGA DE CLIENTES EN UNA COMPAÑÍA DE TELECOMUNICACIONES,» Memoria para optar al título de ingeniero civil industrial, Santiago, 2011.

[13]. Introducción al modelo de regresión logística
http://www.seqc.es/es/Varios/7/40/Modulo_3:_Regresion_logistica_y_multiple/

[14]. J. Sempere, «Aprendizaje de árboles de decisión, » Universidad Politécnica de Valencia, Valencia.

[15]. D. Riano, «Árboles de decisión id3-c4.5»
http://banzai-deim.urv.net/~riano/teaching/id3-m5.pdf

[16]. L. Breiman «Random Forests. Machine Learning,»2001.

[17]. A. Liaw, M. Wiener, «Classification and Regression by Random Forest,» 2002.

[18]. T. Khoshgoftaar, M. Golawala, J.Van Hulse, «An Empirical Study of Learning from Imbalanced Data Using Random Forest,» 19th IEEE International Conference on Tools with Artificial Intelligence.

[19]. U. Murad, G.Pinkas, «Unsupervised Profiling For Identifying Superimposed Fraud,»1999.

[20]. Xu, R. y Wunsch II, D. C.: «Survey of clustering algorithms,». Neural Networks, IEEE Transactions on, 2005, 16(3), pp. 645–678.

[21]. M. Goldstein, M. Amer, «Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner,»2012

12. Anexos

Anexo A. Ranking de variables mediante árbol de decisión

Importancia de las variables independientes

Variable independiente	Importancia	Importancia normalizada
falla_medidor_CL_7-12	,000	100,0%
Cmin7-12	,000	67,0%
prom_C7-12	,000	61,3%
Ratio_C6m	,000	60,1%
coef_varicion_C7-12	,000	59,3%
suma_mesesceroconsumo_7-12	,000	51,0%
zona	,000	50,9%
Cmax7-12	,000	37,8%
corte7-12	,000	36,5%
var%_Factura	,000	35,5%
prom_factur_7-12	,000	29,2%
coef_variacion_C1-6	9,951E-5	25,9%
prom_C1-6	9,696E-5	25,2%
Cmin1-6	9,301E-5	24,2%
Cmax1-6	7,142E-5	18,6%
sospecha_fraude_CL_7-12	6,480E-5	16,9%
suma_mesesceroconsumo_1-6	6,372E-5	16,6%
prom_saldo_7-12	6,319E-5	16,5%
prom_factur_1-6	5,727E-5	14,9%
Prom_ECM_mmovil_consumo	5,558E-5	14,5%
sospecha_fraude_CL_1-6	4,501E-5	11,7%
Cmin_dif	4,098E-5	10,7%
prom_saldo_1-6	2,881E-5	7,5%
falla_medidor_CL_1-6	2,264E-5	5,9%
lectura_estimada_CL_7-12	1,140E-5	3,0%
Cmax_dif	1,074E-5	2,8%
consumo_atipico_CL_7-12	9,688E-6	2,5%
consumo_atipico_CL_1-6	1,519E-6	,4%
pearson_Cprom_zona-rubro	2,035E-7	,1%

Métodos de crecimiento: CRT

Variable dependiente: Fraude

Tabla 34: Ranking importancia de variables

Anexo B. Tablas de contingencia con probabilidades condicionales

coef_varicion_C7-12	0	1	Total general	PR(fraude=1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 0.07]	7545	71	7616	0.009	0.423	-0.860	0.860	10%
[0.07 - 0.09]	7538	78	7616	0.010	0.465	-0.766	0.766	10%
[0.09 - 0.11]	7546	70	7616	0.009	0.417	-0.874	0.874	10%
[0.11 - 0.13]	7520	96	7616	0.013	0.572	-0.558	0.558	10%
[0.13 - 0.16]	7515	101	7616	0.013	0.602	-0.507	0.507	10%
[0.16 - 0.21]	7460	156	7616	0.020	0.930	-0.073	0.073	10%
[0.21 - 0.31]	7406	210	7616	0.028	1.252	0.225	0.225	10%
[0.31 - 0.49]	7381	235	7616	0.031	1.401	0.337	0.337	10%
[0.49 - 0.81]	7305	311	7616	0.041	1.854	0.617	0.617	10%
[0.81 - ?]	7194	301	7495	0.040	1.823	0.601	0.601	10%
Total general	74410	1675	76039	0.022	1.000			54%

