Trabajo fin de grado

Predicción del momento de recarga

01/02/2013

Roberto Canduela Luengo

0.- Abstract

El presente trabajo aborda la estimación del momento futuro de la recarga, para líneas de telefonía móvil. Con el objetivo de, teniendo una estimación de cuando efectuará el cliente la próxima recarga, poder realizar acciones comerciales para incrementar el importe recargado y/o disminuir el tiempo entre recarga.

Se han empleado las variables históricas de recargas y saldo de la línea.

Para abordar esta tarea se han usado dos métodos de clasificación supervisada: regresión logística y redes neuronales, en concreto el Perceptrón Multicapa.

Se han probado segmentaciones de las series históricas de datos con el objetivo de encontrar grupos con comportamientos homogéneos que faciliten el ajuste de los modelos de clasificación. En el proceso de segmentación se empleará la distancia euclídea y la DTW.

Se ha descartado el uso de los métodos de segmentación puesto que no aportan un valor extra significativo y añaden complejidad al proceso. El mejor ajuste se ha conseguido con un Perceptrón Multicapa, usando las recargas históricas de la línea y el saldo del último día antes de la clasificación.

El ajuste obtenido mejora el modelo preexistente y es suficiente como para poder ser usado comercialmente.

1.- Contexto de negocio

En la actualidad las operadores de telecomunicaciones son algunos de los actores más relevantes en el contexto empresarial español, el nivel de cobertura de la población en comunicaciones hace años que ha superado, en accesos, a la población total del país. No es raro encontrar personas que cuentan con línea fija, ADSL o fibra y línea móvil, así como algunas que además disponen de televisión a la carta e incluso de tarjetas para dispositivos móviles como tabletas.

Las empresas de telecomunicaciones cuentan con grandes bases de datos en las que reside la información de los clientes, en cuanto a su comportamiento en llamadas, patrones de consumo, gastos diarios, mensuales, uso de servicios…

Todos estos factores han colaborado al desarrollo de los modelos analíticos en este sector económico, así como a la creación de grupos de trabajo especialmente dedicados a la aplicación de técnicas estadísticas para la mejora de los resultados corporativos.

2.- Objetivos del proyecto

El presente proyecto tiene como fin predecir, con suficiente antelación, el momento en el que el cliente de prepago va a recarga su tarjeta.

Si centramos más el problema el objetivo último será determinar cuándo, una línea determinada, realizará la primera recarga dentro del período de predicción, o en su caso no hará ninguna recarga. Esta puntualización es importante, porque en un período dado, por ejemplo un mes, no es deseable contactar a una misma línea para realizarle la misma acción comercial más de una vez. Por tanto, en caso de que una línea recargue más de una vez en el período nos centraremos solamente en la primera recarga.

Introducción al problema de negocio.

Desde el punto de vista de la compañía tiene interés conocer el momento futuro en el que el cliente introducirá saldo en su tarjeta prepago por varios motivos:

- Conocer cuándo, nos permite realizar acciones comerciales de comunicación con el cliente para aminarle a recargar un mayor importe, normalmente incentivado a través de regalo de saldo.
- Conocer cuándo, nos permite sugerir al cliente cómo y dónde debe realizar la recarga.

Como se explicará en el presente trabajo este último punto es de gran importancia si se tiene en cuenta la comisión que paga el operador dependiendo del canal.

- Recarga efectuada en un cajero.
- Recarga efectuada en un distribuidor propio.
- Recarga efectuada en un distribuidor no propio.
- Recarga efectuada a través del portal web del operador.

Hay que tener en cuenta que entre algunas de estas formas de hacer una recarga el operador tiene una diferencia de ingresos reales que ronda el 5%, si esto lo extendemos a millones de recargas al año se evidencia la importancia del problema.

3.- Antecedentes

Hace años el operador para el que se desarrolla el proyecto diseñó un modelo de previsión del momento de la recarga basado en la información de los importes recargados por las líneas. Como se explicará más adelante la compañía no cuenta con la información del saldo diario de la línea en su entorno DW, obviamente sí en los sistemas operacionales.

La falta de la información de saldos y las urgencias de negocio propiciaron que se hiciera una aproximación al problema de bajo coste, siendo su eficacia suficiente para la realización de acciones comerciales.

La aproximación planteada consistía, a grande rasgos, en la evaluación de los tiempos entre recarga para cada línea, de forma que para cada uno de estos intervalos se calculaba la diferencia de saldos entre los extremos del intervalo, lo que nos daba una cantidad consumida, que al ser dividida por la amplitud del intervalo nos proporcionaba un consumo medio diario, particularizado para cada línea. Disponiendo de dicha información se buscaba la última recarga de la línea y se aplicaba el ratio, haciendo una estimación de los días que tardaría la línea en consumir dicho saldo y por tanto en recargar.

El problema de la primera aproximación es que solo nos da información para líneas que tuvieran un histórico de al menos 10 recargas. Dado que las líneas puede tener gran variabilidad en las recargas y que la recargas promocionales pueden afectar en gran medida a su comportamiento, era necesario disponer de un número suficientemente grande de recargas. El problema surge al exigir un mínimo de diez recargas, puesto que excluye a un gran número de líneas de prepago que por no tener suficiente antigüedad no cuentan con tantos eventos.

Por otro lado el actual modelo identifica las líneas que van a recargar en los próximos 30 días, los intentos con la citada metodología para reducir el tiempo de previsión, con márgenes de acierto razonables, fue infructuoso.

Por tanto en el actual proyecto pretendemos:

- Generalizar el modelo actual a todas las líneas, independientemente del número de recargas que hayan realizado, exigiendo, eso sí una antigüedad mínima como cliente.
- Reducir, si es posible, el umbral de previsión, pasando de treinta días a quince o hasta donde mantengamos un acierto razonable para dar soporte a acciones comerciales.
- Trabajar con información de saldos, que a todas luces debería proporcionar mejores resultados que la información de recargas.

4.- Fuentes de información

4.1.- Entorno de trabajo

4.1.1.- Base de Datos

El entorno de trabajo para el presente proyecto será una Base de Datos Oracle, en la que residen todas las fuentes de información consultadas. Se crearán todas las estructuras de información en dicho entorno con SQL Oracle.

Los códigos empleados se presentarán en la parte de anexos, tiene especial interés los códigos de trasposición de información para la modelación de las series de recargas y de saldos. Se debe hacer hincapié en la complejidad de construir la variable de saldo, dado que no reside en el DW y ha tenido que ser recreada a partir de las comunicaciones de la línea.

4.1.2- Entorno de modelización estadística

Para realizar la fase de predicción se ha empleado la herramienta Octave, software de libre distribución que emula a Matlab. Los algoritmos empleados han sido programados o modificados por el alumno con el objetivo de adatarlos de la mejor forma a los propósitos del proyecto.

La herramienta Octave, al igual que Matalab, están diseñadas para trabajar de manera muy eficiente con cálculo matricial, lo que facilita en gran medida el desarrollo de los algoritmos específicos de modelado que se usarán en el proyecto.

Todos los códigos han sido realizados con la versión de Octave 3.2.4.

4.2- Muestra aleatoria

Seleccionamos una muestra de líneas de prepago, representativas que están activas a cierre del mes de noviembre y que tengan una antigüedad en la compañía superior a 6 meses, por las siguientes razones:

1. Pretendemos obtener una muestra de información de la mejor calidad y con el menor ruido posible. Entendemos que los clientes con antigüedad superior a 6 meses ya son maduros en el servicio y por tanto tienen su patrón de recargas estabilizado.
2. A los seis meses deberían haber superado la fase previa en la que la compañía inyecta a las líneas saldo promocional en el momento del alta.
3. Lo ideal sería alejarse más en el tiempo, pero estaríamos sesgando demasiado la muestra dado que los clientes de prepago tienen una gran volatilidad y hay un gran número de líneas que se activan y se desactivan todos los meses. El objetivo del

modelo es servir para el mayor número de líneas posibles, si nos alejamos más de seis meses empezamos a descartar un gran número de líneas.

Para la selección se ha empleado un muestreo aleatorio simple, basado en la generación de un número aleatorio para cada una de las líneas de la planta de Prepago del Operador, la ordenación de estos valores y la elección de las primeras n líneas.

En el colectivo de líneas activas no se encuentran las que:

1. Han agotado el saldo, ya sea por consumo o por caducidad.
2. Han migrado a contrato.
3. Han portado a otro operador.
4. La línea ha sido suspendida de uso por diversas razones (por ejemplo robo).

Es importante destacar el punto 1, las líneas que agotan saldo, ya sea por consumo o por caducidad, tienen un tratamiento especial por parte del Operador y, por tanto, no son objeto de este estudio. Nos centraremos solo en las líneas que no agotan el saldo y que recarga de forma periódica, con el objetivo de predecir el momento de dicha recarga.

A partir de este momento todos los datos que se muestren se calcularán sobre la muestra del parque de líneas de prepago, por motivos de confidencialidad no se explicitará el número de clientes que hay en la muestra, para evitar extrapolaciones que proporcionen información de negocio.

El % de líneas eliminadas del total en base a los criterios anteriores es del 19,37%.

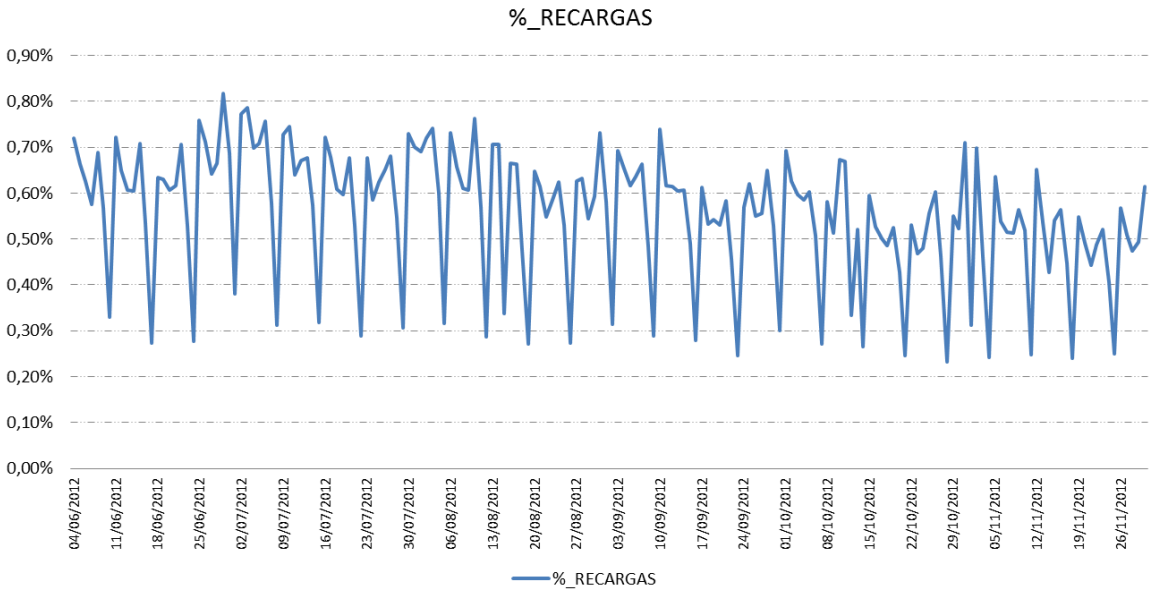
4.3- Variables

El presente estudio se ha basado en el uso de dos variables, que pretenden facilitarnos la predicción del momento futuro de recarga por parte de cada línea de prepago.

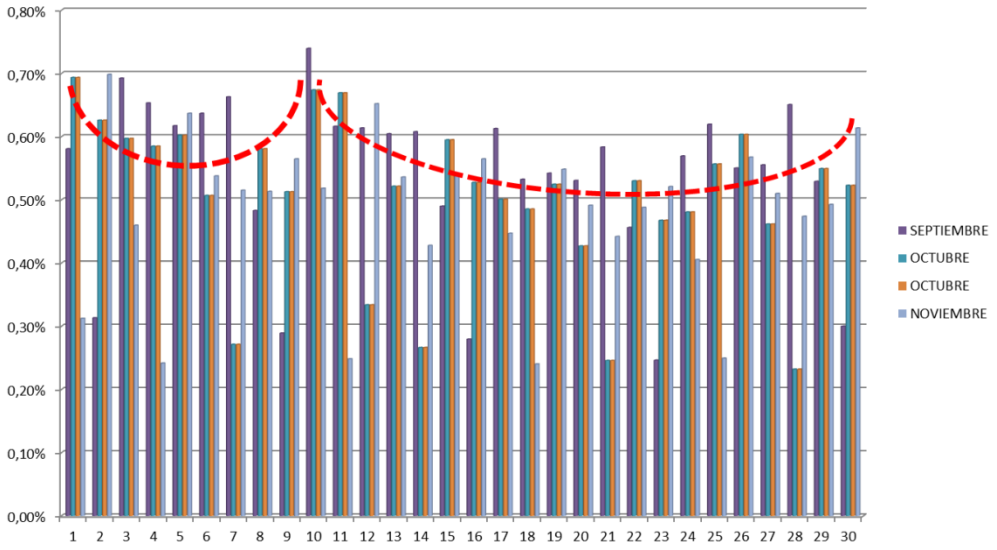
4.3.1.- Importe recargado por día

Indicamos el importe recargado (no promocional) por la línea en el día t, para el análisis descriptivo de los datos se han recuperado 180 días. Esta longitud de estudio es de interés, dado que el ciclo de vida de una tarjeta de prepago es como máximo, si el cliente no recarga de 6 meses, al pasar ese tiempo, y en el caso de que no haya agotado el saldo con anterioridad, perdería el importe que tuviera disponible en la tarjeta.

Si visualizamos la serie del % recargas por día en el período de 180 días tenemos:

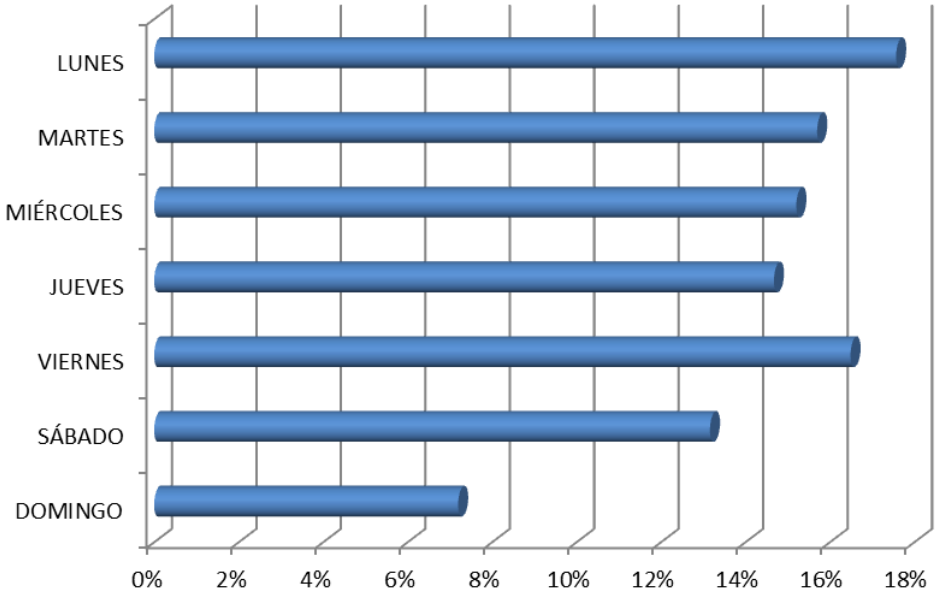


Aparecen ciertas estacionalidades dentro del mes, posiblemente debidas a los momentos de cobro de la nómina por parte de los usuarios. Los valores más altos aparecen en los primeros días, descendiendo de forma progresiva, con un aumento en torno al día diez del mes. A partir de este momento se produce un claro descenso en el número de recargas hasta final de mes.



Hay que destacar que se ha detectado una especial sensibilidad en el comportamiento de las recargas en función del día de la semana.

Ejemplo recargas por día de la semana:



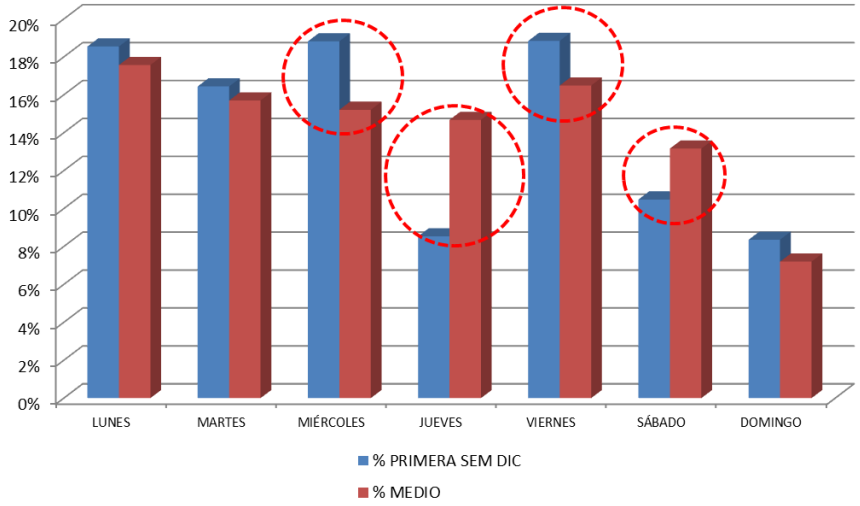
Los datos porcentuales de distribución de las recargas por día de la semana muestran como los días de mayor actividad son los lunes, después del fin de semana, y los viernes, antes del fin de semana, parece claro que estos dos días están concentrando la actividad que falta en el sábado y el domingo. Sobre todo hay que tener en cuenta que el domingo hay una importante bajada de las recargas originado, aparte de por el descenso de tráfico, por el cierre de los distribuidores, lo que origina que un gran número de personas que habitualmente los usan para recargar deban esperarse al lunes o adelantarse al viernes para poder hacerlo por su método habitual.

Nota de negocio: a partir de la información obtenida en el gráfico anterior podemos determinar, que en el caso de que hubiera que elegir algún momento de la semana específico para realizar la acción comercial, los mejores serían los domingos-lunes, para intentar influir en las posibles recargas que se producen los lunes, o los jueves-viernes, para hacerlo en las que se producen los viernes.

También hay que destacar que se ha detectado una gran sensibilidad de las recargas los días de fiesta, se aprecia como las recargas habituales de ese día se traspasan al día anterior, adelantando el momento de recarga. Este punto es de gran importancia a la hora de seleccionar la muestra de entrenamiento, intentando evitar en la medida de lo posible que en los días que forman el target haya festivos. En las variables independientes esto será imposible dado que, con el histórico que tenemos, necesariamente caerá algún día festivo, si bien al ser un histórico amplio se espera que su efecto sea asumible.

Analizaremos como ejemplo lo que ocurrió durante las dos primeras semanas del mes de diciembre de 2012, con los festivos.

Ejemplo de comportamiento de las recargas en días festivos:



Comparando los datos medios extraídos de la muestra con los obtenidos durante la primera semana de diciembre que va del día tres hasta el nueve, sabiendo que los días 6 y 9 fueron fiesta, podemos observar como las recargas en volumen se trasladan, cuando hay un día de fiesta, al día anterior o posterior. Hay que tener en cuenta que muchos clientes realizan la recarga a través de distribuidor y en los festivos al estar cerrado se ve imposibilitado para hacerlo.

Nota de negocio: Dado que uno de nuestros objetivos es modificar el punto de recarga de los clientes que acuden al distribuidor, con el objetivo de que pasen a cajero o portal web de la Operador, con el objetivo de mejorar los márgenes, parece interesante la opción de hacer algún tipo de comunicación a los clientes que vayan a necesitar recargar con alta probabilidad durante los días festivos. Recordándoles que los cajeros y la recarga online están siempre a su disposición.

4.3.2.- Saldo diario

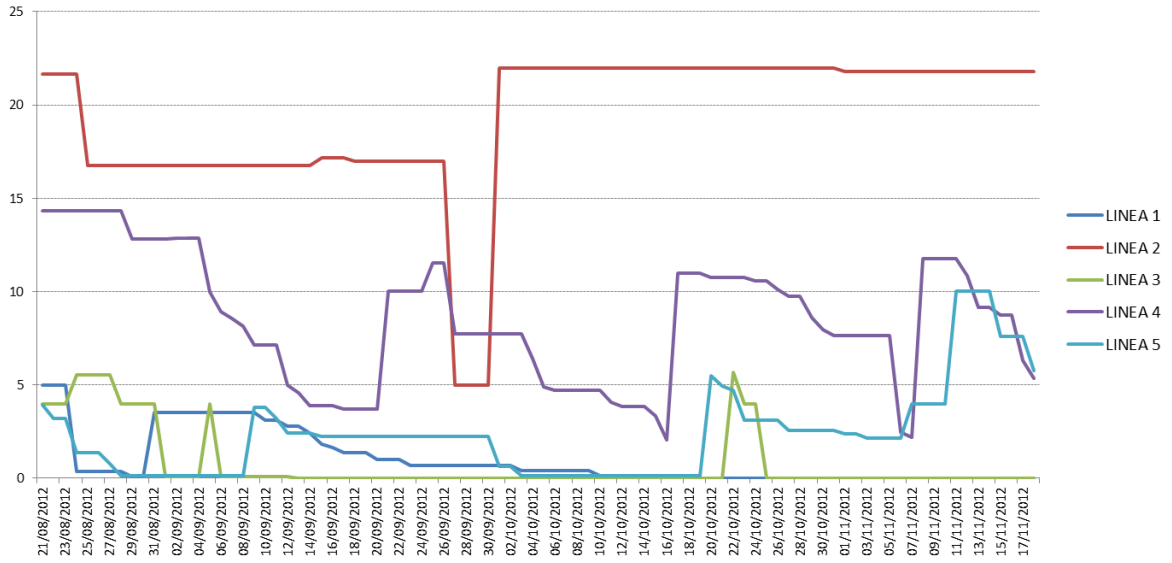
Se ha reconstruido el saldo diario de cada línea de la muestra.

Dispondremos del saldo de la línea a cierre de día, esto es, a las 23:59 horas. Esta variable debería ser de gran importancia puesto que una línea sin saldo no puede realizar llamadas, aunque si recibirlas durante cierto tiempo, no obstante el umbral de recarga varía con cada usuario, mientras que hay algunos que recargan de forma sistemática un determinado día del mes o la semana, otros lo hacen al llegar a un umbral saldo y otros cuando agotan.

La variable de saldo diario se podrá recrear como máximo para tres meses, 90 días. Al no disponer de dicha variable calculada directamente hay que realizar un costoso proceso, que implica analizar todas las comunicaciones de la línea e ir verificando el saldo posterior a las

mismas. Por tanto esta variable dependerá funcionalmente de la existencia de histórico de comunicaciones de la línea, que es de tres meses, de ahí que no se pueda igualar la estructura temporal de las recargas y el saldo diario.

Ejemplo: saldo para 5 líneas:



4.3.3.- Estructura temporal de la información

Inicialmente se ha planteado una estructura de información basada en seis meses para las recargas y tres para el saldo, esta asimetría viene dada por la disponibilidad de la información en el DW corporativo de la Operadora. Mientras que de recargas se dispone de todo el histórico el saldo hay que reconstruirlo a partir de las comunicaciones, lo que hace depender esta variable del histórico disponible.

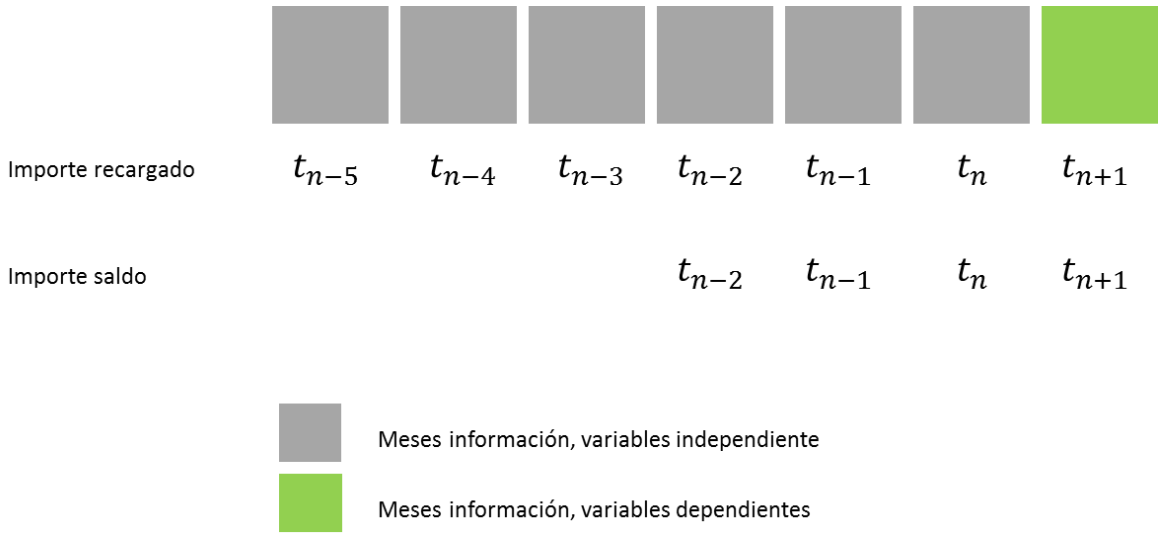
Más adelante en el proyecto se analizará la necesidad de emplear seis meses de recargas, estudiando la posibilidad de emplear solamente tres, con lo que usaríamos el mismo espacio temporal de recargas y de saldo.

4.3.4.- Planteamiento temporal inicial

El planteamiento inicial para la recogida de información consiste en construir una serie de variables para cada uno de los días del período analizado, y cada variable genérica (saldo, importe recargado).

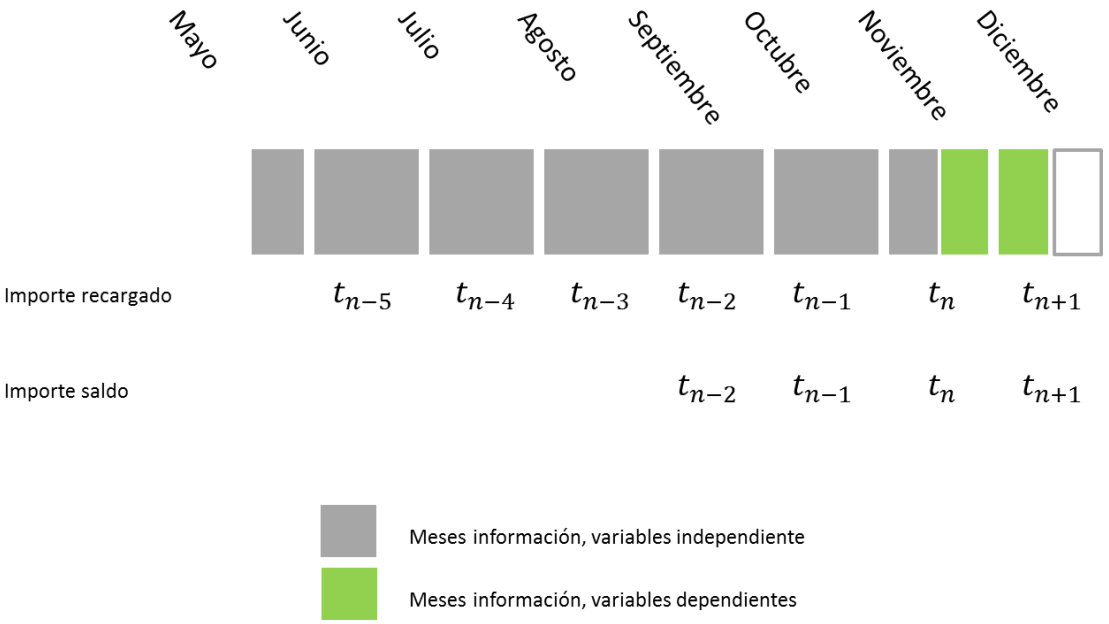
Hemos comenzado con una estructura diaria por ser la mínima unidad accionable comercialmente, actualmente el DW no trabaja en tiempo real y por tanto no tendría sentido

bajar a un nivel informacional inferior, complicando el procesado de construcción de variables, dado que incrementaría de forma considerable su número.

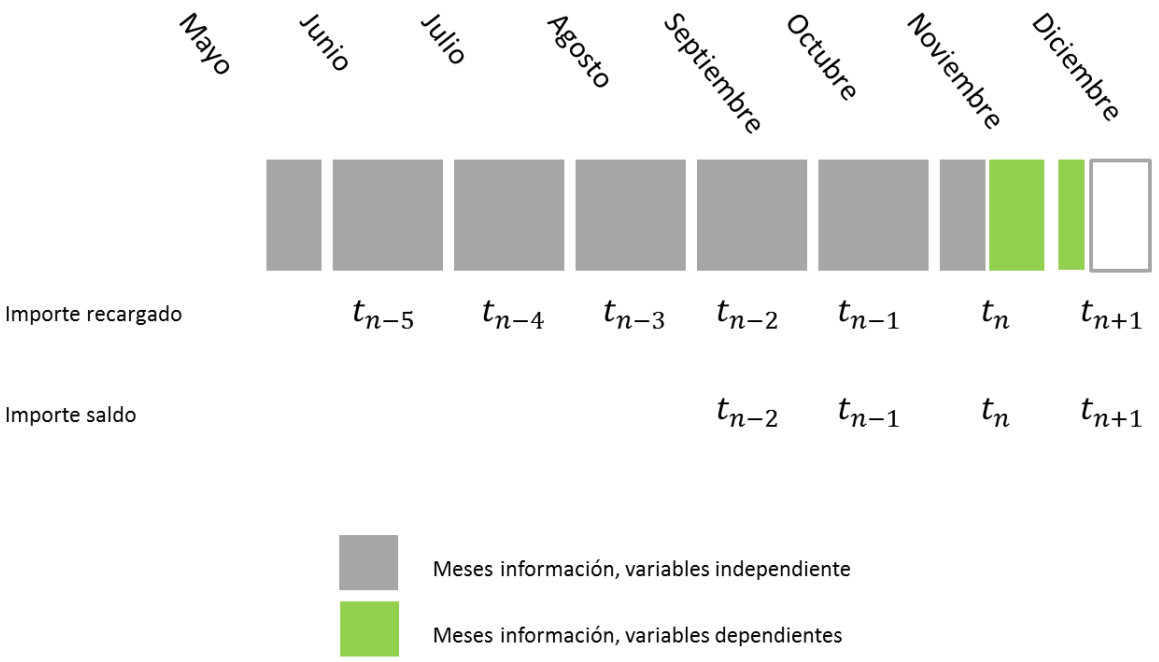


Como ya hemos visto a través del análisis de la información de las recargas y el saldo, en semanas con días festivos se modifica de forma sustancial el comportamiento de las líneas, por un lado desplazan las recargas a los días no festivos y por otro parece que se produce una disminución de las recargas que puede ser consecuencia de una disminución de las llamadas, al reducirse el número de comunicaciones en días festivos.

Por todo ello se ha descartado el planteamiento inicial que implicaba llegar con la variable dependiente hasta el 15 de Diciembre.



Hemos desplazado toda la extracción una semana antes para evitar el período festivo de principios de diciembre.



Seleccionamos seis meses de recargas y tres de saldo, partiendo de la segunda quincena del mes de mayo hasta la primera quincena del mes de noviembre cómo variables independientes de entrada en el modelo y 30 días como variable dependiente, de la segunda quincena del mes de noviembre a la primera quincena del mes de Diciembre.

Se ha optado por este planteamiento con meses divididos para no entrar en el período de vacaciones navideñas, que altera de forma significativa los datos, tanto de recargas como de saldos, al aumentarse de forma considerable las comunicaciones.

No tiene importancia a la hora de la estructuración del modelo, dado que las variables de entrada irán siempre a nivel diario y las de salida (target) se estructuraran convenientemente en diversas formaciones con el objetivo de obtener la mayor granularidad posible, manteniendo unos aciertos suficientes como para poder dar soporte a acciones comerciales.

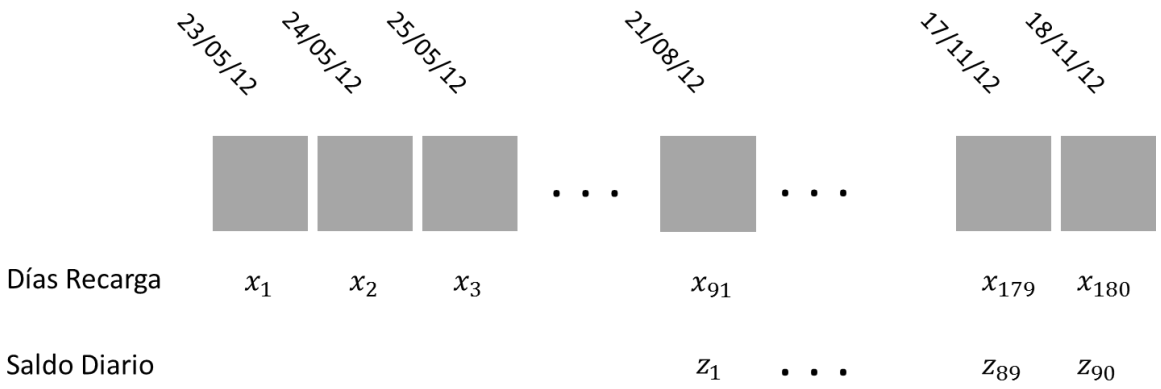
Como ya hemos visto anteriormente debido a las fiestas de las primeras semanas de Diciembre se altera considerablemente el patrón de recarga e incluso el volumen de estas, se ha comprobado que esa semana el volumen total de recargas disminuyó casi un 13% con respecto a la semana media calculada en los 6 últimos meses. Si bien las semanas anteriores y posteriores parecen absorber esta bajada. Por tanto consideramos que el uso de esta información agregada a nivel mensual puede ser válida, si bien su uso a nivel semanal es más discutible.

4.3.5.- Estructuración de las variables

El concepto de tratamiento de las variables consiste en generar una variable para la información de cada uno de los días, tanto en las variables independientes como en la explicativa. Como se ha comentado con anterioridad se ha seleccionado la unidad día, por ser aquella de máxima granularidad que tiene utilidad en el negocio. Por tanto dispondremos de variables consecutivas que nos expliquen la serie de acciones de cada línea, tanto en la parte independiente como en la dependiente.

4.3.5.- Variables independientes

Por motivos de histórico disponible se han seleccionado los datos de importe diario recargado con una amplitud de 180 días, por tanto hemos construido 180 variables, una por día. Así mismo, dado que no se puede disponer de un histórico del saldo diario superior a 90 días se ha construido una variable para cada día con el saldo a cierre de día. Por tanto las variables independientes quedarán estructuradas de la forma:



4.3.6.- Variables dependientes

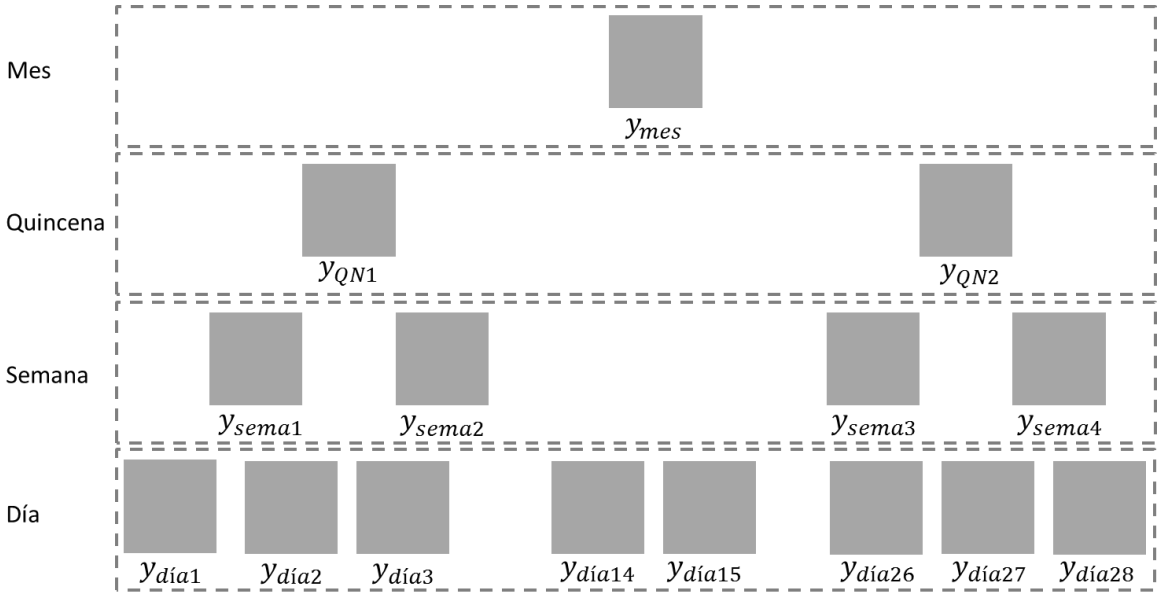
La variable dependiente estará formada por un multitarget, dependiendo del objetivo de la previsión.

Nos hemos planteado ir desde el problema más básico, el que cubre el modelo de previsión del momento de la recarga actual, que hace una estimación a nivel mensual (identifica las líneas que con cierta probabilidad van a recargar los próximos 28 días), hasta el problema más sofisticado que sería intentar predecir el día exacto de la recarga. Las diferentes posibilidades de la previsión son:

- Previsión de las líneas que van a recargar en los próximos veintiocho días. Generamos dos categorías, una para las líneas que recarga y otra para las que no lo hacen.

- Previsión de las líneas que van a recargar en períodos de dos semanas, en los primeros catorce días y en los siguientes catorce. Generamos tres categorías, una para el primer período, otra para el segundo y una tercera para las líneas que no recargan.
- Previsión de las líneas que van a recargar en las próxima cuatro semanas por semana. Generamos una categoría por semana y una última para las líneas que no recargan, en total cinco categorías.
- Previsión de las líneas que van a recargar en los próximos veintiocho días, a nivel diario. Generamos veintinueve categorías, una por día más la categoría de no recarga.

La estructura de información para los cuatro target posibles es:



A la hora de analizar el target tendremos en cuenta diferentes configuraciones de la información de salida, con el objetivo de ver hasta dónde podemos descender en la granularidad manteniendo unos aciertos razonables. Conviene recordar que el actual modelo en uso genera una marca que indica si el cliente va a recargar o no en los próximos 30 días. Por tanto cualquier planteamiento que lo mejore aportará valor a la compañía.

4.3.7.- Muestra aleatoria

Seleccionamos una muestra de líneas representativas que están activas a cierre del mes de noviembre y que tengan una antigüedad en la compañía superior a 6 meses. El resto de condiciones se pueden ver en el apartado 4.2.

El % de líneas eliminadas, en base a los criterios del punto 4.2, es del 19,37% sobre el total de la muestra inicial.

4.3.8.- Muestra de entrenamiento, validación y test

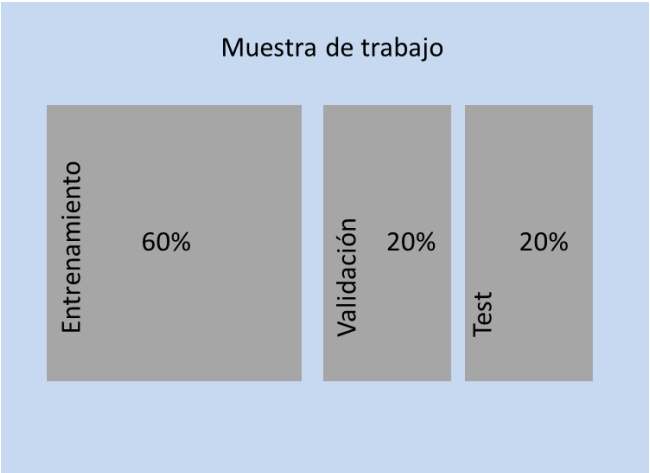
Hemos seguido la estructura habitual de modelación, dividiendo la muestra en tres grupos, a saber:

- Entrenamiento.
- Validación.
- Test.

La submuestra de entrenamiento se empleará para entrenar los algoritmos de clasificación.

La submuestra de validación se empleará para ajustar los parámetros de regularización de los modelos y para seleccionar las configuraciones que presentan un mejor acierto.

La submuestra de test se empleará para seleccionar el modelo óptimo, sin que haya participado en ninguna de las fases de ajuste del modelo.



5.- Tratamiento de los datos

5.1.- Tratamiento de las recargas promocionales

Por recarga promocional entendemos toda aquella aportación de saldo a la línea por parte del Operador sin que para el cliente tenga coste económico. Determinadas campañas comerciales se basan en el regalo de saldo extra a cambio de esfuerzos económicos por parte del cliente.

Hay una gran variedad de tipologías de campañas que conllevan la aportación de saldo por parte del Operador, las más significativas son:

1. Campañas para la estimulación del consumo.
2. Campañas de fidelización.

Ejemplo: Campaña comercial en la que se regala el mismo importe que recargue la línea, siempre que sea superior a 20€, habitualmente para clientes con bajos importes promedios recargados. Por tanto si la línea recargara 30€ el operador le regalaría otros 30€ extra, haciendo un total de 60€ de los que podría disponer en sus comunicaciones.

La recarga promocional puede o no ser una decisión voluntaria del cliente, si es voluntaria suele conllevar una recarga previa por su parte y estaremos evaluando dicha recarga y si no conlleva una recarga previa no ha sido el cliente el que la ha ocasionado y suele provenir de una estrategia comercial del Operador.

Nuestro objetivo es predecir la siguiente recarga, no promocional, por parte del cliente, por ello para realizar el entrenamiento deberemos contar solo con dichas recargas, no obstante las recargas promocionales aportan saldo extra, lo que prolonga el tiempo entre las recargas. Para resolver esto se ha procedido a eliminar las recargas promocionales de la muestra de entrenamiento, asignando su saldo a la recarga no promocional anterior. De esta forma mantenemos la muestra coherente con el objetivo del estudio y al mismo tiempo tenemos en cuanta todo el saldo disponible.

Ejemplo:

Datos originales

Teléfono	Tipo Recarga	Día Recarga	Importe
6XXXXXXXX	No promocional	01/10/12	10€
6XXXXXXXX	Promocional	01/10/12	15€

Datos procesados

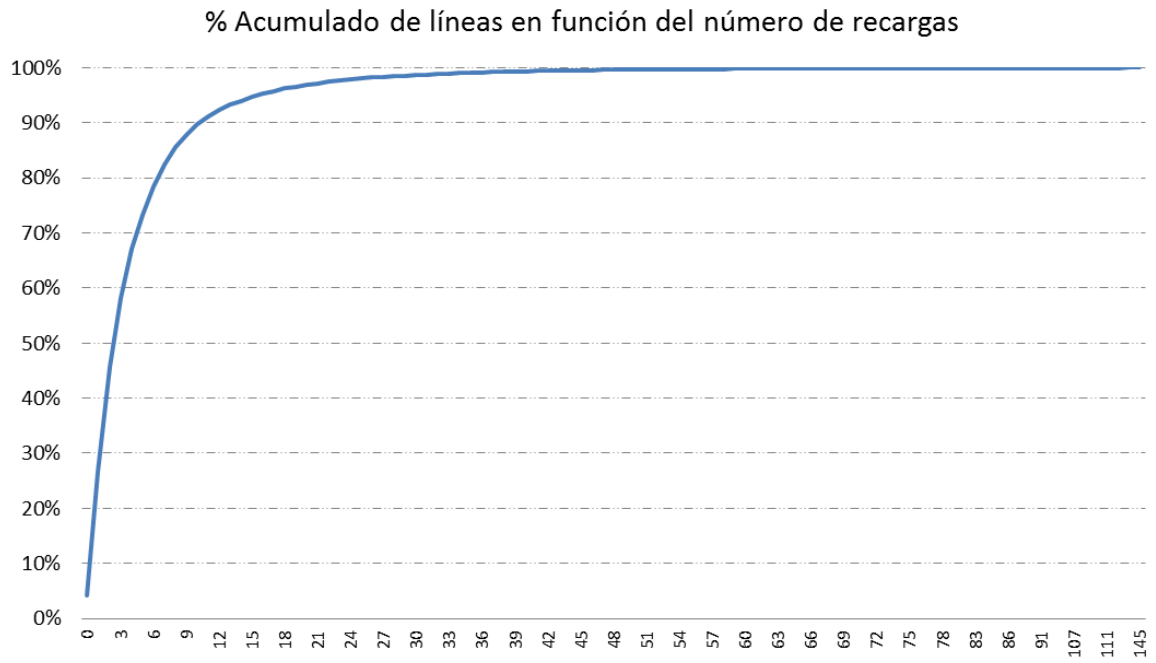
Teléfono	Tipo Recarga	Día Recarga	Importe
6XXXXXXXX	No promocional	01/10/12	25€

5.1.1.- Tratamiento de los outliers en recargas

Si estudiamos el número de recargas de las líneas en el período analizado, podremos observar como la distribución es fuertemente asimétrica, esto puede ocurrir por dos razones:

1. Los datos del número de recargas por línea al igual que una gran cantidad de métricas en el sector de las comunicaciones se ajustan a *distribuciones long-tail.*
2. Se puede estar produciendo fraude con esas líneas, es habitual que desde locutorios se usen tarjetas de prepago.

Si acudimos a una gráfica en la que podamos ver el % de clientes, del total de la muestra, en función del número de recargas que hacen en el período de estudio veremos como con un bajo número de recargas acumulamos una gran cantidad de líneas (con 9 recargas acumulamos casi el 90% de las líneas).



El objetivo del estudio es ser capaz de predecir el momento de la recarga futura para la generalidad de la planta de líneas de prepago, por tanto decidimos eliminar líneas que consideramos se comportan de forma anómala. Si el objetivo fuera la detección de anomalías o de fraude trabajaríamos precisamente con esas líneas.

La media de recargas en el período de estudio (180 días) es de 5,56 por línea, con una desviación típica de 8,13. Si seleccionamos un intervalo con alta confianza, por ejemplo 6σ tendremos que las recargas válidas en la muestra de entrenamiento estarán dentro del intervalo [0,54].

$\bar{X}$	σ	$\hat{\sigma}$	$\bar{X} + n.\sigma$	% Líneas excluidas
5,56	1	8,13	14	7,26%
5,56	2	16,26	22	3,53%
5,56	3	24,39	30	1,92%
5,56	4	32,52	38	1,13%
5,56	5	40,65	46	0,70%
5,56	6	48,78	54	0,45%

Decidimos eliminar las líneas que realizan 54 recargas o más en el periodo de estudio.

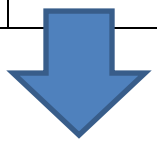
5.1.2.- Consolidación de las recargas diarias

Puede ocurrir que una misma línea haga más de una recarga al día, en ese caso para unificar la información y dado que la medida de agrupamiento temporal que vamos usar es el día, se ha decidido consolidar el importe recargado dicho día y contabilizar solamente una recarga.