Figura 41: Variable Coeficiente variación Consumo 7-12

Cmin_7-12	0	1	Total general	PR(fraude=1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 0.0]	9382	455	9837	0.046	2.100	0.742	0.742	13%
[0.0 - 17.5]	5047	209	5256	0.040	1.805	0.591	0.591	7%
[17.5 - 44.5]	7431	251	7682	0.033	1.483	0.394	0.394	10%
[44.5 - 66.5]	7351	182	7533	0.024	1.097	0.092	0.092	10%
[66.5 - 87.5]	7482	123	7605	0.016	0.734	-0.309	0.309	10%
[87.5 - 108.5]	7519	92	7611	0.012	0.549	-0.600	0.600	10%
[108.5 - 132.5	7506	83	7589	0.011	0.496	-0.700	0.700	10%
[132.5 - 166.5	7733	75	7808	0.010	0.436	-0.830	0.830	10%
[166.5 - 232.5	7543	77	7620	0.010	0.459	-0.779	0.779	10%
[232.5 - ?]	7370	128	7498	0.017	0.775	-0.255	0.255	10%
Total genera	74364	1675	76039	0.022	1.000			54%

Figura 42: Variable Consumo mínimo 7-12

coef_varicion_C1-6	0	1	Total general	PR(fraude=1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 0.07]	7512	92	7604	0.012	0.549	-0.599	0.599	10%
[0.07 - 0.08]	7521	81	7602	0.011	0.484	-0.726	0.726	10%
[0.08 - 0.10]	7512	93	7605	0.012	0.555	-0.589	0.589	10%
[0.10 - 0.12]	7500	104	7604	0.014	0.621	-0.477	0.477	10%
[0.12 - 0.15]	7495	109	7604	0.014	0.651	-0.430	0.430	10%
[0.15 - 0.19]	7456	148	7604	0.019	0.884	-0.124	0.124	10%
[0.19 - 0.27]	7410	194	7604	0.026	1.158	0.147	0.147	10%
[0.27 - 0.44]	7351	253	7604	0.033	1.510	0.412	0.412	10%
[0.44 - 0.74]	7283	321	7604	0.042	1.916	0.650	0.650	10%
[0.74 - ?]	7324	280	7604	0.037	1.672	0.514	0.514	10%
Total genera	74364	1675	76039	0.022	1.000			47%

Tabla 35: Variable coeficiente variación C_1-6

Ratio_C6m	0	1	Total general	PR(fraude= 1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 0.58]	7206	398	7604	0.052	2.376	0.865	0.865	10%
[0.58 - 0.79]	7376	228	7604	0.030	1.361	0.308	0.308	10%
[0.79 - 0.88]	7451	153	7604	0.020	0.913	-0.091	0.091	10%
[0.88 - 0.94]	7480	123	7603	0.016	0.734	-0.309	0.309	10%
[0.94 - 0.99]	7508	97	7605	0.013	0.579	-0.546	0.546	10%
[0.99 - 1.04]	7515	89	7604	0.012	0.531	-0.632	0.632	10%
[1.04 - 1.11]	7505	99	7604	0.013	0.591	-0.526	0.526	10%
[1.11 - 1.21]	7478	125	7603	0.016	0.746	-0.293	0.293	10%
[1.21 - 1.43]	7434	170	7604	0.022	1.015	0.015	0.015	10%
[1.43 - ?]	7411	193	7604	0.025	1.152	0.142	0.142	10%
Total general	74364	1675	76039	0.022	1			37%

Tabla 36: Variable Ratio consumo 6m

Cmax_dif	0	1	Total general	PR(fraude= 1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - -0.44]	7267	349	7616	0.046	2.080	0.732	0.732	10%
[-0.44 - -0.23	7385	231	7616	0.030	1.377	0.320	0.320	10%
[-0.23 - -0.13	7463	146	7609	0.019	0.871	-0.138	0.138	10%
[-0.13 - -0.06	7484	121	7605	0.016	0.722	-0.325	0.325	10%
[-0.00 - 0.06]	7599	97	7696	0.013	0.572	-0.558	0.558	10%
[-0.06 - -0.00	7397	115	7512	0.015	0.695	-0.364	0.364	10%
[0.06 - 0.13]	7534	121	7655	0.016	0.718	-0.332	0.332	10%
[0.13 - 0.25]	7511	104	7615	0.014	0.620	-0.478	0.478	10%
[0.25 - 0.56]	7444	175	7619	0.023	1.043	0.042	0.042	10%
[0.56 - ?]	7280	216	7496	0.029	1.308	0.269	0.269	10%
Total genera	74364	1675	76039	0.022	1.000			36%