Ejemplo:

Datos originales

Teléfono	Día Recarga	Importe
6XXXXXXXX	01/10/12	10€
6XXXXXXXX	01/10/12	15€

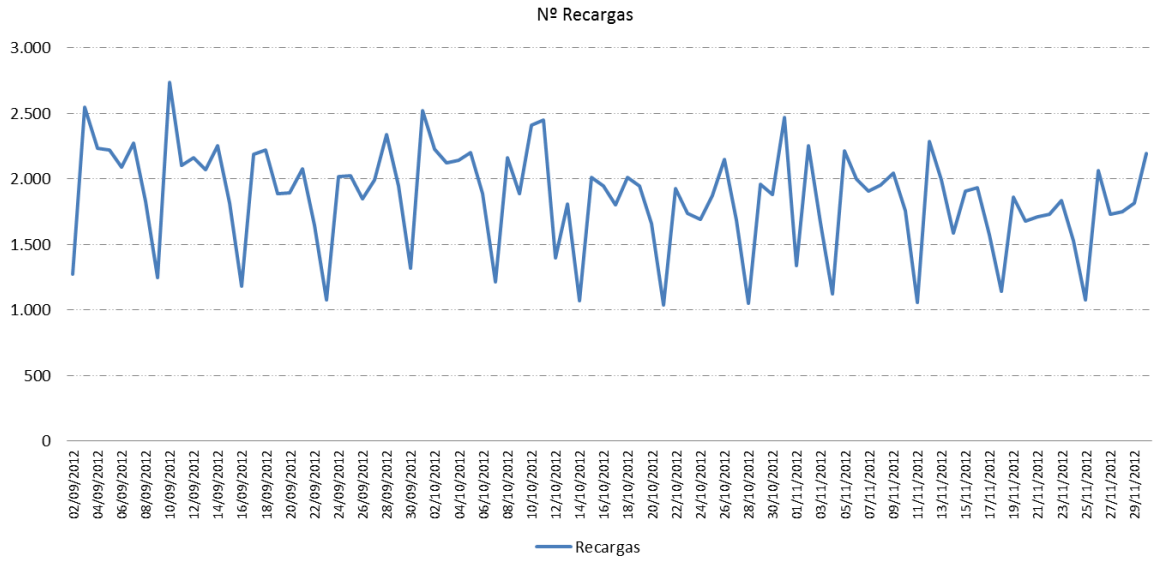


Datos procesados

Teléfono	Día Recarga	Importe
6XXXXXXXX	01/10/12	25€

5.1.3.- Análisis exploratorio de las recargas

Visualizamos las recargas cumuladas, para toda la muestra, para un período de tiempo de 90 días.



5.1.3.1.- Tendencia de la serie de recargas

Partiendo de los datos del número de recargas realizadas, en 90 días, por las líneas de la muestra hemos procedido a analizar la tendencia ajustando una regresión lineal por mínimos cuadrados.

Regresión lineal sobre la serie de recargas acumuladas:

y: número de recargas

x: tiempo en días.

La formulación de la regresión vendrá dada por [1]:

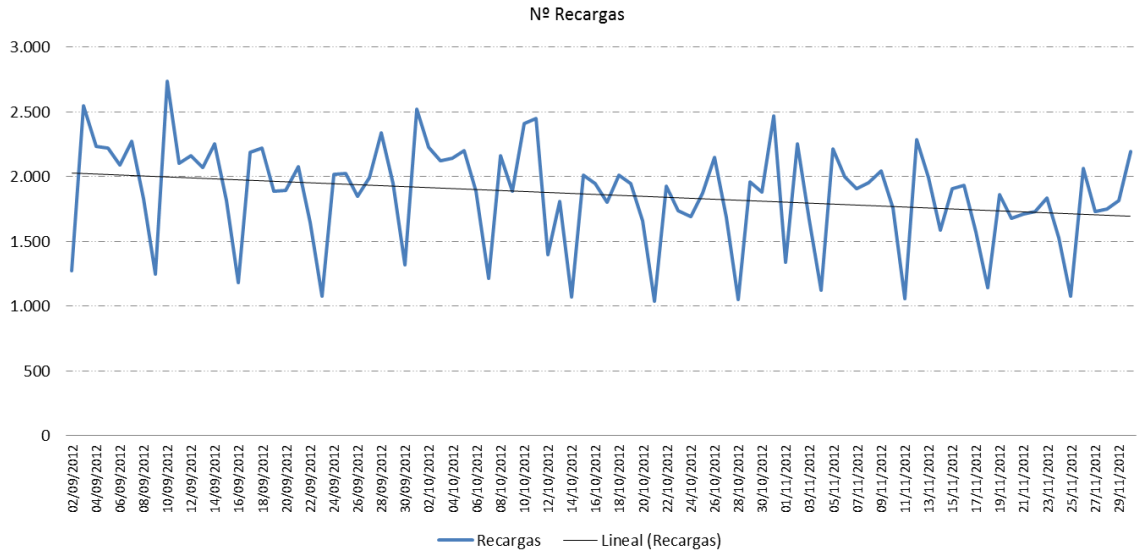
$$y = \beta_1 + \beta_2 . x$$

$$\hat{\beta}_1 = \bar{y}_n - \hat{\beta}_2 . \bar{x}_n \qquad \hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i . y_i - n . \bar{x}_n . \bar{y}_n}{\sum_{i=1}^{n} \bar{x}_i^2 - n . \bar{x}_n^2}$$

Obtenemos una ecuación de regresión:

$$y = 2031 - 3{,}71 . x$$

Como podemos observar la tendencia de la serie es decreciente, la gráfica queda:

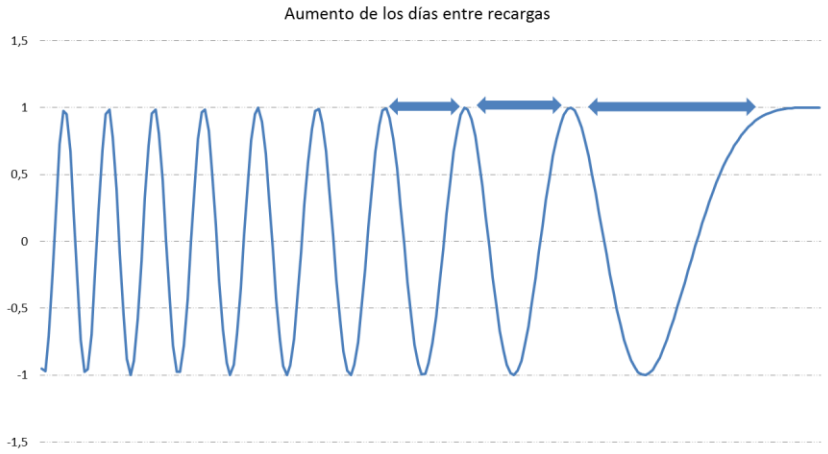


Parece claro que en el último período analizado, concretamente el mes de noviembre, han disminuido el número total de recargas por día. Dado que en esta época del año no se ha identificado, a lo largo del histórico, ninguna componente estacional puede estar ocurriendo:

1. Haya disminuido el número de clientes y por tanto el número global de recargas.
2. Se haya mantenido el número de clientes pero haya disminuido el número de recargas.

Dado que la muestra de clientes se mantiene activa durante todo el período de estudio, incluida la fase de target, la única explicación posible es que se estén produciendo menos recargas, lo que llevado a nivel de línea implica que el tiempo entre recargas está aumentando, al menos en un conjunto considerable de líneas.

Entendemos que se puede estar produciendo un comportamiento del tipo:

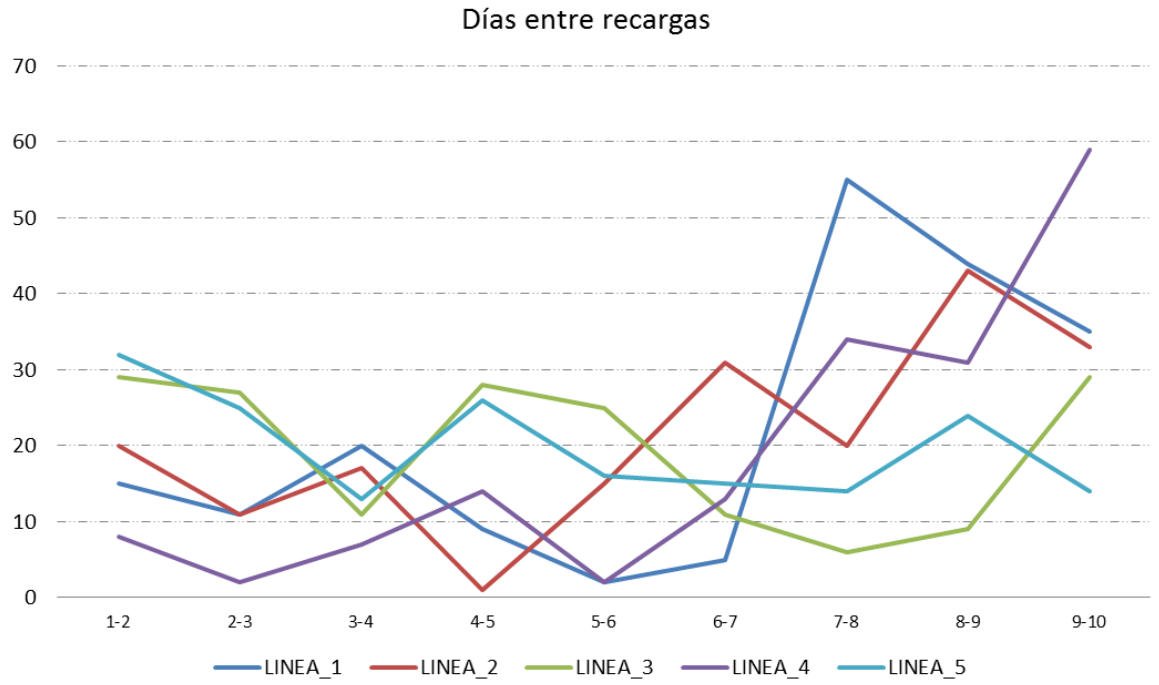


El comportamiento identificado implica que habrá que recalibrar el modelo en un plazo de tiempo no excesivamente elevado, no más de 4 meses, para evitar posibles problemas en el ajuste debidos a la marcada tendencia.

5.1.3.2.- Tiempo entre recargas

Dado que la tendencia de las recargas acumuladas por día es negativa, pasamos a estudiar los días medios, por línea, que transcurren entre una recarga n (consolidada por día) y la siguiente n+1.

En el siguiente gráfico podemos ver cinco ejemplos aleatorios de líneas que han realizado 10 recargas en el período de estudio. Las diez recargas se convierten en 9 diferencias de días entre una recarga y la siguiente.



Nota de negocio: existe una tendencia a aumentar el número de días entre recargas.

5.2.- Tratamiento de los saldos diarios

Dado que el Operador no dispone de una estructura de información consolidada en el DW que permita acceder a la información de saldo diario, de las líneas de prepago, nos hemos visto forzados a recrearlo a partir de diversas fuentes de información:

- Saldo disponible al finalizar llamadas salientes de las líneas.
- Saldo disponible después de enviar un SMS.
- Información de saldo quincenal.

- Información de saldo después de recargar.

Comentaremos brevemente cada una de las fuentes de información y la metodología que se ha empleado para la reconstrucción del saldo diario de las líneas. Empezaremos con una definición.

Definición: se entiende por saldo diario de una línea de prepago el importe del que dispone es su tarjeta telefónica para poder realizar llamadas a cierre del día, esto es a las 23 horas 59 min.

La metodología que hemos seguido para la reconstrucción del saldo diario es auditar todas las fuentes de información del operador, localizando aquellas en las que se haga referencia al saldo, ya sea a través de una foto fija o a través de alguna de las interacciones que puede tener la línea con la compañía (llamadas, SMS, MMS, datos, recargas).

Para la reconstrucción final se han tenido en cuenta las fuentes que aportan información relevante para un % amplio de planta de líneas, descartando las que presentan gran coste de procesamiento en la BD y que afectan a un % bajo de las líneas. En concreto hemos decidido no tener en cuenta los datos y los MMS. El % de líneas que acceden a datos es suficientemente bajo como para poder trabajar sin dicha información. Además hay que tener en cuenta que esas disminuciones del saldo las estamos controlando a través de otras métricas, en concreto:

- Información de recargas.
- Información quincenal del saldo.

En el DW del operador hay información del saldo, en fotos fijas, que se toma los días uno y quince de cada mes, dicha información nos sirve para regularizar los saldos diarios.

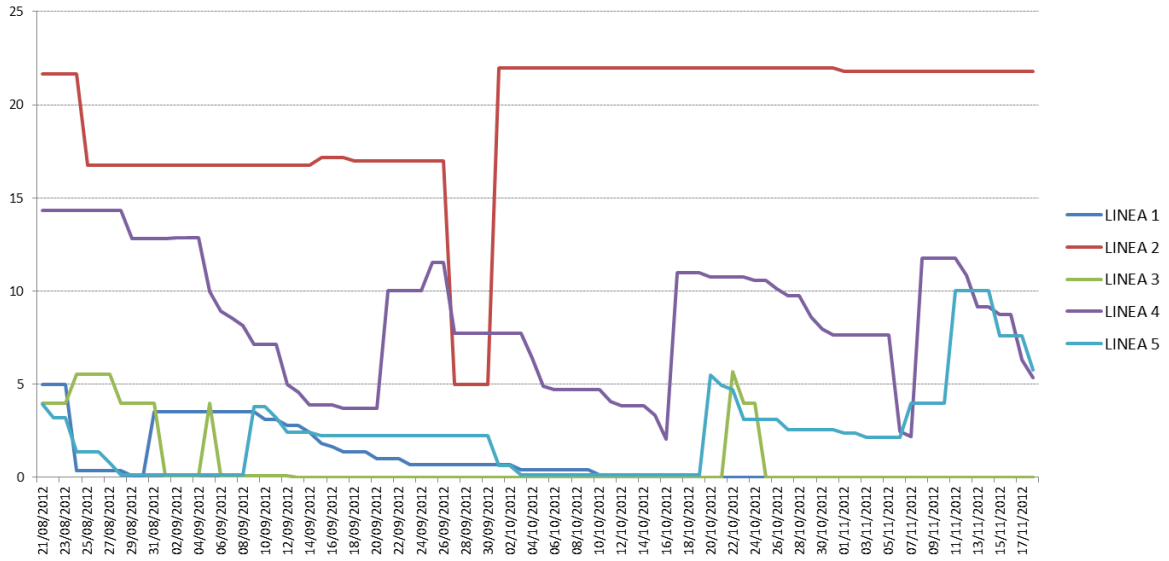
Por otro lado contamos también con la información del importe recargado por la línea y el saldo del que dispone justo después de realizar dicha operación, como suma del importe de la recarga y el saldo del que disponía antes de efectuarla.

Por tanto contamos con información fija, con fotos quincenales e información dinámica (recargas, llamadas…). Nuestro objetivo es llenar los huecos que quedan entre las fotos fijas con la información dinámica. De forma que para cada línea y cada día nos quedaremos con la información del saldo disponible en la hora más tardía.

Una vez que hemos completado los días para los que tenemos información de saldo deberemos rellenar los huecos, esto es, día para los que la línea, al no haber realizado ninguna operación no tiene registrado su saldo. La forma de hacerlo es sencilla, si por ejemplo el martes no tenemos saldo pero el lunes si lo teníamos arrastramos dicho valor del lunes al martes, por tanto tendremos el mismo valor en los dos días. Esto nos da el saldo que tenía la línea el lunes, después de su última interacción, y dado que no ha realizado ninguna otra, que tengamos registrada, entendemos que el saldo deberá ser el mismo que el del cierre del lunes, hasta la próxima interacción.

De forma iterativa vamos rellenando los huecos para todas las líneas, hay que tener en cuenta que una línea, como mucho puede tener 16 huecos, los que van de la foto de saldo del día 15 de un mes hasta la foto de saldo del día 1 del mes siguiente.

Como hemos visto con anterioridad, un ejemplo del seguimiento del saldo para unas líneas de ejemplo durante 90 días, puede ser:



En la gráfica se empiezan a atisbar, aunque de forma sutil, algunos comportamientos típicos de las líneas de prepago en cuanto a las estacionalidad sistemática de algunas de ellas (línea 4), el mantenimiento del saldo en valores altos aunque no haya consumo (línea 2) o el gasto del saldo hasta agotar (línea 3 y línea 5).

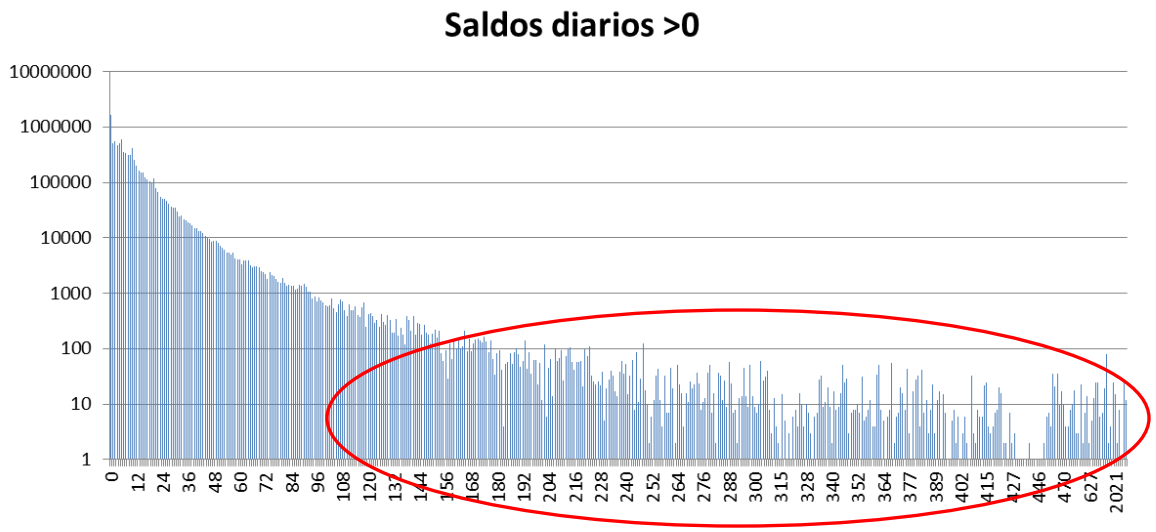
Hay que tener en cuenta que estos comportamientos pueden dificultar la previsión del modelo. Haremos una aproximación a ello en la segunda parte del trabajo.

Como ya se ha comentado con anterioridad, en otro punto, al tener que recrear el saldo diario a partir de otras variables, es necesario disponer de suficiente historio en todas ellas, lo que dificulta aún más su construcción, después de analizarlo se ha llegado a la conclusión de que solo se dispone de una ventana de 90 días en las que todas las variables estén informadas.

5.2.1.- Consolidación de los saldos diarios

Si analizamos los valores que pueden tomar los saldos diarios para cada una de las líneas, hay que tener en cuenta que una proporción muy elevada de las líneas presenta el saldo a cero, en el análisis diario.

En el siguiente gráfico visualizamos los saldos posibles que toman las líneas en el total de los días de estudio. Aplicamos una escala logarítmica en el eje de líneas para que se pueda visualizar la larga cola de la distribución.



5.2.2.- Tratamiento de outliers

Podemos ver que hay líneas que presentan saldos diarios muy elevados, por encima de 100€, que para una líneas de prepago se antoja un saldo alto. Podría tratarse de locutorios o del algún tipo de fraude.

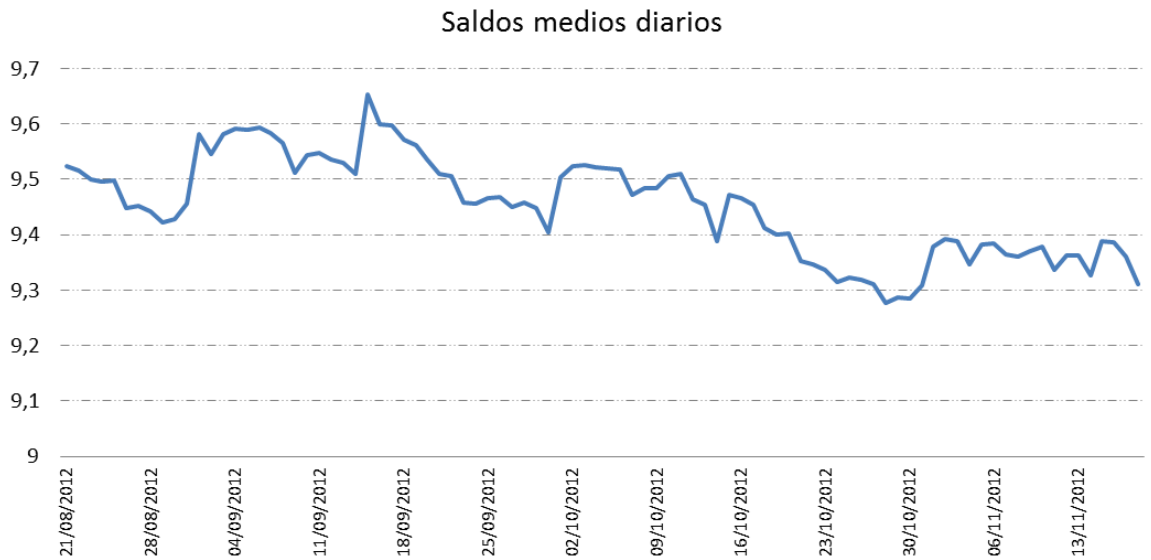
SALDO	LÍNEAS
50€	3161
100€	494
200€	77
300€	31

En la tabla podemos ver las líneas que en algún momento de los 90 días estudiados de saldo presentan valores superiores a {50, 100, 200, 300}.

A la vista de la información decidimos eliminar las líneas, del estudio de saldo, que tienen en algún momento saldo acumulados superiores a 100€.

Si estudiamos la tendencia de los saldos medios diarios apreciamos un ligero descenso, pasando en los primeros días del estudio de 9,5 a 9,35 en los últimos, si bien habría que descartar tendencias estacionales en el período, cuestión que no es posible abordar por la

falta de histórico. En cualquier caso el descenso es leve y no debería afectar, en demasía, al estudio.



6.- Aproximación a la modelización

6.1.- Trabajando con las recargas

Se ha decidido hacer una primera aproximación a lo resolución del problema contando únicamente con la información de las recargas, el objetivo es comparar los resultados de este planteamiento con el del anterior modelo, mucho más näive, que trabajaba solo con recargas y para un subconjunto de las líneas, como se ha explicado en los primeros apartados.

Para esta primera aproximación se usarán las técnicas de regresión logística y redes neuronales, con diferentes ponderaciones del conjunto de entrenamiento y con datos aportados siempre sobre el conjunto de test.

Para la evaluación de los datos con las diferentes técnicas se han creado procesos específicos en Octave, tanto para la regresión logística como para la red neuronal (MLP), para más información ver anexos.

6.1.1.- Recargas, planteamiento mensual

Para esta primera aproximación, sobre las recargas, emplearemos para cada línea l_i las $x_1 \dots x_{180}$ variables diarias creadas sobre las recargas y usaremos el target y_i, que contendrá la información agregada mensual esto es:

$$y_i = \begin{cases} \text{1 si } l_i \text{ recarga en 28 días.} \\ \\ \text{0 si } l_i \text{ no recarga en 28 días.} \end{cases}$$

Los datos iniciales presentan una distribución para la variable dependiente y_i de:

y_i	Porcentaje
0	56%
1	44%
TOTAL	100%

La muestra presenta un balanceo bueno para poder realizar el entrenamiento, no obstante probaremos con diferentes balanceos para los positivos, con el objetivo de detectar hasta dónde puede llegar el modelo ($y_i = 1$), con la variable de recargas.

6.1.1.1.- Regresión logística

La información sobre el algoritmo empleado se puede consultar en los anexos.

En la tabla X se muestran los resultados para el ajuste de la regresión logística sobre la muestra de test con los datos sin balancear.

P. CORTE	MATTHEWS	ACCURACY	PRECISION	RECALL	N	SELECCION	P SEL TOTAL
0.50	0.35	68.06%	73.43%	42.85%	12000	3079	25.66%
0.60	0.33	66.16%	79.59%	30.97%	12000	2053	17.11%
0.70	0.29	63.69%	83.13%	21.85%	12000	1387	11.56%
0.80	0.25	61.61%	86.48%	15.03%	12000	917	7.64%
0.90	0.20	59.56%	89.53%	9.08%	12000	535	4.46%

Podemos ver como en la primera columna se muestran las diferentes probabilidades de corte para las que se han obtenidos los datos de ajuste, 0.5 nos da la probabilidad de asignación por defecto de las clases {0,1} en la variable de salida.

En la segunda columna mostramos el índice de correlación de Matthews (ver anexo).

En la tercera tenemos la acuracidad o (accuracy) como acierto total en la predicción, entendido como total de ceros y unos clasificados de forma correcta.

En la columna de precisión podemos ver los datos del porcentaje de unos que el modelo clasifica correctamente, este dato es muy importante puesto que nos serve para determinar el punto de corte en la probabilidad óptima para la realización de campañas comerciales. Hay que tener en cuenta que este tipo de modelos dan soporte a acciones en las que no se ve implicada toda la planta de clientes y por tanto si se pueden determinar colectivos de menor tamaño pero con un acierto superior el resultado de la acción comercial será mejor.

En la columna de recall, presentamos la captura, esto es, del total de unos que hemos marcado en nuestra estimación que porcentaje suponen frente al total de unos de la muestra de test.

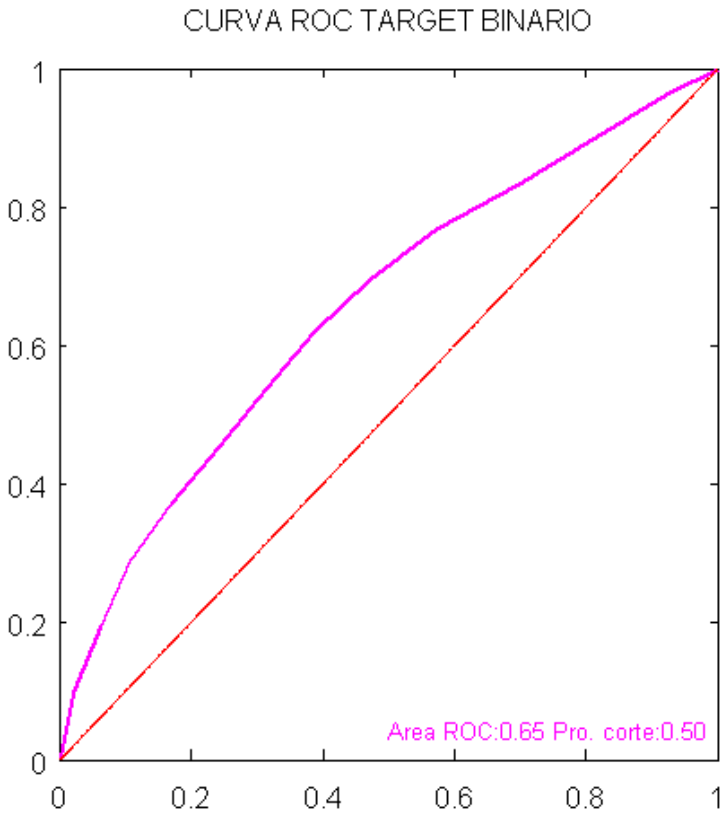
En la columna N visualizamos el tamaño de la muestra de test {0,1}.

En la columna selección tenemos la información del número de líneas que se marcarían como uno en función de la probabilidad de corte de la primera columna. Este sería el Publico Objetivo (PO) de nuestra acción comercial.

Por último tenemos una columna que nos indica el % de la selección para la campaña comercial frente a la base de líneas de la muestra de test. Es un valor interesante porque nos permite extrapolar los resultados a todas las líneas del operador y hacernos una idea del número de clientes que podrían participar en una acción comercial. Si por ejemplo la Operadora tuviera 1M de líneas y decidiéramos usar el 0.5 de probabilidad como punto de corte tendríamos un PO potencial, antes de filtrados de negocio, de 256K líneas.

Este punto es muy importante porque las unidades de negocio o Marketing, obviamente siempre están interesadas en conocer el acierto del modelo y la previsión de contratación a un producto, baja, recarga… pero también en conocer el dimensionamiento de la campaña comercial, esto es, a cuantos clientes habrá que hacer la oferta. Será necesario para dimensionar las plataformas y servicios que deberán garantizar que la operativa de la campaña funcione correctamente, además de permitir dimensionar el impacto económico de la misma, permitiendo ajustar las ofertas o incluso las probabilidades de selección en función de criterios presupuestarios.

La curva ROC para este primer modelo de regresión logística es:



6.1.1.2.- Regresión logística para muestra balanceada

Vamos a realizar una prueba de balanceo de la muestra de entrenamiento. El balanceo consiste, habitualmente, en sobreponderar la proporción de unos del target en la muestra de entrenamiento, con el objetivo de obtener modelos que maximicen la probabilidad de acierto en los unos. Esta técnica se emplea de forma habitual cuando los conjuntos de entrenamiento presentan proporciones de unos muy desequilibradas con respecto a los ceros. En mi caso la he empleado de forma habitual con proporciones de unos por debajo del 30%.

Dado que en este caso disponemos de pocas variables, en concreto por ahora estamos trabajando solo con las recargas, vamos a realizar una prueba de balanceo, aunque la proporción de unos sea suficientemente elevada en la muestra de entrenamiento. El objetivo es analizar si conseguimos modelos con mayor capacidad predictiva de los positivos, a costa de un recall inferior.

Atendiendo al ajuste de las curvas ROC y al resultado obtenido en las tablas individuales (para más información consultar Anexos) llegamos a la conclusión de que el mejor ajuste se consigue con la configuración inicial, sin balancear la muestra.

CONCLUSIÓN: Atendiendo a los resultados obtenidos con la regresión logística, sin balanceo, ya podemos afirmar que hemos conseguido mejorar los resultados del modelo existente hasta estos momentos. Dado que tenemos un acierto en los positivos del 73,4%, para toda la población, frente al 69% que se lograba con el anterior modelo, pero solo para un subnconjunto de las líneas (aquellas que habían realizado, históricamente, más de diez recargas).

6.1.1.3.- Perceptrón multicapa

Vamos a probar el mismo ajuste que en el caso anterior pero esta vez con una red neuronal, en concreto con un Perceptrón Multicapa. Las muestras de entrenamiento, validación y test son las mismas que para el apartado anterior. Las variables de entrada serán las $x_1 \dots x_{180}$ correspondientes a los días en los que se han medido las recargas para cada línea, y el target contendrá los valores {0,1} en función de que las líneas haya recargado o no en los treinta días siguientes.

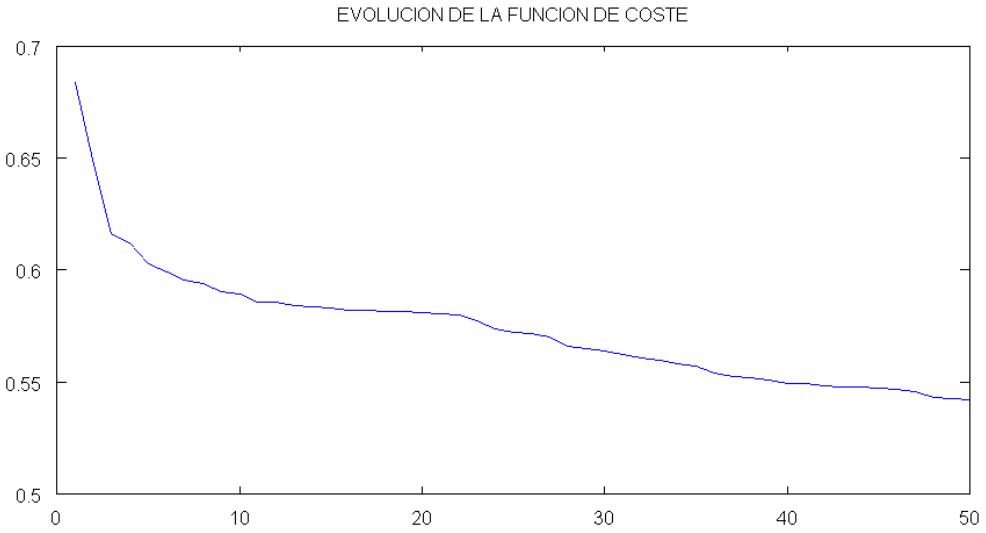
Si visualizamos la misma tabla resumen que para la regresión logística tenemos:

P. CORTE	MATTHEWS	ACCURACY	PRECISION	RECALL	N	SELECCION	P SEL TOT
0.50	0.39	70.07%	70.59%	54.72%	12000	4090	34.08%
0.60	0.38	69.24%	75.91%	44.01%	12000	3059	25.49%
0.70	0.35	67.44%	80.74%	34.08%	12000	2227	18.56%
0.80	0.32	64.90%	86.24%	24.00%	12000	1468	12.23%
0.90	0.25	61.26%	91.86%	13.04%	12000	749	6.24%

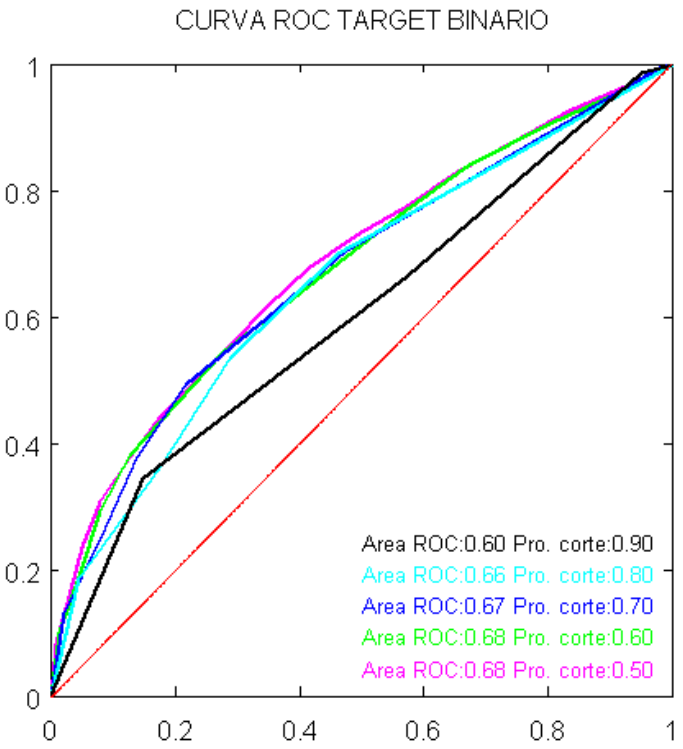
Si analizamos los resultados y los comparamos con los obtenidos para la regresión logística vemos que:

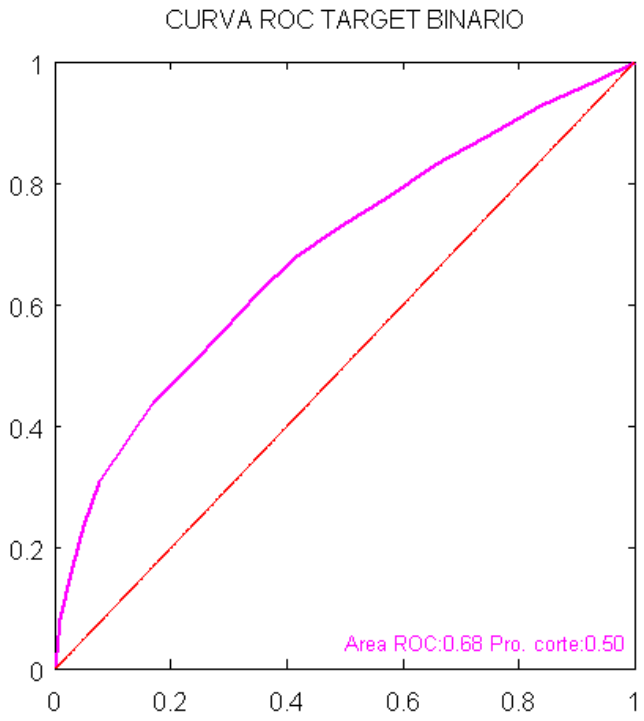
- El indicador de correlación de Matthews es mejor en la red neuronal 0.38 frente a 0.35, si bien no hay una gran diferencia.
- Los valores de precisión para 0.5 son mejores en la regresión logística que en la red neuronal, 70% de acierto en la red frente a 73,43% de unos bien clasificados.
- En cuanto a recall tenemos que la red es considerablemente mejor que la regresión, ya que consigue una captura del 55.65% frente al 42.85 de la logística.

En el siguiente gráfico podemos ver la evolución del error cometido por la red neuronal en cada una de las iteraciones de ajuste. El aspecto es el habitual, no parece haber comportamientos extraños si bien el error en el ajuste es considerable.



La curva ROC de la red neuronal presenta valores algo mejores que la regresión logística, el área debajo de la curva es ligeramente superior, pasando de un 0.65 en la logística a un 0.68 en la red.


CURVA ROC TARGET BINARIO
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
Area ROC:0.60 Pro. corte:0.90
Area ROC:0.66 Pro. corte:0.80
Area ROC:0.67 Pro. corte:0.70
Area ROC:0.68 Pro. corte:0.60
Area ROC:0.68 Pro. corte:0.50


CURVA ROC TARGET BINARIO
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
Area ROC:0.68 Pro. corte:0.50

6.1.2.- Recargas, planteamiento quincenal

Una vez que hemos comprobado que el ajuste mensual, tanto por parte de la regresión logística como por parte de la red neuronal presentan la suficiente calidad como para poder ser usados por las unidades de negocio pasaremos a probar el ajuste en el siguiente grado de la variable target, la clasificación multitarget para el ajuste de pares de semanas que puede tomar los valores.

$$y_i = \begin{cases} 1 \text{ si } l_i \text{ recarga en las dos primeras semanas.} \\ 2 \text{ si } l_i \text{ recarga en la tercera o cuarta semana.} \\ 3 \text{ si } l_i \text{ recarga no recarga en los 28 días.} \end{cases}$$

Para más información sobre cómo se ha realizado el ajuste multitarget acudir al anexo.

Los datos iniciales presentan una distribución para la variable dependiente y_i de:

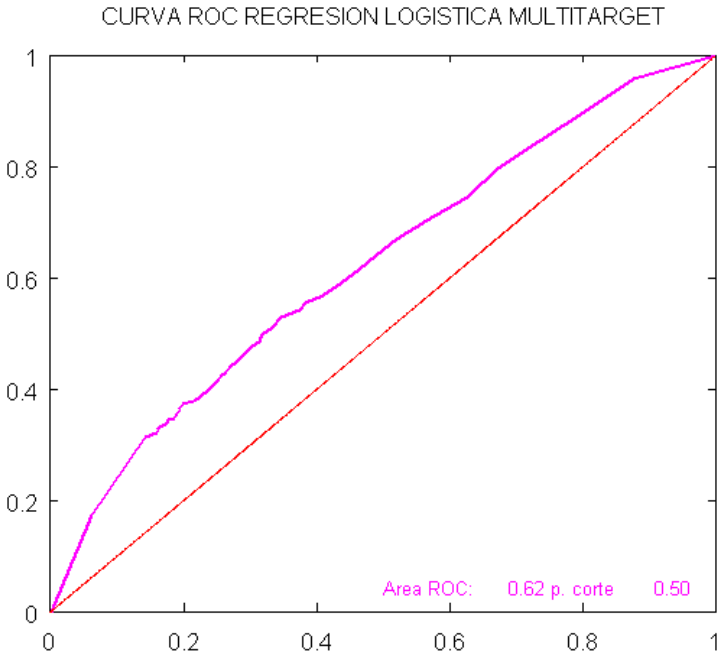
y_i	Porcentaje
1	27%
2	17%
3	56%
TOTAL	100%

Si analizamos la distribución de las recargas vemos que está desbalanceada, en concreto en el % de líneas que recargan las dos primeras semanas, frente al que lo hace en las dos últimas. Esto se debe a que en la construcción de la variable y_i estamos marcando solamente la primera recarga de la línea, por tanto si en las cuatro semanas el cliente recargara dos veces, la primera en las dos primeras semanas y la segunda en las dos últimas solo lo marcaríamos como {1}, esto origina que la base de clientes que pueden recargas en la categoría {2} sea menor que la base general de líneas, puesto que hay que descontar las que ya lo han hecho en {1}. Por tanto la categoría {2} lo que nos estará midiendo será la probabilidad de que una línea que no ha recargado las dos primeras semanas lo haga en las dos siguientes.

6.1.2.1.- Regresión logística

Vamos a empezar trabajando con un modelo logístico ajustado, con la muestra de entrenamiento sin balancear.

Obtenemos la curva ROC de ajuste del modelo:



Como podemos ver el área debajo de la curva ROC es muy inferior a la que teníamos en los modelos anteriores con el target binario (recarga en 28 días o no recarga). Si acudimos a la información del modelo multitarget tenemos:

P. CORTE	F1 SCORE	ACCURACY	PRECISION	RECALL	N	SELECCION	P SEL TOT
0.5	32.57	62.7%	62.3%	22%	12000	1866	15.55%

Para obtener más información sobre el F1 score, como medida de la calidad de ajustes, consultar anexos.

La precisión del modelo, ha bajado al 63,33%, frete a los valores de más del 70% que teníamos con el target "mensual".

Si analizamos la matriz de dispersión del modelo, tenemos:

		REAL		
		3	2	1
PREDICCIÓN	3	6360	1751	2023
	2	2	1	1
	1	362	338	1162

Podemos ver como el modelo se está centrando sobre todo en el ajuste de las etiquetas {1,3}, haciendo desaparecer, casi por completo, la etiqueta {2} en la predicción, que equivale a las líneas que efectúan la recarga en las dos últimas semanas del período contemplado en el target.

Para intentar mejorar esta situación recurriremos a balancear la muestra de entrenamiento.

6.1.2.2.- Regresión logística balanceada

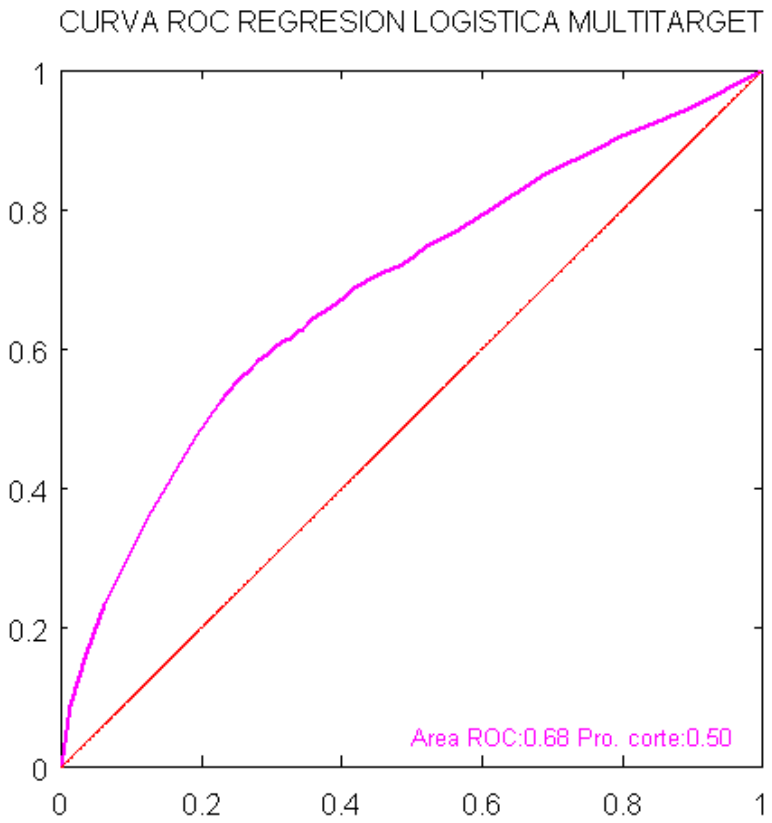
Modelo logístico balanceado al 30%

Modificamos la proporción inicial de negativo, en este caso con la etiqueta {3}, con el fin de estudiar si el algoritmo es capaz de ajustar, con mayor precisión, los positivos {1,2}.

En la siguiente tabla podemos ver cómo se distribuyen las líneas en cada una de las etiquetas para la muestra de entrenamiento balanceada y para la muestra de test, aplicando un balanceo del 30%.

CATEGORÍAS	% ENTRENAMIENTO	% TEST
1	43.09%	26.55%
2	27.03%	17.42%
3	29.88%	56.03%

La curva ROC generada por el modelo presenta unos valores considerablemente mejores que los del experimento inicial, pasando de 0.62 a 0.68, pero sin llegar a los valores de modelación obtenidos con el target “mensual”



Si analizamos los datos generales del modelo vemos que el índice de F1 score ha mejorado considerablemente, la precisión ha bajado pero recall se ha duplicado. El porcentaje de colectivo seleccionado ha aumentado de un 15% a un 46%.

P. CORTE	F1 SCORE	ACCURACY	PRECISION	RECALL	N	SELECCION	P SEL TOT
0.5	41.79	57.11%	40.53%	43.14%	12000	5616	46.80%

Si analizamos la matriz de dispersión podemos ver como se ha incrementado la clasificación de la clase {2}, si bien sigue estando muy por debajo de los porcentajes de la muestra de test.

		REAL		
PREDICCIÓN	0	3	2	1
	3	4577	890	917
	2	75	27	20
	1	2072	1173	2249

Modelo logístico balanceado al 20%

Probamos un balanceo menor para comprobar si ajusta mejor el modelo.

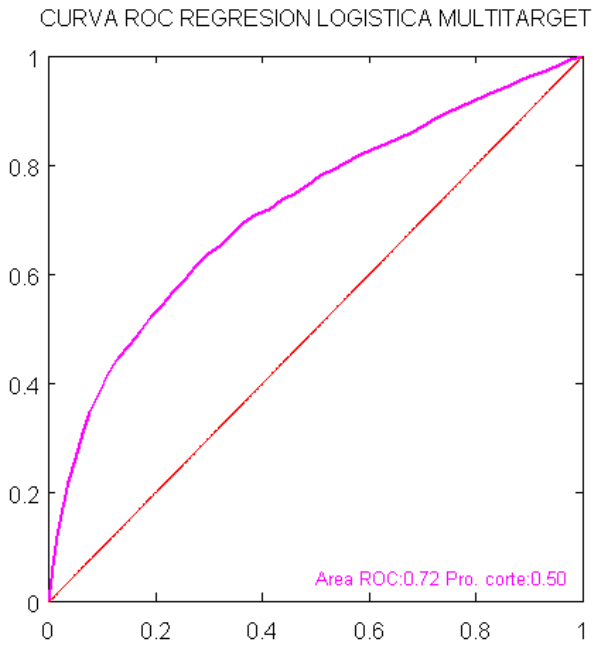
ENTRENAMIENTO	% ENTRENAMIENTO	% TEST
1	49.09%	26.55%
2	30.80%	17.42%
3	20.12%	56.03%

Hemos balanceado los negativos, etiqueta {3}, a un 20% del total de la muestra de entrenamiento, eliminando registros hasta alcanzarlo.

Los resultados son peores que con el balanceo del 30%, hay un gran aumento en recall debido a que estamos seleccionando casi el 95% de la muestra, esto impacta claramente en la precisión que se desploma.

P. CORTE	F1 SCORE	ACCURACY	PRECISION	RECALL	N	SELECCION	P SEL TOT
0.5	37.25	29.32%	27.28%	58.70%	12000	11351	94.59%

La curva ROC presenta valores mejores debido al gran aumento del % de recall.



La matriz de confusión ahora si muestra una distribución más uniforme en las tres categorías, si bien el error es muy elevado.