Tabla 37: Variable diferencia entre consumos máximos

Cmin_1-6	0	1	Total general	PR(fraude= 1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 0.0]	7606	290	7896	0.037	1.667	0.511	0.511	10%
[0.0 - 25.5]	7036	284	7320	0.039	1.761	0.566	0.566	10%
[25.5 - 50.5]	7234	211	7445	0.028	1.287	0.252	0.252	10%
[50.5 - 71.5]	7376	173	7549	0.023	1.040	0.040	0.040	10%
[71.5 - 91.5]	7523	151	7674	0.020	0.893	-0.113	0.113	10%
[91.5 - 112.5]	7585	107	7692	0.014	0.631	-0.460	0.460	10%
[112.5 - 136.5	7435	110	7545	0.015	0.662	-0.413	0.413	10%
[136.5 - 170.5	7642	92	7734	0.012	0.540	-0.616	0.616	10%
[170.5 - 236.5	7566	88	7654	0.011	0.522	-0.650	0.650	10%
[236.5 - ?]	7361	169	7530	0.022	1.019	0.019	0.019	10%
Total genera	74364	1675	76039	0.022	1.000			36%

Tabla 38: Variable Consumo mínimo 1-6

Prom_C7_12	0	1	Total general	PR(fraude=1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 32.6]	7259	325	7584	0.043	1.945	0.665	0.665	10%
[32.6 - 61.4]	7319	276	7595	0.036	1.650	0.501	0.501	10%
[61.4 - 83.6]	7443	178	7621	0.023	1.060	0.059	0.059	10%
[83.6 - 104.2]	7412	159	7571	0.021	0.953	-0.048	0.048	10%
[104.2 - 124.8	7497	130	7627	0.017	0.774	-0.256	0.256	10%
[124.8 - 147.6	7468	117	7585	0.015	0.700	-0.356	0.356	10%
[147.6 - 175.1	7493	112	7605	0.015	0.669	-0.403	0.403	10%
[175.1 - 215.4	7515	107	7622	0.014	0.637	-0.451	0.451	10%
[215.4 - 300.9	7513	110	7623	0.014	0.655	-0.423	0.423	10%
[300.9 - ?]	7445	161	7606	0.021	0.961	-0.040	0.040	10%
Total genera	74364	1675	76039	0.022	1.000			32%

Tabla 39: Variable Consumo promedio 7-12

Var%_factura	0	1	Total general	PR(fraude=1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - -0.7]	7254	362	7616	0.048	2.158	0.769	0.769	10%
[-0.7 - -0.5]	7442	174	7616	0.023	1.037	0.036	0.036	10%
[-0.5 - -0.3]	7423	193	7616	0.025	1.150	0.140	0.140	10%
[-0.3 - -0.2]	7483	133	7616	0.017	0.793	-0.232	0.232	10%
[-0.2 - -0.1]	7495	121	7616	0.016	0.721	-0.327	0.327	10%
[-0.1 - 0.0]	7505	111	7616	0.015	0.662	-0.413	0.413	10%
[0.0 - 0.2]	7519	97	7616	0.013	0.578	-0.548	0.548	10%
[0.2 - 0.4]	7490	126	7616	0.017	0.751	-0.286	0.286	10%
[0.4 - 1.2]	7469	147	7616	0.019	0.876	-0.132	0.132	10%
[1.2 - ?]	7284	211	7495	0.028	1.278	0.245	0.245	10%
Total genera	74364	1675	76039	0.022	1.000			31%

Tabla 40: Variable Variación % facturación

Zona	0	1	Total general	PR(fraude=1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
1	18282	331	18613	0.018	0.807	-0.214	0.214	24%
2	17417	285	17702	0.016	0.731	-0.314	0.314	23%
3	7430	178	7608	0.023	1.062	0.060	0.060	10%
4	15157	288	15445	0.019	0.846	-0.167	0.167	20%
5	9565	375	9940	0.038	1.713	0.538	0.538	13%
6	6513	218	6731	0.032	1.470	0.385	0.385	9%
Total genera	74364	1675	76039	0.022	1.000			27%

Tabla 41: Variable zona

Prom_ECM-mmovil_consumo	0	1	Total casos	PR(fraude=1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 88.5]	7433	171	7604	0.022	1.021	0.021	0.021	10%
[88.5 - 184.9]	7487	117	7604	0.015	0.698	-0.359	0.359	10%
[184.9 - 313.1]	7499	105	7604	0.014	0.627	-0.467	0.467	10%
[313.1 - 502.1]	7455	149	7604	0.020	0.890	-0.117	0.117	10%
[502.1 - 802.8]	7506	98	7604	0.013	0.585	-0.536	0.536	10%
[802.8 - 1321.5]	7461	143	7604	0.019	0.854	-0.158	0.158	10%
[1321.5 - 2357.6]	7429	175	7604	0.023	1.045	0.044	0.044	10%
[2357.6 - 4741.9]	7428	176	7604	0.023	1.051	0.049	0.049	10%
[4741.9 - 12151.7]	7342	262	7604	0.034	1.564	0.447	0.447	10%
[12151.7 - ?]	7324	279	7603	0.037	1.666	0.510	0.510	10%
Total general	74364	1675	76039	0.022	1.000			27%