		REAL		
PREDICCIÓN	0	3	2	1
	3	421	102	126
	2	942	222	185
	1	5361	1766	2875

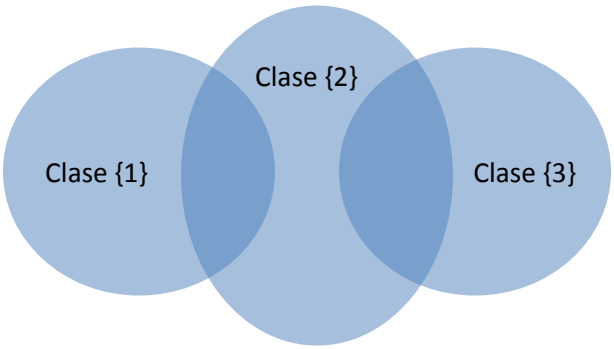
Al sobrebalancear la muestra de entrenamiento estamos provocando un gran aumento en la asignación de la categoría {1}, sobre todo frente a la {3}.

Los resultados no son utilizables operativamente.

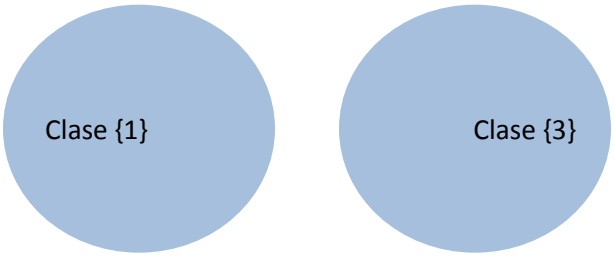
Modelo logístico última prueba

Vamos a realizar una última prueba con el modelo logístico. Al realizar la prueba inicial sobre el target quincenal que toma los valores {1,2,3} siendo {3} el valor correspondiente a los negativos, hemos verificado el ajuste para los unos y los ceros no era malo, pero presentaba problemas en la clase {2}.

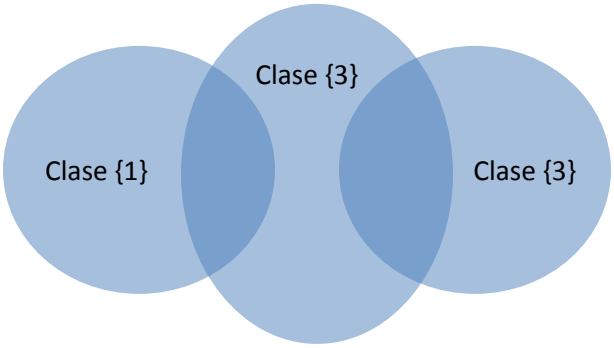
El modelo parece que está teniendo problemas para identificar la frontera entre las líneas que están en la clase {1}, recarga en las dos primeras semanas, y la clase {2}, recarga en las dos últimas semanas. El área de decisión podría ser de la forma:



Vamos a generar un conjunto de entrenamiento eliminando las líneas que estén marcadas con la clase {2}, de la forma:



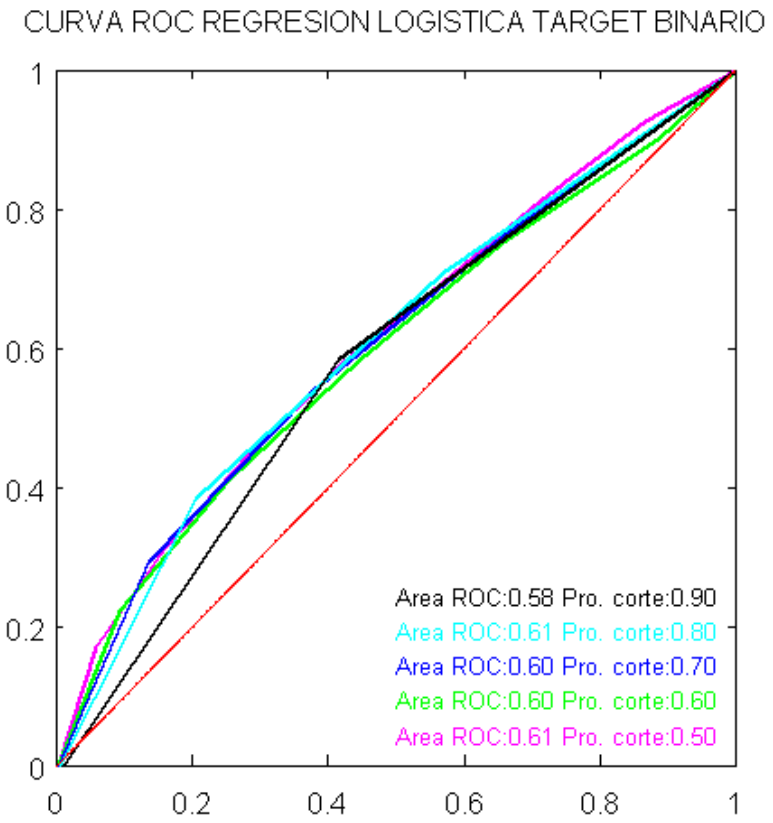
En cambio en la muestra de test tendremos una distribución del tipo:



Asignaremos en test los elementos de la clase {2}, a la clase {3}.

Con esto pretendemos conseguir que el algoritmo de entrenamiento trabaje con registros con las clases lo más limpias posible, de forma que no haya una frontera tan difusa entre ellas, con la esperanza de que al extrapolarlo en test ese aprendizaje ayude a diferenciar los grupos. Es de esperar que se produzca un reparto de la clase {2} entre la {1} y la {3}.

Las curvas ROC que obtenemos en función de la probabilidad de corte son muy similares entre sí, oscilando en áreas en torno a 0.6.



Los resultado del modelo directo, sin balancear

P. CORTE	MATTHEWS	ACCURACY	PRECISION	RECALL	N	SELECCION	P SEL TOT
0.50	0.35	77.37%	62.37%	37.19%	12000	1900	15.83%
0.60	0.33	77.23%	66.62%	28.56%	12000	1366	11.38%
0.70	0.29	76.69%	69.28%	21.94%	12000	1009	8.41%
0.80	0.26	76.17%	72.21%	16.64%	12000	734	6.12%
0.90	0.23	75.52%	77.56%	10.95%	12000	450	3.75%

Los resultados son claramente peores que el entrenamiento realizado con target binario para la variable de recarga en cuatro semanas {0,1}.

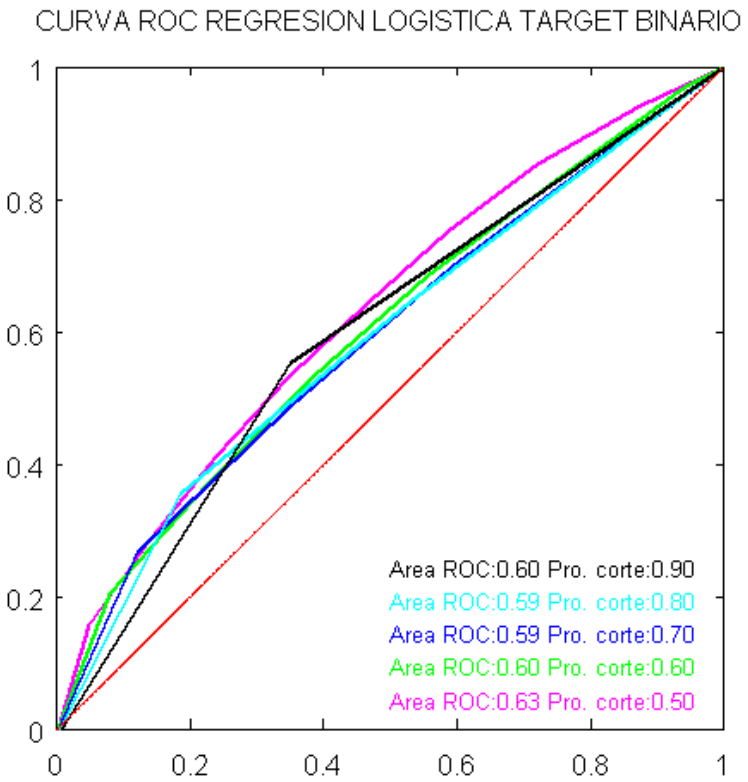
Vamos a probar con un balanceo de datos, dado que la proporción de la clase {0} en la muestra de entrenamiento es elevada y el nivel de recall obtenido es bajo.

Modelo logístico última prueba con balanceo

Parece que el balanceo no está aportando demasiado, una vez más la proporción inicial de {1} es suficientemente alta, está cercana al 30%, con lo que el balanceo no mejora los resultados iniciales, como ejemplo veamos qué ocurre si balanceamos los unos para contar con un 35% de ellos en la muestra de entrenamiento.

P. CORTE	MATTHEWS	ACCURACY	PRECISION	RECALL	N	SELECCION	P SEL TOT
0.50	0.35	76.99%	59.92%	40.30%	12000	2143	17.86%
0.60	0.33	77.31%	65.10%	31.32%	12000	1533	12.78%
0.70	0.30	76.89%	68.65%	23.85%	12000	1107	9.22%
0.80	0.27	76.33%	71.57%	18.02%	12000	802	6.68%
0.90	0.23	75.53%	75.51%	11.61%	12000	490	4.08%

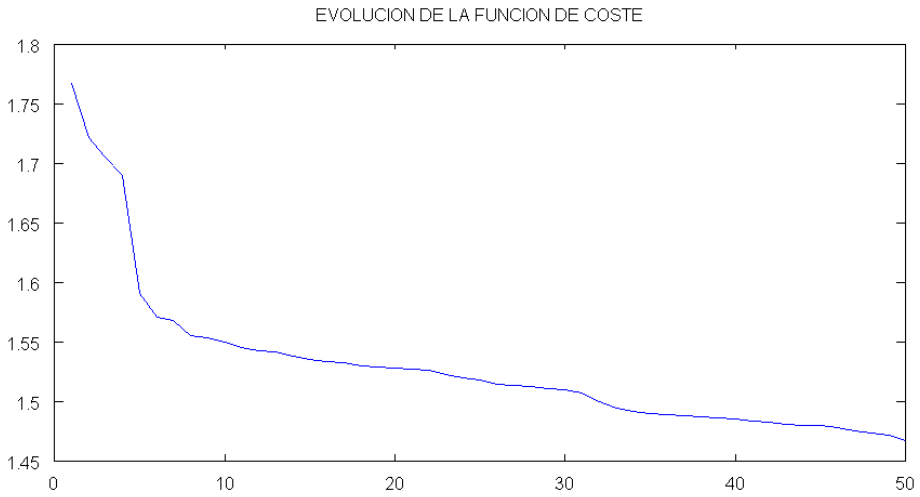
El mejor valor para la correlación de Matthews los obtenemos con el corte en probabilidad 0.5, con la precisión en el 60% y el recall en el 40%. Por tanto los resultados obtenidos con el target sin procesar son mejores.



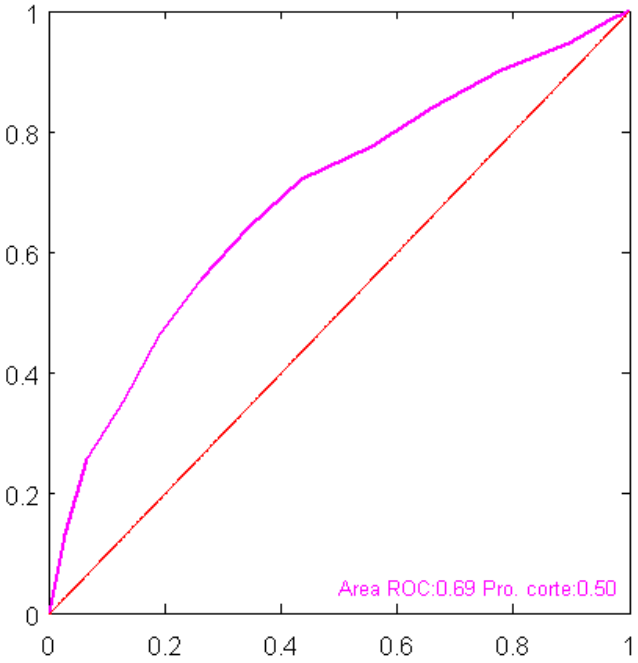
6.1.2.3.- Perceptrón multicapa

Se ha ejecutado la red neuronal con una configuración sin balanceo, para un target múltiple con las categorías {1,2,3}, donde el {3} representa la no recarga en las cuatro semanas de estudio, {1} la recarga en las dos primeras semanas, y {2} la recarga en las dos últimas.

La curva de ajuste del modelo a través de la función de coste presenta un aspecto usual y no se aprecian problemas.



La curva ROC presenta valores claramente mejores que los obtenidos en la regresión logística.



Los resultados generales del modelo son:

P. CORTE	F1 SCORE	ACCURACY	PRECISION	RECALL	N	SELECCION	P SEL TOT
0.5	38.64	63.81%	55.01%	29.78%	12000	2856	23.80%

La precisión presenta valores bajos y el recall muy bajos, si analizamos la matriz de confusión vemos que tenemos el mismo problema que en la regresión logística, el modelo tiene problemas a la hora de diferenciar la clase {2}.

		REAL		
PREDICCIÓN	0	3	2	1
	3	6086	1444	1614
	2	0	0	1
	1	638	646	1571

Nos decidimos por aplicar un balanceo a los 0 intentando que el modelo tenga mayor sensibilidad a la clase {2}.

6.1.2.3.- Perceptrón multicapa balanceado

Nos decidimos por aplicar un balanceo a los ceros intentando que el modelo tenga mayor sensibilidad a la clase {2}.

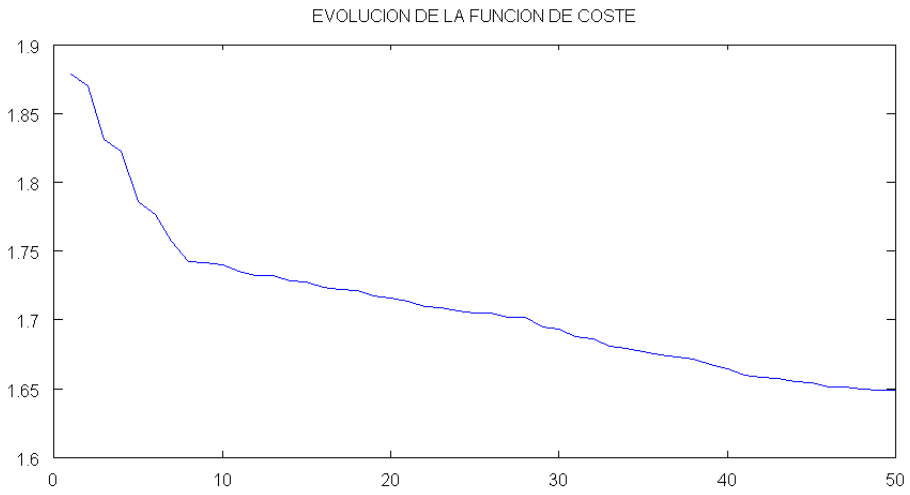
La distribución inicial de las clases en la muestra de entrenamiento se exponen en la siguiente tabla.

y_i	Porcentaje
1	27%
2	17%
3	56%
TOTAL	100%

Balancearemos la clase {3} al 30%.

y_i	Porcentaje
1	42.97%
2	26.96%
3	30.07%
TOTAL	100%

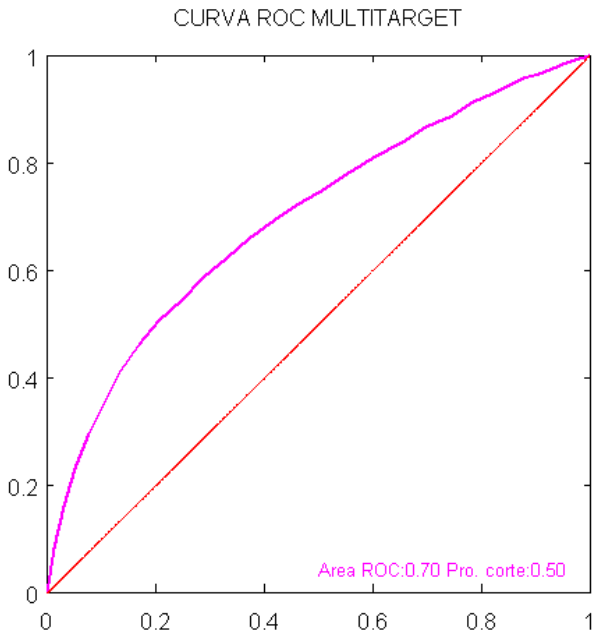
La función de coste evoluciona de la forma:



Los datos generales del modelo son:

P. CORTE	F1 SCORE	ACCURACY	PRECISION	RECALL	N	SELECCION	P SEL TOT
0.5	42.20	56.36%	39.47%	45.32%	12000	6057	50.48%

La curva ROC presenta valores mejores que la logística



La matriz de coincidencia presenta los datos:

		REAL		
PREDICCIÓN	0	3	2	1
	3	4372	779	792
	2	467	179	182
	1	1885	1132	2212

Los resultados siguen sin ser lo suficientemente buenos como para poder usar el modelo en campañas comerciales, al menos con el corte de probabilidad en 0.5.

CONCLUSIONES

A la vista de los resultados no tiene sentido seguir profundizando en la granularidad del target, al menos empleando solamente la variable recargas, volveremos a hacer el mismo ejercicio con la variable saldo y con la combinación de las dos, con la intención de analizar el punto crítico hasta el que podemos descender con el modelo manteniendo unas garantías de acierto y recall.

6.2.- Reducción de dimensionalidad

Se ha estudiado la posibilidad de aplicar un PCA (Análisis de Componentes Principales) a los datos para reducir su dimensión, esto es condensar la variabilidad de la información de entrada (variables independientes) en un número menor de variables.

Dada la particularidad del problema se ha descartado, dado que:

- Trabajamos con variables temporales, cada variable de entrada equivale a un día.
- El comportamiento en un día de una línea puede no tener nada que ver con el de otra y por tanto cada variable (día) tienen sentido dentro del vector de variables de una línea pero no lo tiene al ser comparado ese día particular con el del resto de líneas, obviando el resto de la serie.
- En caso de realizar una estandarización de la información, a nivel de registro, en vez de a nivel de variable para introducir la información en el PCA, nos encontraríamos con una posible reducción de información a nivel temporal, pero esto conllevaría asumir que dicha condensación se mantendrá cuando ejecutemos en el futuro el modelo. Esto solo sería viable si en nuestro proceso no existiera estacionalidad ni continuidad temporal, esto es, que nos pudiéramos colocar siempre en el inicio de la serie (por ejemplo un electrocardiograma).
- Asumiríamos que en todo momento t en el que se ejecutara el modelo el comportamiento de los X180 días anteriores sería el mismo que en el entrenamiento (lunes del entrenamiento coincidiría con lunes en ejecución), y esto en un modelo que se deberá ejecutar, como poco, una vez cada mes no tiene sentido, puesto que estaríamos asumiendo que el comportamiento de los 180 días anteriores al mes de enero es el mismo que, por ejemplo, los 180 anteriores al mes de septiembre.

6.3.- Segmentación

6.3.1.- Segmentación sobre recargas

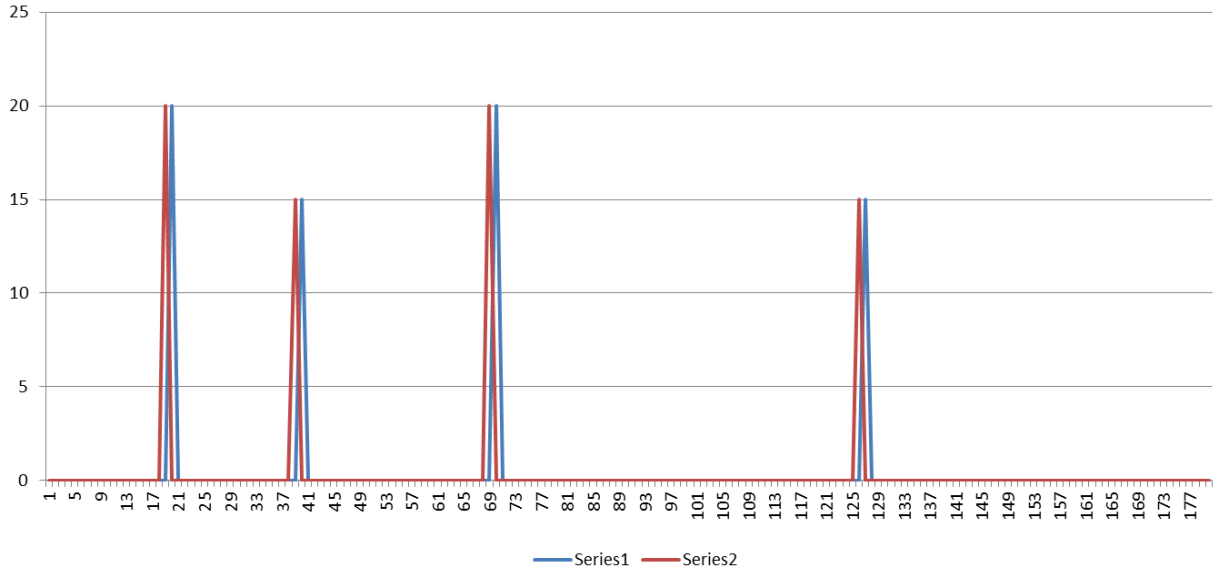
El proceso de segmentación consiste en la agrupación, no supervisada, de los individuos de la muestra [5] [6].

El objetivo sería identificar comportamientos similares entre los distintos patrones de recargas, una vez identificados procederíamos a modelizar el comportamiento futuro de cada uno de los grupos por separado, reduciendo la variabilidad que recibe el algoritmo en el entrenamiento.

El problema surge a la hora de elegir la distancia para realizar la evaluación de las distintas líneas al ejecutar un algoritmo de segmentación, por ejemplo un k-means.

Hay que tener en cuenta que las medidas habituales, por ejemplo la euclídea no tienen en cuenta los posibles desplazamientos, aunque sean pequeños de la serie de datos.

Pongamos un ejemplo:



Como podemos ver las series, que se muestran en el gráfico, son iguales, con la particularidad de que una de ellas tiene un desplazamiento de un día. El comportamiento de las dos series en cuanto a la distancia entre recargas y el importe recargado es exactamente igual, con un retardo de un días entre ellas.

Si sobre las dos series aplicamos una distancia euclídea, día a día, obtendríamos una distancia entre ellas muy elevada, que no se ajusta a la realidad.

La distancia euclidea para las dos series de referencia seria 50 unidades mientras que la distancia DTW sería de cero.

En nuestro caso es muy interesante identificar estas dos series como una sola, puesto que un retardo pequeño, como el del ejemplo, es de suponer que conllevará una diferencia pequeña en el momento de la recarga y por tanto el ajuste del modelo de predicción para ambas series podría ser el mismo.

Para resolver el problema de la similitud entre este tipo de series emplearemos de Alineamiento Temporal Dinámico (DTW), para más información ver Anexos.

El objetivo con la técnica de k-means con distancia DTW es encontrar grupos homogéneos, identificar sus centroides y emplearlos como series modelo para clasificar los nuevos registros que nos lleguen.

Dado que el cálculo de la distancia DTW es muy pesado vamos a realizar una primera aproximación con 1.000 registros aleatorios de la muestra de entrenamiento y con tres clusters.