Tabla 42: Variable promedio ECM media móvil 6m

Prom_Factura_7-12	0	1	Total general	PR(fraude=1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 5231.4]	7329	275	7604	0.036	1.642	0.496	0.496	10%
[5231.4 - 842	7385	219	7604	0.029	1.307	0.268	0.268	10%
[8429.8 - 111	7458	146	7604	0.019	0.872	-0.137	0.137	10%
[11175.4 - 13	7461	143	7604	0.019	0.854	-0.158	0.158	10%
[13917.9 - 17	7490	114	7604	0.015	0.681	-0.385	0.385	10%
[17166.5 - 21	7465	139	7604	0.018	0.830	-0.187	0.187	10%
[21350.2 - 27	7481	122	7603	0.016	0.728	-0.317	0.317	10%
[27193.8 - 36	7464	140	7604	0.018	0.836	-0.179	0.179	10%
[36287.8 - 53	7455	149	7604	0.020	0.890	-0.117	0.117	10%
[53642.0 - ?]	7376	228	7604	0.030	1.361	0.308	0.308	10%
Total genera	74364	1675	76039	0.022	1.000			26%

Tabla 43: Variable Facturación promedio 7-12

Cmax_7-12	0	1	Total general	PR(fraude=1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 54.5]	7263	325	7588	0.043	1.944	0.665	0.665	10%
[54.5 - 86.5]	7194	178	7372	0.024	1.096	0.092	0.092	10%
[86.5 - 112.5]	7578	188	7766	0.024	1.099	0.094	0.094	10%
[112.5 - 137.5	7558	131	7689	0.017	0.773	-0.257	0.257	10%
[137.5 - 161.5	7279	129	7408	0.017	0.791	-0.235	0.235	10%
[161.5 - 191.5	7653	121	7774	0.016	0.707	-0.347	0.347	10%
[191.5 - 228.5	7556	104	7660	0.014	0.616	-0.484	0.484	10%
[228.5 - 283.5	7459	141	7600	0.019	0.842	-0.172	0.172	10%
[283.5 - 414.5	7523	161	7684	0.021	0.951	-0.050	0.050	10%
[414.5 - ?]	7301	197	7498	0.026	1.193	0.176	0.176	10%
Total genera	74364	1675	76039	0.022	1.000			26%

Tabla 44: Variable Consumo máximo 7-12

Prom_C1_6	0	1	Total general	PR(fraude=1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 40.1]	7310	273	7583	0.036	1.634	0.491	0.491	10%
[40.1 - 65.2]	7382	228	7610	0.030	1.360	0.308	0.308	10%
[65.2 - 86.2]	7434	171	7605	0.022	1.021	0.021	0.021	10%
[86.2 - 105.9]	7465	145	7610	0.019	0.865	-0.145	0.145	10%
[105.9 - 126.1	7442	160	7602	0.021	0.955	-0.046	0.046	10%
[126.1 - 148.4	7452	149	7601	0.020	0.890	-0.117	0.117	10%
[148.4 - 176.1	7476	112	7588	0.015	0.670	-0.400	0.400	10%
[176.1 - 216.6	7518	112	7630	0.015	0.666	-0.406	0.406	10%
[216.6 - 301.9	7475	131	7606	0.017	0.782	-0.246	0.246	10%
[301.9 - ?]	7410	194	7604	0.026	1.158	0.147	0.147	10%
Total genera	74364	1675	76039	0.022	1.000			23%

Tabla 45: Variable consumo promedio 1-6

Cmax_1-6	0	1	Total general	PR(fraude= 1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 63.5]	7171	250	7421	0.034	1.529	0.425	0.425	10%
[63.5 - 92.5]	7498	191	7689	0.025	1.128	0.120	0.120	10%
[92.5 - 115.5]	7309	138	7447	0.019	0.841	-0.173	0.173	10%
[115.5 - 138.5	7655	138	7793	0.018	0.804	-0.218	0.218	10%
[138.5 - 162.5	7475	146	7621	0.019	0.870	-0.140	0.140	10%
[162.5 - 190.5	7363	138	7501	0.018	0.835	-0.180	0.180	10%
[190.5 - 225.5	7518	145	7663	0.019	0.859	-0.152	0.152	10%
[225.5 - 281.5	7650	117	7767	0.015	0.684	-0.380	0.380	10%
[281.5 - 405.5	7425	189	7614	0.025	1.127	0.119	0.119	10%
[405.5 - ?]	7300	223	7523	0.030	1.346	0.297	0.297	10%
Total genera	74364	1675	76039	0.022	1.000			**22%**