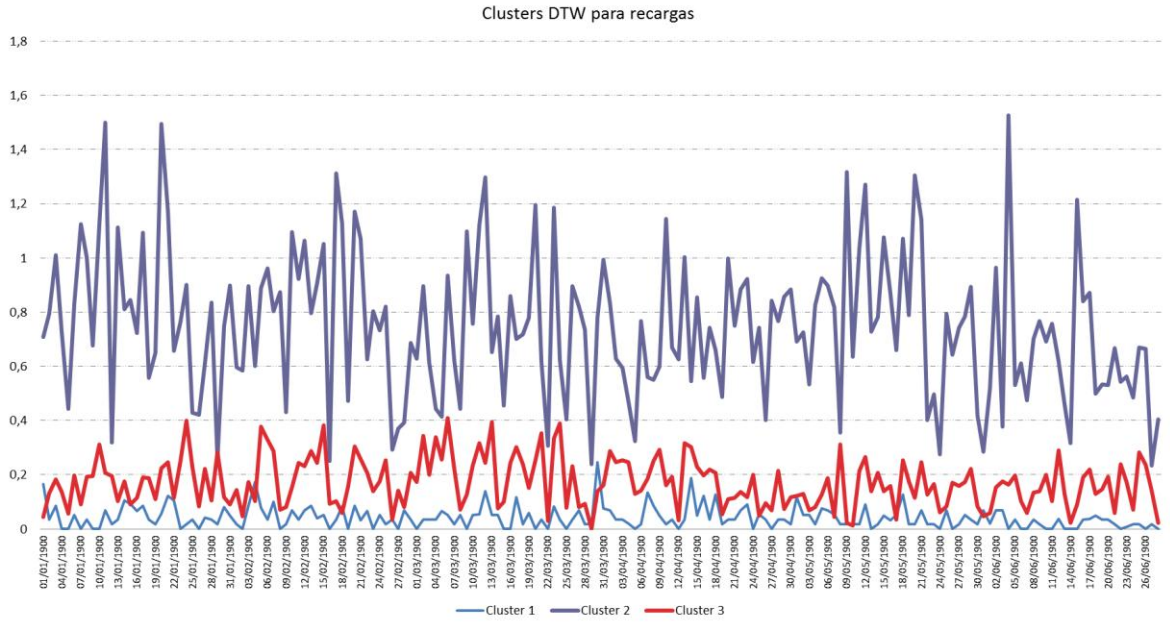
El objetivo será compara los resultados obtenidos mediante la técnica DTW, para cada uno de los clusters creados frente a los resultados obtenidos al realizar la modelación con toda la muestra sin segmentar, y contra los resultados obtenidos con distancia euclídea.

6.3.2.- Segmentación sobre recargas DTW

Se han realizado diferentes pruebas de segmentación con tres, cuatro y cinco grupos, iterando cada una de las opciones cinco veces para evitar posibles problemas en el algoritmo k-means al hacer la selección aleatoria inicial de los centroides de partida.

Hemos decido empezar por el caso más simple, que es el de tres clusters, además en este caso ya se observa que dos de los grupos aparecen relativamente juntos lo que conlleva una pobre separación entre los casos que se asignan a cada uno.

Si observamos la gráfica de los centroides medios para cada uno de los clusters tenemos:



Podemos ver como el cluster 2 está claramente diferenciado del 1 y el 3. Las gráficas presentan una gran aleatoriedad con medias estables y sin tendencia aparente, la varianza parece mantenida en el 1 y el 3, en el 2 hay mayores diferencias.

Los valores parecen muy fluctuantes y se asemejan a un ruido blanco, lo que no es una buena señal a la hora de ser empleados para la agrupación del resto de las series.

Una vez calculados los centroides clasificamos cada uno de los registros con la distancia DTW, tanto para la muestra de entrenamiento, validación como la de test.

	%	Media	Std	% Imp Rec
Cluster 1	42,70%	12,7	9,5	11,00%
Cluster 2	17,10%	109	109	54,53%
Cluster 3	40,20%	45,5	20,99	34,46%
Total	100%	53,3	73,9	100%

Se puede observar como el Cluster 1, asociado a las líneas que hacen menos recargas, es el que acumula una mayor cantidad de casos, en concreto más del 42%. Esto responde a la distribución de las recargas en el negocio, la mayoría del parque de prepago hace pocas recargas, se mantienen con el mínimo importe recargado durante los meses de validez del saldo.

El Cluster 2 corresponde a las líneas con una mayor actividad en la recarga, es el que agrupa un menor número de líneas, tan solo un 17%. La detección de estos clientes tienen interés por

sí misma para el negocio, dado que son clientes VIP en prepago, hacen muchas recargas y con importe superior a la media.

Por último el Cluster 3, muy parecido al 1, presenta un 40% de las observaciones.

Si atendemos a los valores de las recargas medias para los individuos de cada cluster podemos ver como el cluster 2 presenta un valor muy superior a la media de los tres y al 1 y el 3. El cluster 1 presenta valores muy por debajo de la media y el dos es el que más cerca está de la media.

Las desviaciones típicas son muy elevadas en el cluster 2.

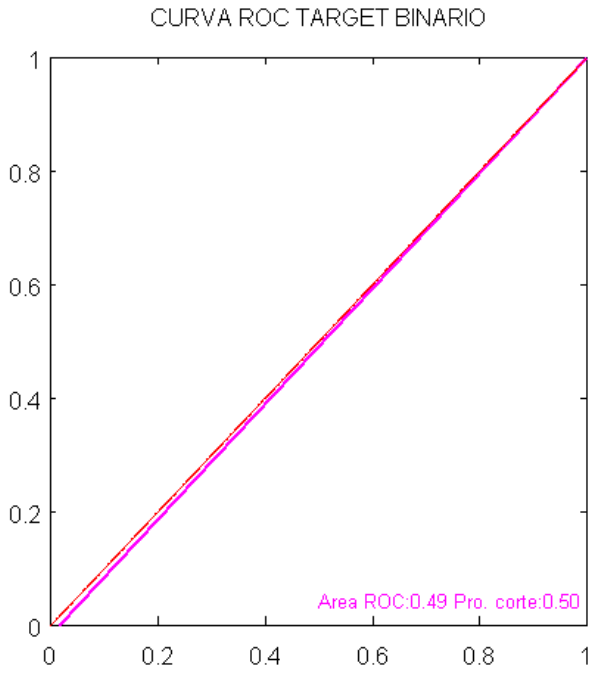
Nota de negocio: el cluster 2 tienen una especial relevancia para el negocio dado que, a pesar de contar tan solo con el 17% de las líneas, concentra el 54% del importe recargado, por tanto cualquier acción comercial sobre este colectivo puede tener un importante impacto en los resultados.

6.3.3.- Clasificación sobre clusters DTW

6.3.3.1.- Regresión logística

Cluster 1

Vamos a realizar un ajuste de regresión logística sobre las líneas clasificadas en el cluster 1, de la misma forma que lo hemos realizado anteriormente para toda la muestra.



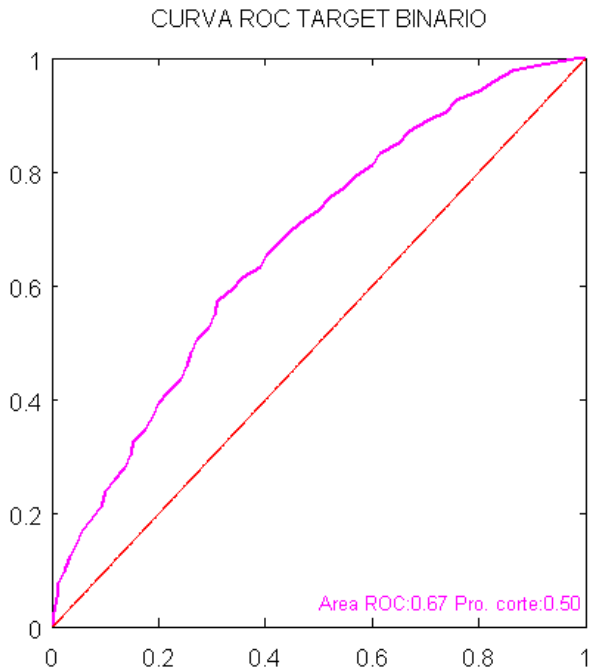
El ajuste obtenido consigue aciertos iguales a lo aleatorio, lo que nos indica que este grupo de clientes no tienen un comportamiento predecible, con la metodología que estamos empleando. Este colectivo se caracteriza, como ya hemos visto en la tabla resumen, por hacer pocas recargas de bajo importe y con un patrón pseudoaleatorio.

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.05	72.89%	43.27%	3.31%	5065	104	2.05%
0.60	0.03	73.05%	41.18%	1.03%	5065	34	0.67%
0.70	-0.01	73.05%	20.00%	0.15%	5065	10	0.20%
0.80	-0.01	73.15%	0.00%	0.00%	5065	1	0.02%
0.90	NaN	73.17%	NaN%	0.00%	5065	0	0.00%

Los valores de ajuste del modelo son aleatorios, no conseguimos explicar el comportamiento de los datos.

Cluster 2

Procedemos a realizar el ajuste de la regresión logística sobre las líneas pertenecientes al cluster 2.

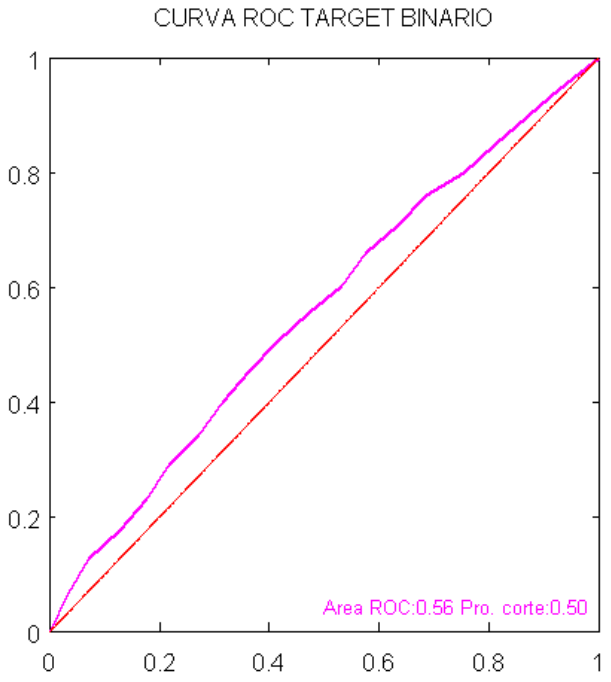


P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.09	78.40%	78.81%	99.08%	2079	2048	98.51%
0.60	0.23	78.98%	81.14%	95.33%	2079	1914	92.06%
0.70	0.22	71.04%	83.94%	77.96%	2079	1513	72.78%
0.80	0.19	54.02%	87.26%	48.37%	2079	903	43.43%
0.90	0.13	34.54%	90.61%	18.35%	2079	330	15.87%

Los resultados de ajuste del cluster 2 parecen mucho mejores, no obstante hay que tener en cuenta que en este cluster la probabilidad a priori de 1 es del 78%, por tanto en el corte de 0.5 estaríamos en un acierto aleatorio. La correlación de Matthews en el caso del ajuste de regresión logística para toda la muestra, sin clusters, es del 0.35, muy por encima del 0.09 que obtenemos.

Cluster 3

Realizamos el ajuste de la regresión logística para el cluster 3.



Los valores de la curva ROC no son buenos, estamos por encima de lo aleatorio pero lejos de los valores del ajuste general del problema.

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.20	60.48%	59.70%	49.65%	4856	1903	39.19%
0.60	0.16	58.22%	64.03%	25.83%	4856	923	19.01%
0.70	0.10	55.00%	67.46%	8.70%	4856	295	6.07%
0.80	0.06	53.46%	78.00%	1.70%	4856	50	1.03%
0.90	-0.02	52.82%	0.00%	0.00%	4856	3	0.06%

Nota de negocio: la identificación de las líneas que pertenecen al cluster 2 puede tener gran interés a la hora de plantear acciones comerciales, dado que la probabilidad a priori de recarga en el mes siguiente está en el 78%.

6.3.3.2.- Perceptrón multicapa

La presencia a priori de unos es del 47%, en la muestra de entrenamiento, por tanto la predicción en este cluster si mejora lo aleatorio.

Los resultados empleando el Perceptrón Multicapa son similares a los obtenidos con la regresión logística. En concreto los resultados para el Cluster 2 son:

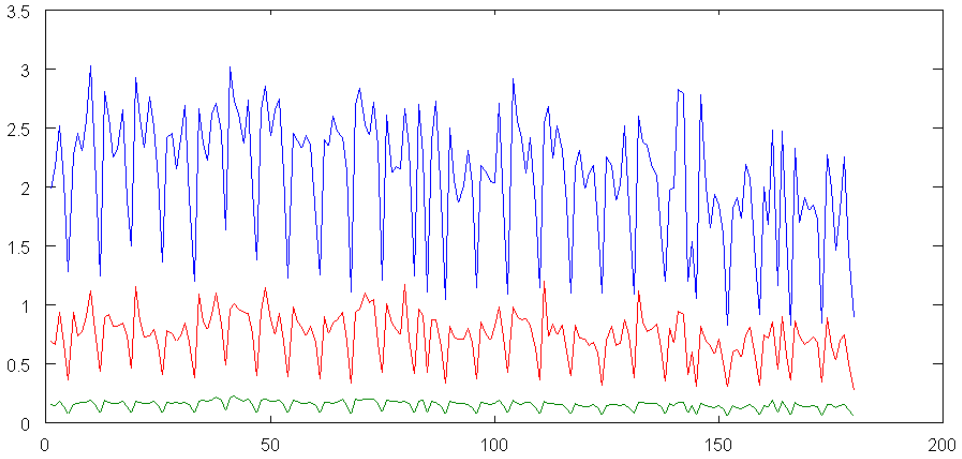
P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.11	76.00%	79.67%	93.12%	2079	1904	91.58%
0.60	0.13	73.59%	80.65%	87.23%	2079	1762	84.75%
0.70	0.12	67.29%	81.28%	75.69%	2079	1517	72.97%
0.80	0.11	57.53%	82.43%	58.20%	2079	1150	55.32%
0.90	0.11	40.16%	86.25%	28.12%	2079	531	25.54%

6.3.4.- Clasificación sobre cluster distancia euclídea

En los puntos anteriores se ha empleado la distancia DTW, sin obtenerse los resultados esperados de mejora en la clasificación. Además, el uso de DTW añadiría una gran dificultad a la hora de hacer el paso a producción del modelo, por lo que se ha decidido hacer una última prueba de segmentación, pero solo con la distancia euclídea que simplifica considerablemente la operativa, al no haber encontrado, anteriormente, una diferencia significativa en los resultados al usar las dos métricas.

6.3.4.1.- Regresión logística

Vamos a desarrollar los mismos pasos que en el estudio del k-means con distancia DTW pero aplicando en esta ocasión la distancia euclídea.

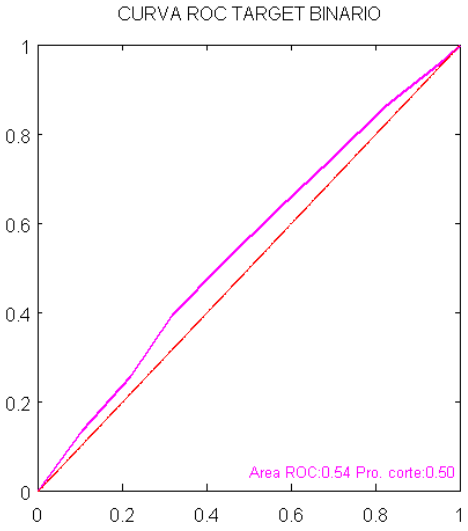


Los patrones son similares a los obtenidos empleando el cluster con distancias DTW

	%	Media	Std	% Imp Rec
Cluster 1	2,53%	374,7	143,3	17,77%
Cluster 2	80,63%	26,85	20,65	40,61%
Cluster 3	16,84%	131,77	46,27	41,62%
Total	100%	53,3	73,9	100%

Comprobamos como con la distancia euclídea los clusters se distribuyen peor que con la DTW, los casos aparecen concentrados, mayoritariamente, en el cluster 2.

Vamos a realizar una prueba de ajuste para la regresión logística con los casos del cluster 2.



La proporción de positivos en la muestra para las líneas del cluster 1 es del 90%, si bien este cluster tan solo selecciona un 2,53% de la muestra inicial, y por tanto un número muy bajo del total de líneas que van a recargar.

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.31	89.39%	92.12%	96.42%	311	292	93.89%
0.60	0.32	87.78%	92.88%	93.55%	311	281	90.35%
0.70	0.33	85.85%	93.68%	90.32%	311	269	86.50%
0.80	0.24	77.49%	93.72%	80.29%	311	239	76.85%
0.90	0.20	66.24%	94.39%	66.31%	311	196	63.02%

6.3.5.- Conclusión clasificación sobre segmentación

La estrategia de segmentar las líneas, tanto mediante la distancia euclídea como con DTW, no ha servido para mejorar la predicción general. Como nota positiva y útil de cara al negocio podemos destacar que se han encontrado grupos de clientes, mediante los dos sistemas que presentan una probabilidad de recarga en las próximas 4 semanas del doble de lo aleatorio, si bien para conjuntos pequeños de clientes.

Estos grupos podrían servir para la realización de una acción comercial puntual.

A la vista de los resultados continuaremos con la estrategia de modelado simple, sin crear subgrupos y distintos modelos de estimación para cada uno de ellos.

7.- Aproximación a través de los saldos

Hemos recreado, a partir de diferentes fuentes de información, la variable saldo diario de la línea. Se ha generado el importe pendiente de gasto del que dispone la línea a cierre de cada uno de los noventa días analizados.

Debido a la multiplicidad de datos históricos la consolidación se ha realizado con una longitud temporal de 90 días, dado que no era posible contar con todas las fuentes de datos para un horizonte temporal mayor.

Por tanto para cada una de las líneas de la muestra, entrenamiento, validación y test, disponemos de la información de saldo para los 90 días anteriores al período de target.

El planteamiento de partida para la construcción de esta variable, que implica un gran coste de proceso, es la idea de que las variaciones del saldo pueden determinar el momento de la recarga de la líneas.

Podemos entender que si una línea dispone de un alto saldo no realizará recarga, así como que líneas con saldo no muy alto pero que se ha mantenido durante días puede que tampoco la realicen. En cambio líneas que en los últimos días han sufrido una importante disminución del saldo, cuando su costumbre es no superar ciertos umbrales tendrán una mayor probabilidad de recargar.

7.1.- Trabajando con saldos

Para comenzar vamos a analizar la variable saldo frente al target, para comprobar si nuestras hipótesis de partida se confirma o no.

La primera H0 será:

El saldo en el último día del estudio influye en la probabilidad de recarga.

Para comprobarlo vamos a montar una tabla con las probabilidades de recarga en los siguientes 28 días en función del saldo el día Z90.

	0	1	TOTAL	0	1
0-1	5727	8636	14363	40%	60%
1-5	7959	9294	17253	46%	54%
5-10	9240	6326	15566	59%	41%
10-20	9841	4265	14106	70%	30%
>20	6652	2173	8825	75%	25%
TOTAL	39419	30694	70113	56%	44%

En la tabla podemos ver como el dato del último días antes del período de target nos identifica, sobre todo, los clientes que no van a recargas. Si nos fijamos en los clientes con saldo mayor de 10 euros, que son más del 40% del total vemos que su probabilidad de no hacer recarga está por encima del 70%.

Los clientes que presentan saldo inferior a 5€ tienen una probabilidad ligeramente superior a la media de recargar, la diferencia se hace más significativa en los que tienen saldo de menos de 1€, pasando de un 44% de recargas para todo el colectivo a un 60%.

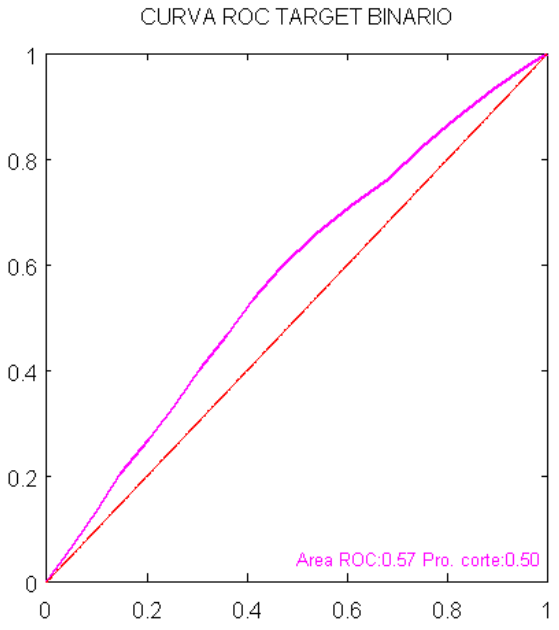
En cualquier caso parece claro que hay muchos clientes que se mantienen en planta del Operador con saldo a cero o casi, por lo que no se puede generar una regla, con alta fiabilidad, a partir del saldo del último día que nos permita determinar si el cliente va a recargar o no.

Recordemos que nuestro objetivo es identificar aquellos clientes que con alta probabilidad vayan a recargar en los próximos días.

Es probable que realizando algún tipo de análisis de la evolución del saldo en los días anteriores al target, podemos mejorar la discriminación.

7.2.- Regresión logística

Vamos a realizar un intento de modelar la variable target {recarga en los siguientes 28 días} introduciendo la información de la variable saldo, con sus 90 días disponibles.



La curva ROC ofrece unos resultados muy pobres.

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.28	65.11%	63.58%	48.31%	12000	4009	33.41%
0.60	0.12	58.12%	70.83%	8.06%	12000	600	5.00%
0.70	0.05	56.41%	66.42%	1.72%	12000	137	1.14%
0.80	0.02	56.10%	60.00%	0.45%	12000	40	0.33%
0.90	0.02	56.09%	73.33%	0.21%	12000	15	0.12%

Los resultados obtenidos son pobres, tanto en precisión como en recall.

A través de pruebas iterativas se ha comprobado que con menos variables de entrada el modelo se comporta mejor, el ajuste con menos error lo obtenemos usando para la predicción el saldo de los 10 últimos días.

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.31	66.58%	67.89%	45.53%	12000	3538	29.48%
0.60	0.09	56.97%	74.89%	3.22%	12000	227	1.89%
0.70	0.04	56.20%	76.32%	0.55%	12000	38	0.32%
0.80	0.02	56.09%	76.92%	0.19%	12000	13	0.11%
0.90	0.02	56.07%	85.71%	0.11%	12000	7	0.06%

La precisión mejora en mayor medida de lo que se reduce el recall, vemos como la correlación de Matthews pasa de 0.28 a 0.31.

Aunque el modelo es mejor la aportación de forma individual del saldo, más si tenemos en cuenta el coste de cálculo, es pobre.

Se han probado diferentes balanceos sin obtener mejoras significativas.

7.3.- Reducción de dimensionalidad

Aplican los mismos comentarios y conclusiones que en el estudio realizado para el importe recargado.

7.4.- Segmentación

Los resultados sobre saldos para la segmentación tanto euclídea como usando DTW son similares, en cuanto a su efectividad, a los obtenidos en la prueba realizada con las variables de recargas.

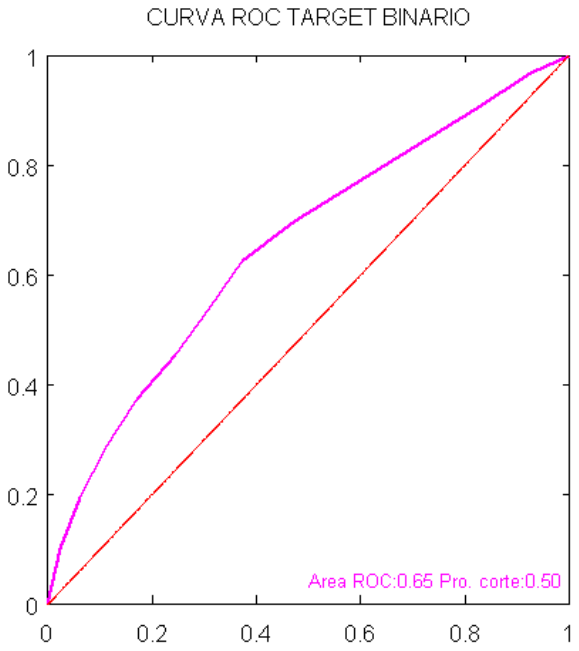
8.- Aproximación mediante recargas semanales

En este apartado vamos a realizar una prueba de ajuste del modelo reduciendo la variabilidad total de las variables independientes de recargas, inicialmente para 180 días, agrupando la información en totales recargados semanales. El objetivo es estudiar si el modelo es capaz de extraer un patrón general de mayor calidad al disponer de información con menos ruido.