Tabla 46: Variable máximo consumo 1-6

Prom_factura_1-6	0	1	Total general	PR(fraude= 1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 5995.5]	7381	222	7603	0.029	1.326	0.282	0.282	10%
[5995.5 - 917	7414	191	7605	0.025	1.140	0.131	0.131	10%
[9176.4 - 119	7458	146	7604	0.019	0.872	-0.137	0.137	10%
[11947.1 - 14	7478	126	7604	0.017	0.752	-0.285	0.285	10%
[14797.3 - 18	7468	136	7604	0.018	0.812	-0.208	0.208	10%
[18237.6 - 22	7463	141	7604	0.019	0.842	-0.172	0.172	10%
[22656.4 - 28	7469	135	7604	0.018	0.806	-0.216	0.216	10%
[28897.8 - 38	7429	175	7604	0.023	1.045	0.044	0.044	10%
[38064.2 - 56	7452	152	7604	0.020	0.907	-0.097	0.097	10%
[56061.8 - ?]	7352	251	7603	0.033	1.499	0.405	0.405	10%
Total genera	74364	1675	76039	0.022	1.000			**20%**

Tabla 47: Variable Facturación promedio 1-6

Prom_saldo 7-12	0	1	Total general	PR(fraude= 1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - 4.8]	7277	135	7412	0.018	0.827	-0.190	0.190	10%
[4.8 - 2684.6]	7584	212	7796	0.027	1.234	0.211	0.211	10%
[2684.6 - 609	7395	208	7603	0.027	1.242	0.217	0.217	10%
[6097.2 - 973	7425	180	7605	0.024	1.074	0.072	0.072	10%
[9731.2 - 135	7448	156	7604	0.021	0.931	-0.071	0.071	10%
[13544.7 - 18	7456	148	7604	0.019	0.884	-0.124	0.124	10%
[18144.8 - 24	7473	131	7604	0.017	0.782	-0.246	0.246	10%
[24401.2 - 34	7464	140	7604	0.018	0.836	-0.179	0.179	10%
[34004.8 - 53	7462	141	7603	0.019	0.842	-0.172	0.172	10%
[53263.5 - ?]	7380	224	7604	0.029	1.337	0.291	0.291	10%
Total genera	74364	1675	76039	0.022	1.000			**18%**

Tabla 48: Variable Saldo promedio 7-12

pearson_Cprom_zona-rubro	0	1	Total general	PR(fraude=1/total)	ODDS	LN(ODDS)	|LN_ODDS|	Pr(total casos)
[-? - -0.4]	7439	165	7604	0.022	0.985	-0.015	0.015	10%
[-0.4 - -0.3]	7419	185	7604	0.024	1.104	0.099	0.099	10%
[-0.3 - -0.2]	7427	177	7604	0.023	1.057	0.055	0.055	10%
[-0.2 - -0.1]	7444	160	7604	0.021	0.955	-0.046	0.046	10%
[-0.1 - 0.0]	7429	175	7604	0.023	1.045	0.044	0.044	10%
[0.0 - 0.1]	7443	160	7603	0.021	0.955	-0.046	0.046	10%
[0.1 - 0.2]	7419	185	7604	0.024	1.104	0.099	0.099	10%
[0.2 - 0.3]	7437	167	7604	0.022	0.997	-0.003	0.003	10%
[0.3 - 0.5]	7439	165	7604	0.022	0.985	-0.015	0.015	10%
[0.5 - ?]	7468	136	7604	0.018	0.812	-0.208	0.208	10%
Total general	74364	1675	76039	0.022	1.000			**6%**

Tabla 49: Coeficiente correlación consumo 12 meses misma zona-rubro

Anexo C. Ganancia en modelo regresión logística mediante incorporación de variables independientes

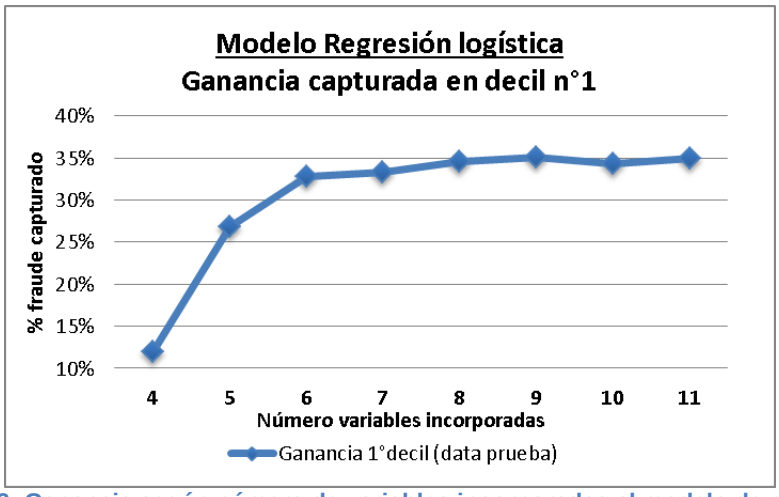

Figura 43: Ganancia según número de variables incorporadas al modelo de regresión

Anexo D. Codificación variables categóricas modelo regresión logística

Codificaciones de variables categóricas							
		Frecuencia	Codificación de parámetros				
			(1)	(2)	(3)	(4)	(5)
zona	5	9940	1,000	,000	,000	,000	,000
	3	7608	,000	1,000	,000	,000	,000
	4	15445	,000	,000	1,000	,000	,000
	6	6731	,000	,000	,000	1,000	,000
	2	17702	,000	,000	,000	,000	1,000
	1	18613	,000	,000	,000	,000	,000
coef_varicion_C7-12	range1	19009	,000	,000	,000		
	range2	19010	1,000	,000	,000		
	range3	19010	,000	1,000	,000		
	range4	19010	,000	,000	1,000		
Cmin_7-12	range1	19181	1,000	,000	,000		
	range2	19058	,000	1,000	,000		
	range3	18951	,000	,000	1,000		
	range4	18849	,000	,000	,000		
Ratio_C6m	range1	19009	1,000	,000	,000		
	range2	19010	,000	1,000	,000		
	range3	19010	,000	,000	1,000		
	range4	19010	,000	,000	,000		

Tabla 50: Codificación variables categóricas modelo regresión logística

Anexo E. Extracto Árbol de decisión ID3 (formato texto)

```
coef_varicion_C7-12 = range1 [-∞ - 0.096]
| zona = 2
| | sospecha_fraude_CL_7-12 = false
| | | Cmin7-12 = range1 [-∞ - 32.500]: 0
| | | Cmin7-12 = range2 [32.500 - 88.500]
| | | | Ratio_C6m = range1 [-∞ - 0.841]
| | | | | cortes_7-12 = false: 0
| | | | | cortes_7-12 = true: 0
| | | | Ratio_C6m = range2 [0.841 - 0.994]
| | | | | cortes_7-12 = false
| | | | | | consumo_atipico_CL_7-12 = false: 0
| | | | | | consumo_atipico_CL_7-12 = true: 0
| | | | | cortes_7-12 = true: 0
| | | | Ratio_C6m = range3 [0.994 - 1.149]
| | | | | consumo_atipico_CL_7-12 = false
| | | | | | cortes_7-12 = false: 0
| | | | | | cortes_7-12 = true: 0
| | | | | consumo_atipico_CL_7-12 = true: 0
| | | | Ratio_C6m = range4 [1.149 - ∞]
| | | | | cortes_7-12 = false
| | | | | | consumo_atipico_CL_7-12 = false: 0
| | | | | | consumo_atipico_CL_7-12 = true: 0
| | | | | cortes_7-12 = true: 0
| | | Cmin7-12 = range3 [88.500 - 148.500]
| | | | Ratio_C6m = range1 [-∞ - 0.841]
| | | | | cortes_7-12 = false
| | | | | | consumo_atipico_CL_7-12 = false: 0
| | | | | | consumo_atipico_CL_7-12 = true: 0
| | | | | cortes_7-12 = true: 0
| | | | Ratio_C6m = range2 [0.841 - 0.994]
| | | | | consumo_atipico_CL_7-12 = false
| | | | | | cortes_7-12 = false: 0
| | | | | | cortes_7-12 = true: 0
| | | | | consumo_atipico_CL_7-12 = true: 0
| | | | Ratio_C6m = range3 [0.994 - 1.149]
| | | | | consumo_atipico_CL_7-12 = false
```