Agruparemos los días en 25 semanas, desde el 18/11/2012 hacia atrás, tomadas de lunes a domingo. Los 5 primeros días no son agrupables en una semana completa y por tanto se prescinde de ellos. Del total de los 180 días hemos agrupado 175.

8.1.- Regresión Logística

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.35	68.20%	73.76%	42.95%	12000	3072	25.60%
0.60	0.33	66.05%	80.05%	30.34%	12000	2000	16.67%
0.70	0.28	63.60%	83.58%	21.42%	12000	1352	11.27%
0.80	0.25	61.56%	87.21%	14.73%	12000	891	7.42%
0.90	0.20	59.57%	90.30%	9.00%	12000	526	4.38%



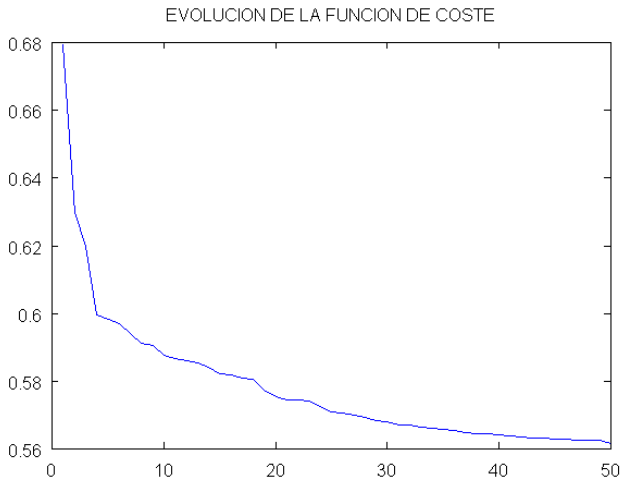
Los resultados del ajuste logístico son muy similares a los del problema sin reducción, esto es, empleando las 180 variables iniciales de recargas por día. Por tanto no aporta valor, si bien a

la hora de realizar la implementación puede resultar más sencilla y reduce el espacio necesario en la base de datos, dado que pasamos de una variable por día a una variable por semana.

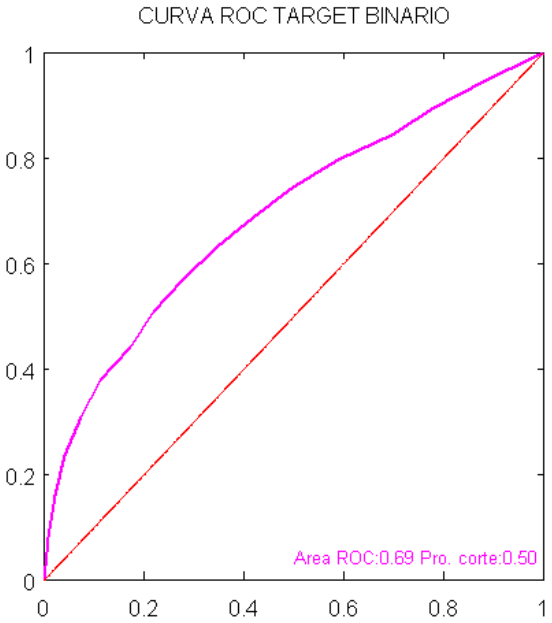
8.2.- Perceptrón multicapa

Hacemos la prueba para el Perceptrón multicapa entrenando con los datos de las recargas agrupadas por semanas, frente al target de recarga en el mes siguiente.

La función de coste presenta una evolución adecuada.



Tenemos una curva ROC con uno de los mejores ajustes obtenidos.



La tabla de resultados muestra valores de correlación de Matthews de los mejores que se han obtenido, solo comparables a los alcanzados con la combinación de las dos variables, recarga y saldos.

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.42	71.76%	73.80%	55.46%	12000	3965	33.04%
0.60	0.40	70.26%	79.48%	43.61%	12000	2895	24.12%
0.70	0.37	67.84%	84.82%	32.71%	12000	2035	16.96%
0.80	0.32	64.73%	90.03%	22.25%	12000	1304	10.87%
0.90	0.24	60.80%	95.25%	11.41%	12000	632	5.27%

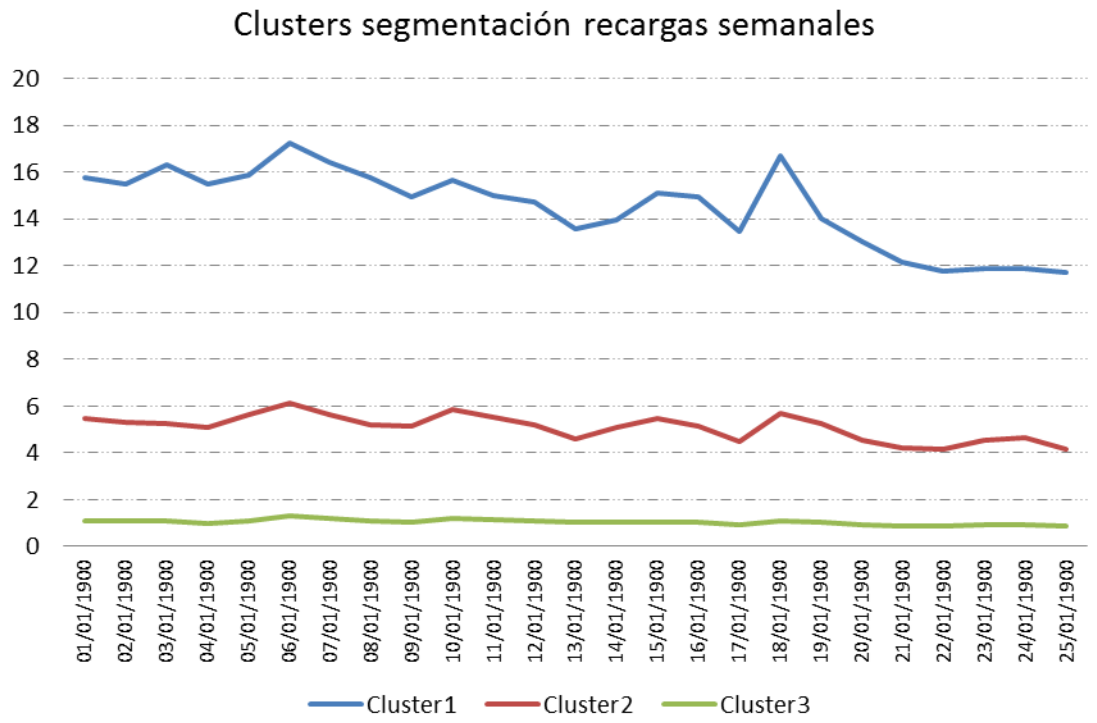
Los resultados del Perceptrón multicapa para la agrupación de variables semanales son buenos, tanto en precisión como en recall.

Por tanto parece que el modelado con información a nivel semanal aporta ventajas frente al trabajo con variables diarias, al menos en las variables de recargas.

8.3.- Segmentación

Al igual que en el caso anterior vamos a probar de nuevo la segmentación reduciendo el número de variables, una reducción de la variabilidad de la muestra podría ayudar al algoritmo a encontrar grupos más claramente diferenciados.

Se han realizado 10 iteraciones del algoritmo k-means sobre el conjunto de datos de las recargas agregadas a nivel semanal, hemos seleccionado el que presentaba una mejor distribución de los centroides, medias y desviaciones, si bien salvo en un par de ocasionas los valores han sido muy similares.



Aplicamos sobre los segmentos los modelos

	%	Media	Std	% Imp Rec
Cluster 1	2,7%	362	141	19%
Cluster 2	17,3%	127	43	42%
Cluster 3	80,0%	26	20	39%
Total	100%	53,3	73,9	100%

Si bien es cierto que la segmentación no está equilibrada, dado que el primer cluster contienen solo el 2,7% de las observaciones, hemos probado para dos clusters y para cuatro y los resultados son claramente peores, formando más grupos pequeños que casi no contienen líneas.

8.3.1- Cluster 1

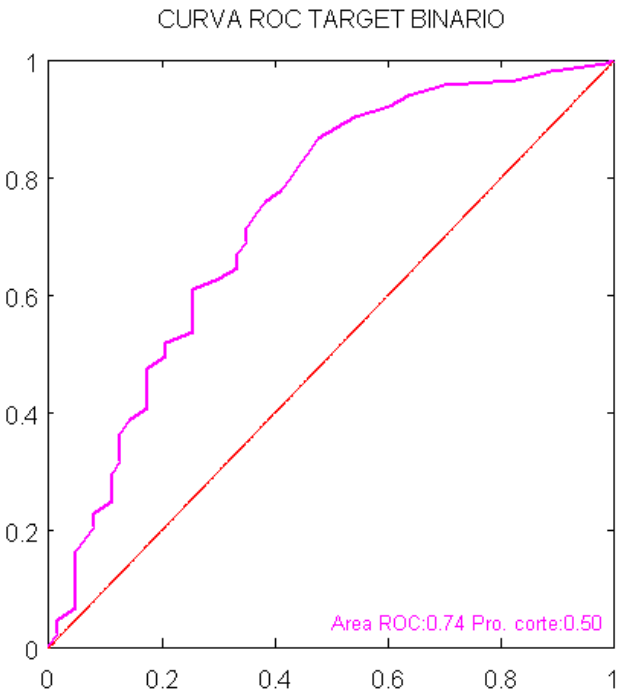
Vamos a realizar una prueba de ajuste para estos tres clusters con la regresión logística.

Hay pocos registros en la muestra de entrenamiento y test, por tanto los valores hay que analizarlos con precaución.

Hay que destacar que el 87,52% de la muestra recarga en los siguientes 28 días. Por tanto el ser capaz, simplemente, de identificar este segmento nos da un acierto, para estos individuos

muy alto. Es evidente que es un colectivo de usuarios de consumo muy alto y por tanto con una gran necesidad de recargas para mantenerlo.

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.36	89.26%	89.71%	99.10%	633	612	96.68%
0.60	0.36	88.31%	90.82%	96.39%	633	588	92.89%
0.70	0.42	87.36%	92.78%	92.78%	633	554	87.52%
0.80	0.38	81.83%	94.16%	84.48%	633	497	78.52%
0.90	0.25	61.14%	95.29%	58.48%	633	340	53.71%

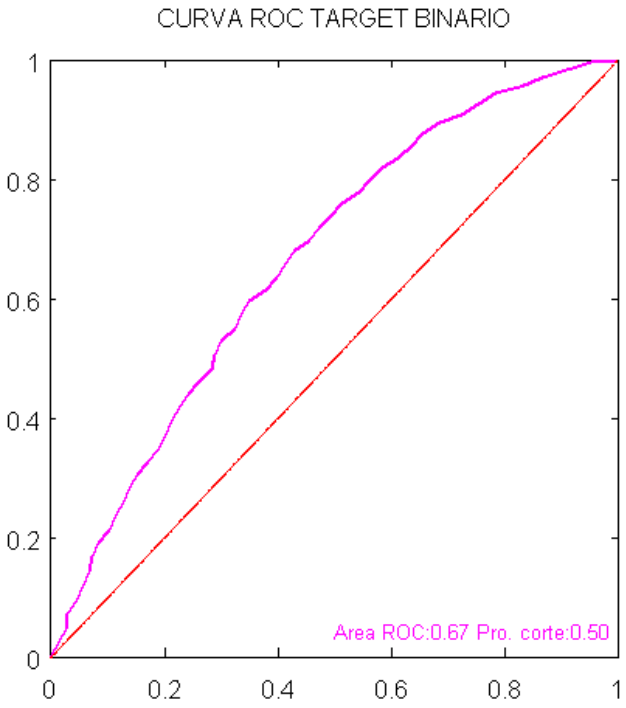


Conclusión de negocio: Es evidente que este colectivo de clientes aun siendo solo el 2,7% de la población presentan un comportamiento claramente diferenciado del resto, son líneas prepago de muy alto consumo y con una frecuencia de recarga elevada. Sería interesante analizar la posibilidad de migrarlos a contrato puesto que con consumos tan elevados son muy vulnerables, si bien es cierto que siempre han existido pequeños colectivos de prepago de consumos muy elevados y que no han querido históricamente pasar a contrato, muchas veces motivado por el tipo de uso de la línea, en algunas ocasiones asociada a fraude, locutorios…

8.3.1- Cluster 2

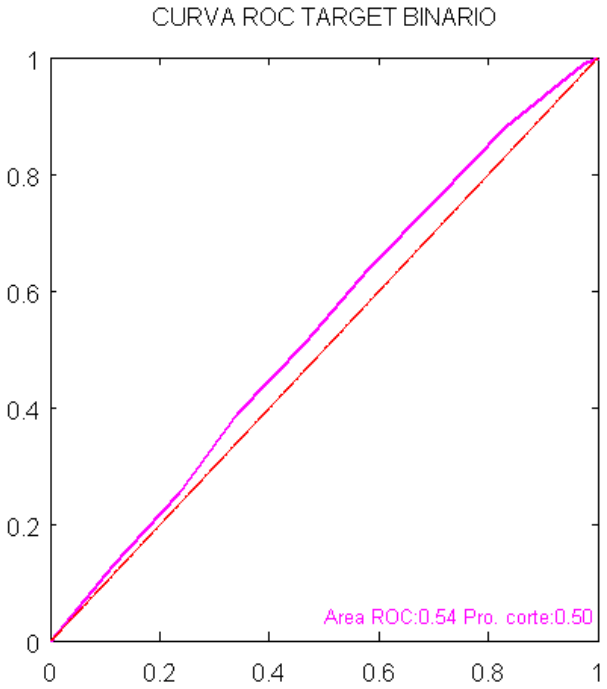
El prior en este cluster es del 74,42%

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.07	74.61%	74.63%	99.81%	2095	2085	99.52%
0.60	0.22	75.75%	77.36%	95.32%	2095	1921	91.69%
0.70	0.22	65.25%	82.08%	68.18%	2095	1295	61.81%
0.80	0.15	45.44%	84.78%	32.52%	2095	598	28.54%
0.90	0.07	30.55%	86.11%	7.95%	2095	144	6.87%



8.3.2- Cluster 3

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.21	66.94%	58.18%	27.08%	9576	1595	16.66%
0.60	0.15	65.81%	61.50%	11.93%	9576	665	6.94%
0.70	0.04	64.35%	57.65%	1.43%	9576	85	0.89%
0.80	0.01	64.22%	100%	0.03%	9576	1	0.01%
0.90	NaN	64.21%	NaN%	0.00%	9576	0	0.00%



9.- Combinación de modelos

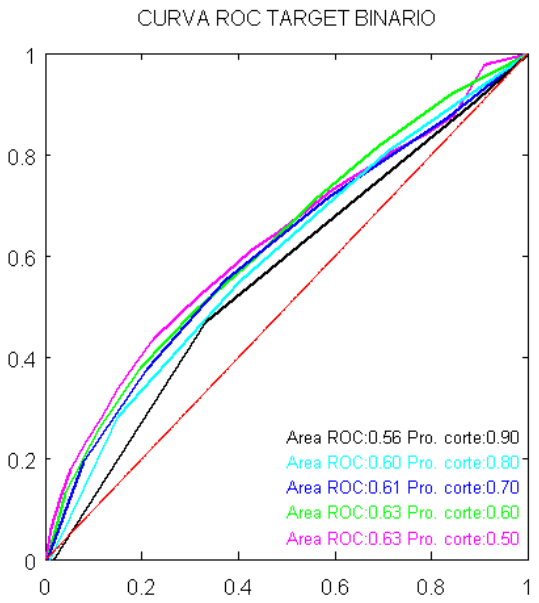
En este apartado vamos a tratar la capacidad predictora de la combinación del modelo de recargas y del modelo de saldos. Para ello nos quedaremos con la mejor aproximación de cada uno de ellos y combinaremos los resultados con el fin de obtener una predicción que mejore los obtenidos individualmente.

9.1.- Criterio de máxima probabilidad

Con este criterio asignamos la probabilidad de 1 en función de:

$$p(y = 1) = \max_{i \in \{1,2\}} \{p(y_i = 1)\}$$

Tendremos una curva ROC:



Con un área bajo la curva de 0.63 para un corte de 0.5 de probabilidad.

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0,50	0,43	72%	68%	69%	12000	5351	45%
0,60	0,33	66%	79%	32%	12000	2171	18%
0,70	0,29	64%	83%	22%	12000	1412	12%
0,80	0,25	62%	86%	15%	12000	927	8%
0,90	0,20	60%	89%	9%	12000	541	5%

El resultado en cuanto a correlación de Matthews y en cuanto a acierto global es el mejor de los encontrados hasta el momento, si comparamos los valores obtenidos con los mejores ajustes del modelo de regresión y MLP sobre la variable de recargas tenemos:

	P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
Recargas losgistica	0.50	0.35	68.06%	73.43%	42.85%	12000	3079	25.66%
Recargas MLP	0.50	0.39	70.07%	70.59%	54.72%	12000	4090	34.08%
Max Probabilidad	0,50	0,43	72%	68%	69%	12000	5351	45%

Los valores del ajuste combinado son claramente mejores en cuanto a correlación de Matthews, que se refleja en el recall mucho más elevado y en un % de selección sobre el total de la muestra para la realización de acciones comerciales mucho más elevado. Si bien la precisión es inferior a la obtenida en el modelo logístico solo para las recargas.

No obstante hay que analizar el coste de implementación y de cálculo que presenta esta solución, frente a la de aplicar la regresión logística sobre las recargas, mucho menos costoso.

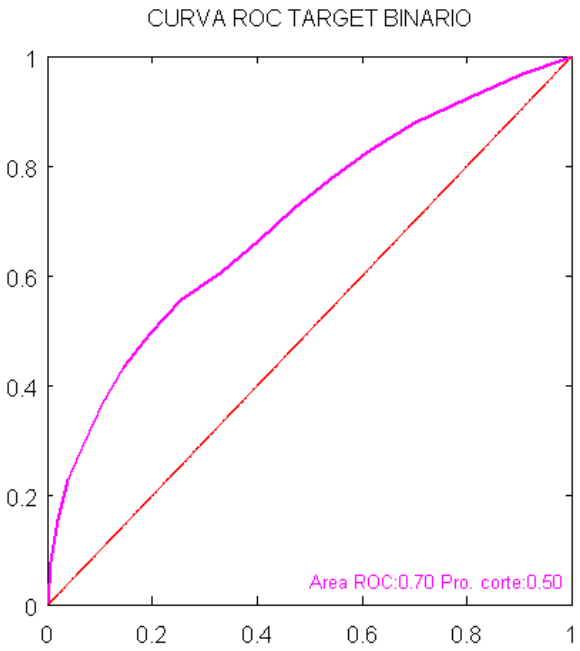
9.2.- Recargas + saldo último día

Con el fin de analizar las capacidades predictivas de la combinación de las dos variables, recargas y saldo, hemos decidido probar la introducción en el modelo de las 25 semanas creadas de forma agregada para las recargas y el saldo del último día con asignación de probabilidad de recarga, según la tabla:

	Probabilidad
0-1	0,6
1-5	0,54
5-10	0,41
10-20	0,3
>20	0,25
TOTAL	0,44

El algoritmo que mejor se comporta con estas variables es el MLP, con los siguientes resultados:

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.43	72.28%	73.66%	57.51%	12000	4119	34.33%
0.60	0.42	71.36%	79.39%	47.08%	12000	3129	26.07%
0.70	0.39	68.90%	82.96%	36.83%	12000	2342	19.52%
0.80	0.35	66.37%	87.99%	27.22%	12000	1632	13.60%
0.90	0.28	62.46%	92.79%	15.85%	12000	901	7.51%



Para el corte de probabilidad en 0,5 tenemos una precisión superior al 73% y una recall del 57%. Hemos aumentado la precisión con respecto al modelo anterior a costa del recall. Estos valores son de mayor interés para el negocio dado que las acciones comerciales a las que dará cobertura este modelo trabajan con grupos reducidos de clientes y por tanto es más interesante disponer de una mayor precisión. Además desde el punto de vista de implementación es mucho más económico generar el saldo en el último día antes de lanzar el modelo que generar el saldo en los 90 días anteriores.

10.- Conclusiones del estudio

10.1.- Modelo seleccionado

Los dos modelos que presentan un equilibrio mejor entre el coste de implementación y los resultados obtenidos son:

Resultado del modelo logístico con 180 variables sobre las recargas diarias:

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.35	68.06%	73.43%	42.85%	12000	3079	25.66%
0.60	0.33	66.16%	79.59%	30.97%	12000	2053	17.11%
0.70	0.29	63.69%	83.13%	21.85%	12000	1387	11.56%
0.80	0.25	61.61%	86.48%	15.03%	12000	917	7.64%
0.90	0.20	59.56%	89.53%	9.08%	12000	535	4.46%

Resultado del modelo MLP con 25 variables sobre las recargas semanales y el saldo del último día:

P. CORTE	MATTHEWS C	ACCURACY	PRECISION	RECALL	N	SELECCIÓN	P SEL TOTAL
0.50	0.43	72.28%	73.66%	57.51%	12000	4119	34.33%
0.60	0.42	71.36%	79.39%	47.08%	12000	3129	26.07%
0.70	0.39	68.90%	82.96%	36.83%	12000	2342	19.52%
0.80	0.35	66.37%	87.99%	27.22%	12000	1632	13.60%
0.90	0.28	62.46%	92.79%	15.85%	12000	901	7.51%

El segundo modelo construido con las dos variables proporciona resultados considerablemente mejores que el logístico, aunque tiene un mayor coste de implementación dado que el MLP es mucho más complicado de programar que la regresión y además implica el cálculo de la variable del saldo en el último día.

Las mejoras que apreciamos entre uno y otro se centran sobre todo en el recall y en él % seleccionado sobre el total. En el caso del MLP en los cortes de mayor probabilidad es de casi el doble. Por ejemplo si decidimos seleccionar el 20% de la muestra, aproximadamente, el MLP tendrá unos valores de precisión del 83% y un recall del 37%, frente al modelo logístico que tendrá un acierto por debajo del 77% y un recall algo por encima del 31%.

Por tanto nos quedaremos con el segundo modelo para su paso a producción.

10.2.- Uso por parte del Operador

El uso que se propone para el modelo consiste en la actualización del mismo una vez al mes para dar soporte a las acciones comerciales de estimulación de la recarga.

Se proporcionará a la unidad de negocio un scoring de propensión a la recarga en los siguientes 28 días para que, en función del presupuesto y de criterios de optimización, se seleccionen los clientes deseados.

La sugerencia sería cortar en probabilidades del modelo cercanas al 0.7, puesto que hemos visto que proporcionan valores buenos de acierto, estando en torno al 83%. Llegando, para los avalores más altos de probabilidad, a casi el 93%. Al mismo tiempo, en estos valores de probabilidad, disponemos de una masa de clientes potenciales para realizarles la acción comercial cercana al 20%, lo que quiere decir que si el Operador dispusiera de 1M de líneas podríamos hacerle la oferta comercial a 200K, volumen suficiente para justificar una acción comercial.

Debido a la tendencia detectada en el estudio exploratorio de las recargas se recomienda actualizar el modelo cada 4 meses.

Hay que destacar que por problemas de histórico y migración entre distintas BBDD no se ha podido hacer una prueba del modelo en un t diferente, esto es realizando una extracción meses después de la empleada para entrenar y comprobar el acierto obtenido. Es muy recomendable para asegurarnos que la estacionalidad u otros factores no afectan al modelo.