| | | | | | cortes_7-12 = false
| | | | | | | sifalla_medidor7-12 = 0: 0
| | | | | | | sifalla_medidor7-12 = 1: 0
| | | | | | cortes_7-12 = true: 0
| | | | | consumo_atipico_CL_7-12 = true: 0
| | | | Ratio_C6m = range4 [1.149 - ∞]
| | | | | cortes_7-12 = false
| | | | | | consumo_atipico_CL_7-12 = false
| | | | | | | suma_mesesceroconsumo_7-12 = false: 0
| | | | | | | suma_mesesceroconsumo_7-12 = true: 0
| | | | | | consumo_atipico_CL_7-12 = true: 0
| | | | | cortes_7-12 = true: 0
| | | Cmin7-12 = range4 [148.500 - ∞]
| | | | Ratio_C6m = range1 [-∞ - 0.841]
| | | | | consumo_atipico_CL_7-12 = false: 0
| | | | | consumo_atipico_CL_7-12 = true: 0
| | | | Ratio_C6m = range2 [0.841 - 0.994]
| | | | | consumo_atipico_CL_7-12 = false
| | | | | | cortes_7-12 = false
| | | | | | | suma_mesesceroconsumo_7-12 = false: 0
| | | | | | | suma_mesesceroconsumo_7-12 = true: 0
| | | | | | cortes_7-12 = true: 0
| | | | | consumo_atipico_CL_7-12 = true: 0
| | | | Ratio_C6m = range3 [0.994 - 1.149]
| | | | | consumo_atipico_CL_7-12 = false
| | | | | | cortes_7-12 = false: 0
| | | | | | cortes_7-12 = true: 0
| | | | | consumo_atipico_CL_7-12 = true: 0
| | | | Ratio_C6m = range4 [1.149 - ∞]
| | | | | consumo_atipico_CL_7-12 = false
| | | | | | cortes_7-12 = false: 0
| | | | | | cortes_7-12 = true: 0
| | | | | consumo_atipico_CL_7-12 = true: 0
| | sospecha_fraude_CL_7-12 = true
| | | sifalla_medidor7-12 = 0
| | | | Cmin7-12 = range1 [-∞ - 32.500]: 0
| | | | Cmin7-12 = range2 [32.500 - 88.500]
| | | | | Ratio_C6m = range1 [-∞ - 0.841]: 0

```
|  |  |  |  |  Ratio_C6m = range2 [0.841 - 0.994]: 0
|  |  |  |  |  Ratio_C6m = range3 [0.994 - 1.149]
|  |  |  |  |  |  consumo_atipico_CL_7-12 = false: 0
|  |  |  |  |  |  consumo_atipico_CL_7-12 = true: 0
|  |  |  |  |  Ratio_C6m = range4 [1.149 - ∞]
|  |  |  |  |  |  consumo_atipico_CL_7-12 = false: 0
|  |  |  |  |  |  consumo_atipico_CL_7-12 = true: 0
|  |  |  Cmin7-12 = range3 [88.500 - 148.500]
|  |  |  |  |  Ratio_C6m = range1 [-∞ - 0.841]: 0
|  |  |  |  |  Ratio_C6m = range2 [0.841 - 0.994]
|  |  |  |  |  |  consumo_atipico_CL_7-12 = false: 0
|  |  |  |  |  |  consumo_atipico_CL_7-12 = true
|  |  |  |  |  |  |  cortes_7-12 = false: 0
|  |  |  |  |  |  |  cortes_7-12 = true: 0
|  |  |  |  |  Ratio_C6m = range3 [0.994 - 1.149]: 0
|  |  |  |  |  Ratio_C6m = range4 [1.149 - ∞]: 0
|  |  |  Cmin7-12 = range4 [148.500 - ∞]
|  |  |  |  |  Ratio_C6m = range1 [-∞ - 0.841]
|  |  |  |  |  |  consumo_atipico_CL_7-12 = false: 0
|  |  |  |  |  |  consumo_atipico_CL_7-12 = true: 0
|  |  |  |  |  Ratio_C6m = range2 [0.841 - 0.994]
|  |  |  |  |  |  consumo_atipico_CL_7-12 = false: 0
|  |  |  |  |  |  consumo_atipico_CL_7-12 = true: 0
|  |  |  |  |  Ratio_C6m = range3 [0.994 - 1.149]: 0
|  |  |  |  |  Ratio_C6m = range4 [1.149 - ∞]: 0
|  |  |  sifalla_medidor7-12 = 1
|  |  |  |  Ratio_C6m = range1 [-∞ - 0.841]: 1
|  |  |  |  Ratio_C6m = range2 [0.841 - 0.994]: null
|  |  |  |  Ratio_C6m = range3 [0.994 - 1.149]: 0
|  |  |  |  Ratio_C6m = range4 [1.149 - ∞]: null
|  zona = 1
...
```

Anexo F. Árbol de decisión C4.5

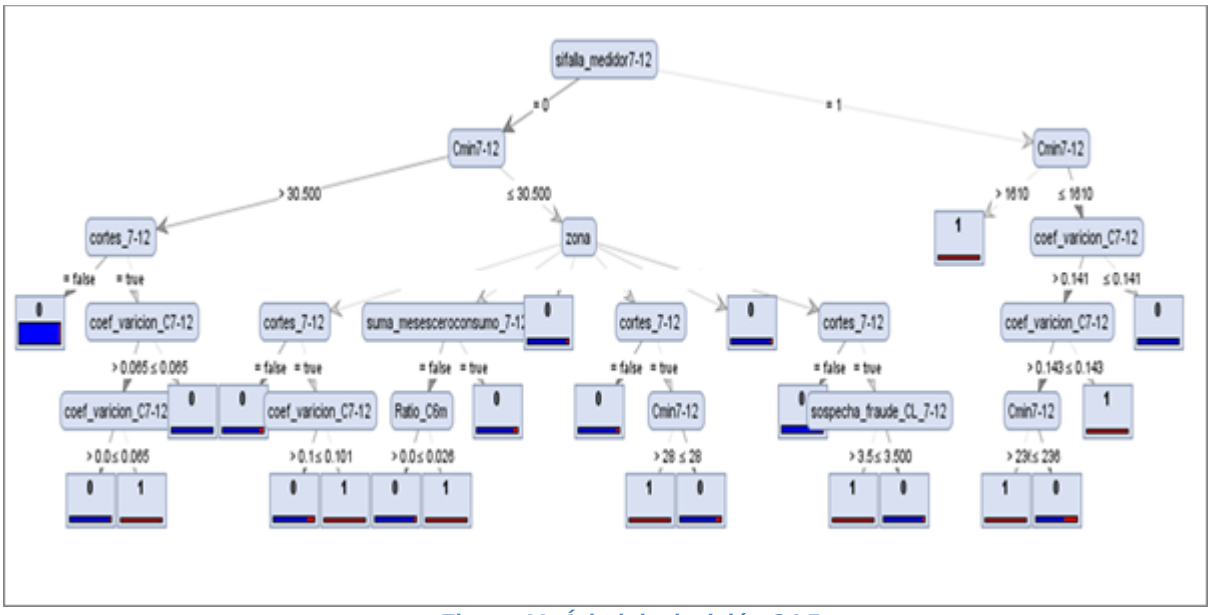

Figura 44: Árbol de decisión C4.5

Anexo G. Variables utilizadas en modelo Random Forest

Variable	Descripción	Tipo
Cmax_1-6	Consumo máximo primer semestre	Entero
Cmax_7-12	Consumo máximo segundo semestre	Entero
Cmax_dif	Diferencia entre consumos máximos ambos semestres	Entero
Cmin1-6	Consumo mínimo primer semestre	Entero
Cmin7-12	Consumo mínimo segundo semestre	Entero
Cmin_dif	Diferencia entre consumos mínimos ambos semestres	Entero
Prom_ECM_mmovil_consumo	Promedio error cuadratico medio media movil consumo	Real
Ratio_C6m	Ratio consumo	Real
coef_variacion_C1-6	Coeficiente variación consumo primer semestre	Real
coef_varicion_C7-12	Coeficiente variación consumo segundo semestre	Real
consumo_atipico_CL_1-6	Marcas por consumo atípico (CL) primer semestre	Entero
consumo_atipico_CL_7-12	Marcas por consumo atípico (CL) segundo semestre	Entero
cortes_1-6	Cortes de suministro primer semestre	Entero
cortes_7-12	Cortes de suministro segundo semestre	Entero
falla_medidor_CL_1-6	Marcas por falla de medidor (CL) primer semestre	Entero
falla_medidor_CL_7-12	Marcas por falla de medidor (CL) segundo semestre	Entero
lectura_estimada_CL_1-6	Marcas de lectura estimada (CL) primer semestre	Entero
lectura_estimada_CL_7-12	Marcas de lectura estimada (CL) segundo semestre	Entero
pearson_zona-rubro	Coeficiente correlación consumo misma zona-rubro	Real
prom_C1-6	Consumo promedio primer semestre	Real
prom_C7-12	Consumo promedio segundo semestre	Real
prom_factur_1-6	Facturación promedio primer semestre	Real
prom_factur_7-12	Facturación promedio segundo semestre	Real
prom_saldo_1-6	Saldo promedio primer semestre	Real
prom_saldo_7-12	Saldo promedio segundo semestre	Real
sospecha_fraude_CL_1-6	Marcas de sospecha fraude(CL) primer semestre	Entero
sospecha_fraude_CL_7-12	Marcas de sospecha fraude (CL) segundo semestre	Entero
mesesceroconsumo_1-6	Meses sin consumo primer semestre	Entero
mesesceroconsumo_7-12	Meses sin consumo segundo semestre	Entero
var%_Factura	Variación porcentual facturación ambos semestre	Real
zona	Zona	Nominal

Tabla 51: Listado completo de variables utilizadas en modelo Random Forest

Anexo H. Puntos de máximo beneficio

# Meses cobro retroactivo	Probabilidad de corte	Tasa fraude esperada	# óptimo de Inspecciones	Máximo beneficio obtenido	Utilidad en máx capacidad de inspección (7800)
1	40.74%	44.94%	12	$ 7.096	**-$ 33,265,284.65**
2	20.30%	29.35%	88	$ 210.217	**-$ 25,330,969.31**
3	13.51%	21.01%	227	$ 670.821	**-$ 17,396,653.96**
4	10.14%	14.92%	643	$ 1.617.417	**-$ 9,462,338.61**
5	8.09%	10.87%	1890	$ 3.445.679	**-$ 1,528,023.27**
6	6.74%	8.67%	4938	$ 7.484.416	**$ 6,406,292.08**
9	4.49%	6.30%	18256	$ 38.789.445	**$ 30,209,238.12**
12	3.37%	5.05%	37573	$ 99.376.806	**$ 54,012,184.16**

Tabla 52: Punto de máximo beneficio económico según plazo de recuperación (cobro retroactivo)