10.3.- Paso a producción

El modelo se implementará en el entorno de BD del operador, automatizando los procesos, con el fin de que se ejecuten una vez al mes, produciendo una tabla de salida que contendrá todas las numeraciones de prepago activas en el momento de la ejecución y con un mínimo de antigüedad de 6 meses en la compañía, para poder generar las variables necesarias, y una variable de scoring que permitirá seleccionar las líneas más propensas.

10.4.- Conclusiones de negocio

A continuación resumiremos las principales conclusiones que hemos obtenido en el presente proyecto y que pueden ser aplicadas en el negocio:

- A partir de la información obtenida podemos determinar, que en el caso de que hubiera que elegir algún momento de la semana específico para realizar la acción comercial, los mejores serían los domingos-lunes, para intentar influir en las posibles recargas que se producen los lunes, o los jueves-viernes, para hacerlo en las que se producen los viernes.

- Dado que uno de nuestros objetivos es modificar el punto de recarga de los clientes que acuden al distribuidor, con el objetivo de que pasen a cajero o portal web de la Operador, con el objetivo de mejorar los márgenes, parece interesante la opción de hacer algún tipo de comunicación a los clientes que vayan a necesitar recargar con alta probabilidad durante los días festivos. Recordándoles que los cajeros y la recarga online están siempre a su disposición.
- El tiempo entre recargas está aumentando, al menos en un conjunto considerable de líneas.
- No tiene sentido profundizando en la granularidad del target más allá de la estructura semanal.
- El cluster 2 sobre recargas, con DTW, tienen una especial relevancia para el negocio dado que, a pesar de contar tan solo con el 17% de las líneas, concentra el 54% del importe recargado, por tanto cualquier acción comercial sobre este colectivo puede tener un importante impacto en los resultados. La probabilidad a priori de recarga de este colectivo está en el 78%, muy por encima de la media.
- Segmentación sobre variable recargas reducida: Es evidente que este colectivo de clientes aun siendo solo el 2,7% de la población presentan un comportamiento claramente diferenciado del resto, son líneas prepago de muy alto consumo y con una frecuencia de recarga elevada. Sería interesante analizar la posibilidad de migrarlos a contrato puesto que con consumos tan elevados son muy vulnerables. Habría que analizar el coste para la compañía por la pérdida de ingresos y el riesgo de fuga en caso de no actuar.

Anexos

A0 Metodología y desarrollo del proyecto.

La metodología del proyecto se puede dividir en las siguientes fases:

1. Obtención de datos a partir de las fuentes de información disponibles en la BD corporativa del Operador, seleccionando el horizonte temporal adecuado para el estudio.
2. Tratamiento de los datos, eliminación de outliers y creación de las variables necesarias para el modelado. Creación de variables independientes a nivel de día, tanto para saldos como para recargas, así como de las variables dependiente a nivel mensual, quincenal, semanal...
3. Empleo de algoritmos de clasificación supervisada, regresión logística y Perceptrón Multicapa.
4. Análisis de la mejora en la clasificación segmentando las líneas y modelando cada segmento por separado.
5. Comparativa de resultados y elección del mejor modelo, en base a criterios de acierto e implementación

Si profundizamos un en las técnicas empleadas

Regresión logística

El algoritmo empleado es un gradiente descendente que optimiza la función de coste [8]:

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} . \log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right). \log(1 - h_\theta\left(x^{(i)}\right))]$$

Si añadimos un término de regularización tenemos [8]

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} . \log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right). \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right] + \frac{\lambda}{2m}\sum_{j=2}^{n} \theta_j^2$$

Dado que la función de coste de la regresión logística es convexa el algoritmo de gradiente descendente siempre alcanza el óptimo global, con un número suficiente de iteraciones.

Perceptrón Multicapa

El algoritmo implementado es un gradiente descendente que optimiza la función de coste [8]:

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} . \log\left(h_\theta\left(x^{(i)}\right)_k\right) + \left(1 - y_k^{(i)}\right) . \log(1 - h_\theta\left(x^{(i)}\right)_k)]$$

Si añadimos un término de regularización tenemos [8]:

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)}.\log\left(h_\theta\left(x^{(i)}\right)_k\right) + \left(1 - y_k^{(i)}\right).\log(1 - h_\theta\left(x^{(i)}\right)_k)] + \frac{\lambda}{2m}\sum_{l=1}^{l-1}\sum_{i=1}^{S_l}\sum_{j=1}^{S_{l-1}}(\theta_j^{(i)})^2$$

Dado que la función no es convexa corremos el riesgo que "caer" en un óptimo local, para evitarlo reordenamos la información de forma aleatoria y repetimos la clasificación diversas veces.

Desarrollo del proyecto

Las principales dificultades que han surgido durante el proyecto han sido:

1. Reconstruir la información del Operador, las variables diarias para el saldo no se encuentran en la BD del operador y ha sido costoso recrear una aproximación de la suficiente calidad. En el anexo 6 se puede ver una parte del código empelado.
2. Automatización del proceso de aprendizaje de los algoritmos de regresión logística y redes neuronales, de forma que devolvieran toda la información requerida y que permitieran ajustar el balanceo y las diferentes métricas de aprendizaje, regularización, número de neuronas en la capa oculta…
3. Utilización del algoritmo DTW. Dado que todo el proceso de modelado se ha realizado en Octave se comenzó usando un algoritmo de DTW en esta herramienta, ante su poco rendimiento se pasó a ejecutar el algoritmo en Matlab y después a investigar en las librerías de procesamiento matricial de Matlab y Octave para mejorar los tiempos. Se hicieron muchas pruebas y se consiguió ejecutar para un número pequeño de registros. Hay que tener en cuenta que el K-means dentro del que iba el DTW necesita un gran número de ejecuciones de la distancia entre centroides y puntos para hacer la asignación de cada línea a un cluster y esto producía unos tiempos altísimos de ejecución para muestras muy pequeñas, lo que hacía inviable su uso sobre toda la muestra e imposibilitaba su paso a producción. Finalmente se empleó el DTW de R, programado con una librería en C y que aporta un buen rendimiento.
4. El código SQL necesario ara obordar el proyecto cuenta con más de 1.600 líneas de código, en los anexos se puedo ver gran parte.
5. En cuanto al código Octave se ha desarrollado con la versión 3.2.4, un cuenta con más de 1.000 líneas de código uniendo el proceso de Regresión Logística y MLP. En los anexos se pueden ver ejemplos del código para los lanzadores (programas de ejecución del proceso, balanceo de la muestra, fichero de texto de salidas, gráficos…).

En total se han empleados dos meses de trabajo continuado para realizar la programación del proyecto, tanto en Octave como en SQL, y las diferentes simulaciones de ajuste, así como la redacción del presente documento.

A1 Ajuste Multitarget

Para realizar los ajustes multitarget del presente trabajo se ha procedido a generar código en Octave que permita, sobre un conjunto de etiquetas {0,1,2,3…} obtener el modelo que mejor ajusta cada una de forma individual.

Creamos tantas variables target de salida como etiquetas tenemos, recodificando por cuestiones prácticas la etiqueta 0 como la n+1, en un conjunto de n etiquetas. Por tanto para el problema de ajuste de la recarga con períodos bisemanales hemos creado tres etiquetas. Por tanto la variable target tomará los valores:

- 1: cuando las líneas hayan efectuado una recarga en las dos primeras semanas.
- 2: cuando las líneas hayan efectuado recarga en las dos últimas semanas.
- 3: cuando las líneas no hayan recargado.

Una vez codificado el target procedemos al ajuste de los modelos multitarget, creando un modelo para estimar cada una de las etiquetas por separado, de forma que en el caso bisemanal entrenaremos tres modelos, de la forma:

- El primer modelo mide la probabilidad de recarga en las dos primeras semanas frente al resto, por tanto codificamos como uno las líneas que han efectuado recarga y como cero al resto.
- El segundo modelo mide la probabilidad de recarga en las dos últimas semanas frente al resto, por tanto codificamos como unos las líneas que han efectuado recarga y como cero el resto.
- El tercer modelo mide la probabilidad de no recarga, codificamos como uno las líneas que no han recargado y como cero el resto.

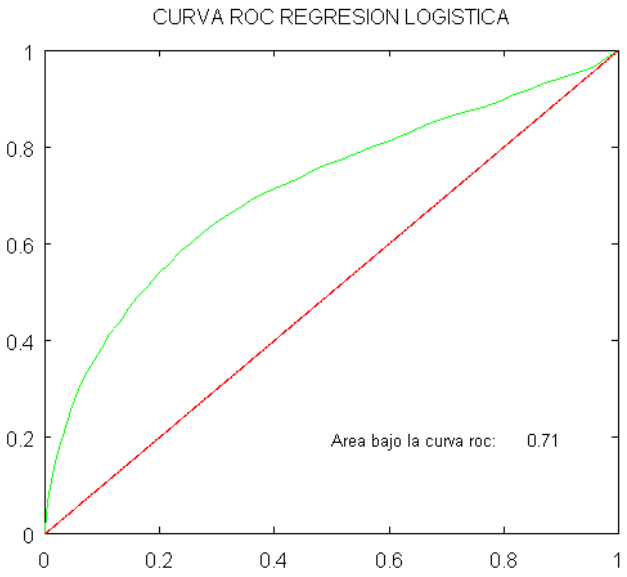
Una vez disponemos de los tres modelos empleamos un sistema de votación para decidir que etiqueta asignamos a la previsión. De forma que la estimación vendrá dada por la etiqueta de mayor probabilidad de entre los modelos generados.

Devolvemos la probabilidad y la etiqueta asignada.

A2 Recargas regresión logística balanceo

La tabla inicial de resultados para el ajuste de la regresión logística sin balancear se muestra en la Tabla 1:

P. CORTE	MATTHEWS	ACCURACY	PRECISION	RECALL	N	SELEC	P SEL TOT
0.50	0.35	68.06%	73.43%	42.85%	12000	3079	25.66%
0.60	0.33	66.16%	79.59%	30.97%	12000	2053	17.11%
0.70	0.29	63.69%	83.13%	21.85%	12000	1387	11.56%
0.80	0.25	61.61%	86.48%	15.03%	12000	917	7.64%
0.90	0.20	59.56%	89.53%	9.08%	12000	535	4.46%

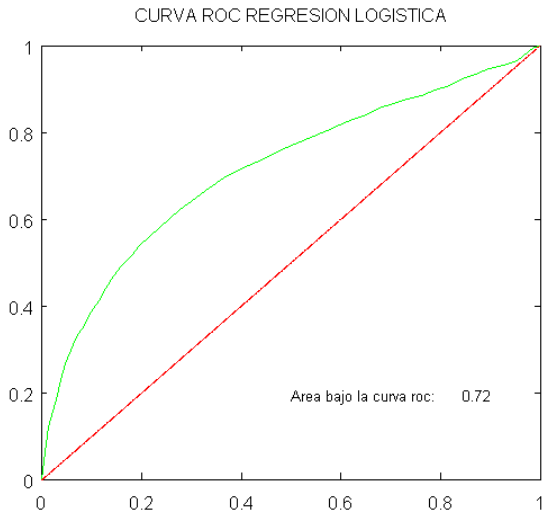


Aplicamos el primer balanceo, forzando una proporción de positivos igual a la de negativos (50% cada uno), los resultados de pueden ver en la Tabla 2:

Balanceo al 50%

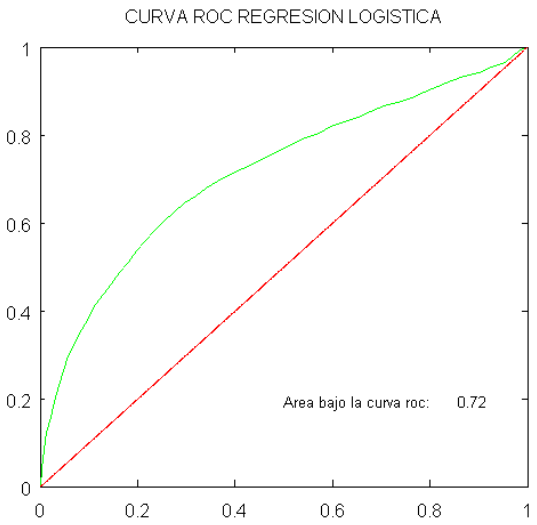
P. CORTE	MATTHEWS	ACCURACY	PRECISIO N	RECALL	N	SELEC	P SEL TOT
0.50	0.36	68.65%	68.52%	53.09%	12000	4088	34.07%
0.60	0.34	67.29%	75.85%	37.57%	12000	2613	21.77%
0.70	0.31	65.14%	81.61%	26.74%	12000	1729	14.41%
0.80	0.26	62.40%	84.35%	17.78%	12000	1112	9.27%
0.90	0.21	60.20%	88.11%	10.96%	12000	656	5.47%

Hay que tener en cuenta que para que los resultados sean comparables a los de la tabla sin balancear, hay que realizar un proceso para la transformación de las probabilidades a posteriori, teniendo en cuenta el sobrebalanceo empleado. Se puede ver con más detalle el proceso empleado en los anexos.



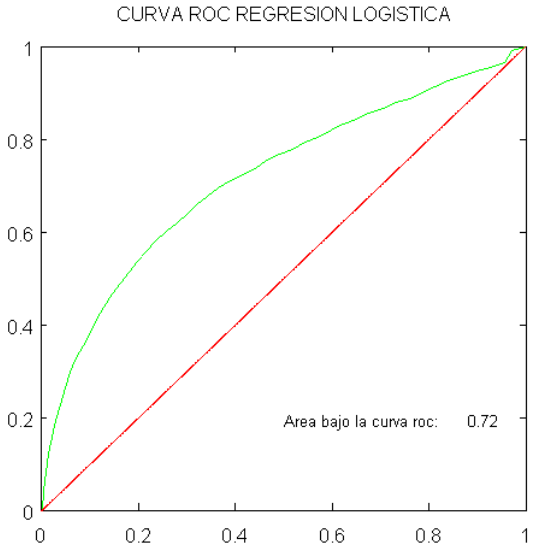
Balanceo al 55%.

P. CORTE	MATTHEWS	ACCURACY	PRECISIO N	RECALL	N	SELEC	P SEL TOT
0.50	0.35	68.01%	63.77%	63.08%	12000	5219	43.49%
0.60	0.35	68.17%	72.40%	44.60%	12000	3250	27.08%
0.70	0.33	66.22%	78.97%	31.60%	12000	2111	17.59%
0.80	0.28	63.43%	83.04%	21.15%	12000	1344	11.20%
0.90	0.23	60.82%	88.38%	12.55%	12000	749	6.24%



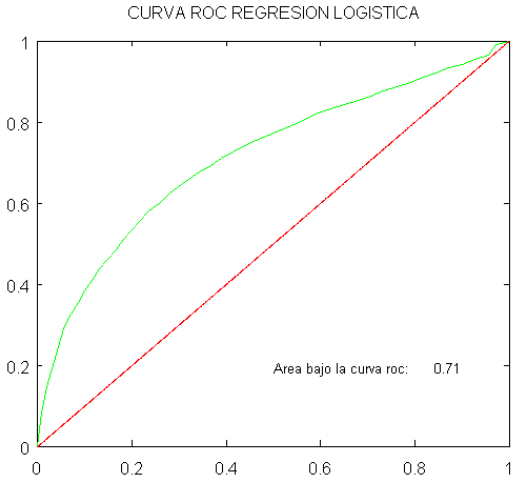
Balanceo al 60%.

P. CORTE	MATTHEWS	ACCURACY	PRECISION	RECALL	N	SELEC	P SEL TOT
0.50	0.30	64.05%	57.10%	73.35%	12000	6778	56.48%
0.60	0.36	68.66%	68.53%	53.11%	12000	4089	34.08%
0.70	0.34	67.24%	75.86%	37.40%	12000	2601	21.68%
0.80	0.30	64.43%	81.90%	24.53%	12000	1580	13.17%
0.90	0.24	61.25%	85.98%	14.18%	12000	870	7.25%



Balanceo al 65%.

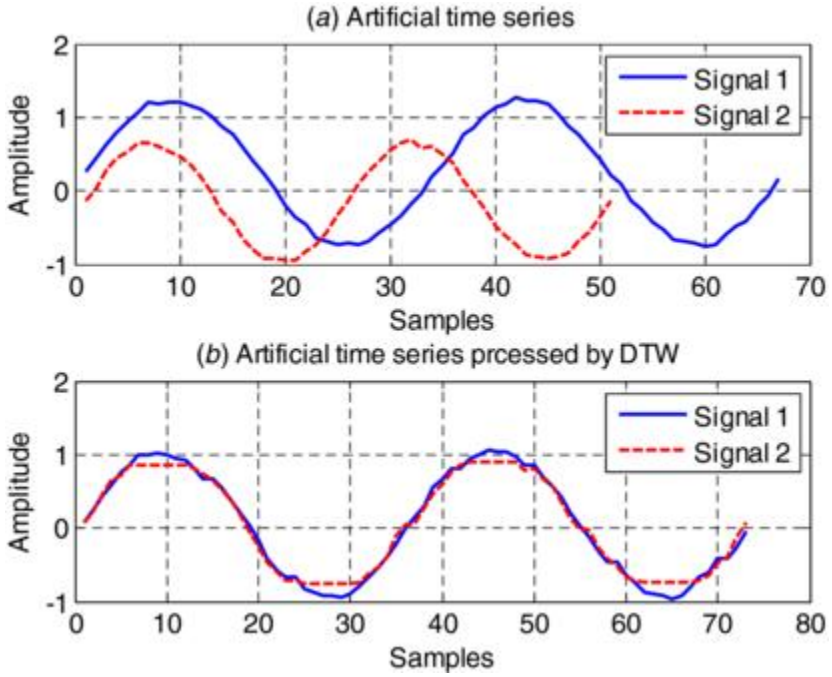
P. CORTE	MATTHEWS	ACCURACY	PRECISION	RECALL	N	SELEC	P SEL TOT
0.50	0.19	54.93%	49.28%	85.67%	12000	9173	76.44%
0.60	0.34	67.50%	62.85%	63.78%	12000	5354	44.62%
0.70	0.34	68.00%	71.76%	44.88%	12000	3300	27.50%
0.80	0.32	65.92%	79.40%	30.38%	12000	2019	16.83%
0.90	0.25	62.09%	84.07%	17.00%	12000	1067	8.89%



A3.- DTW (Dynamic Time Warping)

El DTW es un algoritmo que mide la similitud entre dos secuencias que varían en el tiempo [7].

El DTW analiza al similitud entre dos series teniendo en cuenta sus puntos singulares, por ejemplo si tenemos dos series iguales pero una de ellas está desplazada hacia la derecha la distancia DTW entre ellas será cero.



En general el método DTW permite encontrar el mapeo óptimo entre dos series temporales, lo que posibilita analizar la distancia entre ellas.

El empleo de la distancia DTW en vez de la euclídea podría permitirnos encontrar grupos de series que, aunque presenten cierto desplazamiento, sean prácticamente iguales en cuanto a amplitud de períodos y forma general.

A4.- Correlación de Matthews

El coeficiente de correlación de Matthews [3] se emplea en machine learning para obtener una medida de la calidad del ajuste de modelos con target binarios. El coeficiente de Matthews es una medida derivada del coeficiente phi y está relacionado con las medidas chi-cuadrado para una tabla de contingencia 2x2.

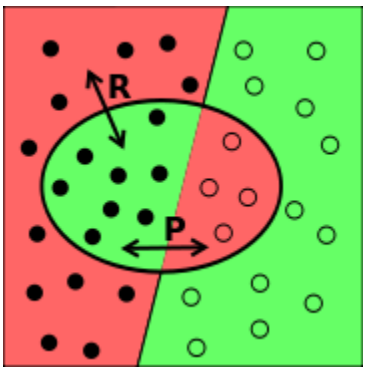
$$|\mathrm{MCC}| = \sqrt{\frac{\chi^2}{\mathrm{n}}}$$

Siendo n el número total de observaciones de la muestra.

MCC se puede calcular directamente en función de la tabla de contingencia a través de la fórmula:

$$\mathrm{MCC} = \frac{\mathrm{TP}.\,\mathrm{TN} - \mathrm{FP}.\,\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP}).\,(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

A5.- F1 score



El F1 score es una medida de la precisión del modelo, basada en dos conceptos, Precision y Recall, que traduciremos como Precisión y Captura [2]. Para introducirnos en el concepto es conveniente analizar las posibilidades que presenta la matriz de confusión, en la que se visualizan las combinaciones de los posibles resultados en un proceso de clasificación, en nuestro caso binario.

		Datos reales: 1	Datos reales: 0
Predichos	1	tp (verdaderos positivos)	fp (falsos positivos)
	0	fn (falsos negativos)	tn (verdaderos negativos)

- **Precisión:** mide el acierto del modelo a la hora de evaluar los verdaderos positivos.

$$Precisión = \frac{tp}{tp + fp}$$

- **Captura:** mide cuantos, del total de positivos que hay en la muestra inicial, estamos clasificando correctamente.

$$Captura = \frac{tp}{tp + fn}$$

El F1 score vendrá dado por la fórmula:

$$F1 = 2.\frac{Precisión.Captura}{Precisión + Captura}$$

El empleo de esta métrica nos permitirá evaluar los diferentes algoritmos empleados para la resolución del problema con consistencia, permitiéndonos seleccionar el mejor.

A6.- Código SQL, extracción de información

```
/*************************************************************************************************************************/
/*************************************CREACIÓN DELA MUESTRA DE TRABAJO*************************************/
/*************************************************************************************************************************/

/*************************************************************************************************************************/
/*********************************************COLECTIVO INICIAL DE LINEAS*********************************************/
/*************************************************************************************************************************/
DROP TABLE MODELO_REC_COLECTIVO_P
/
CREATE TABLE MODELO_REC_COLECTIVO_P
-- Consulta eliminada, confidencial
/
/*************************************************************************************************************************/
/**********************ELIMINAMOS LAS LÍNEAS QUE QUE NO ESTÉN ACTIVAS EL DÍA 14**********************************/
/*************************************************************************************************************************/
DROP TABLE MODELO_REC_MOV_SALIDA_L
/
-- Consulta eliminada, confidencial
/
DROP TABLE MODELO_REC_MOV_SALIDA_DET
/
-- Consulta eliminada, confidencial
/

DROP TABLE MODELO_REC_COLECTIVO_P_F
/
-- Consulta eliminada, confidencial
/
DROP TABLE MODELO_REC_COLECTIVO_P
/
```

```
/*************************************************************************************************************************************/
/************************OBTENEMOS UNA MUESTRA DEL COLECTIVO PARA ENTRENAR Y VALIDAR************************/
/*************************************************************************************************************************************/
DROP TABLE MODELO_REC_COLECTIVO_P_MUES
/
CREATE TABLE MODELO_REC_COLECTIVO_P_MUES AS
SELECT *
FROM
(SELECT *
FROM MODELO_REC_COLECTIVO_P_F
WHERE ID_DIA_ALTA<TO_DATE('01/12/12','DD/MM/YY')-180-12
ORDER BY DBMS_RANDOM.VALUE)
WHERE ROWNUM<=100000
/
CREATE INDEX MODEL_REC ON MODELO_REC_COLECTIVO_P_MUES (NUM_TEL)
/
ANALYZE TABLE MODELO_REC_COLECTIVO_P_MUES ESTIMATE STATISTICS
/
/*************************************************************************************************************************************/
/*********************************OBTENCIÓN Y PROCESAMIENTO DE LAS RECARGAS*********************************/
/*************************************************************************************************************************************/
DROP TABLE MODELO_REC_COLECTIVO_P_M_REC
/
-- Consulta eliminada, confidencial
/
/*************************************************************************************************************************************/
/*************************************TRATAMIENTO DE LAS RECARGAS PROMOCIONALES*************************************/
/*************************************************************************************************************************************/
DROP TABLE MODELO_REC_COLECTIVO_P_M_REC_P
/
CREATE TABLE MODELO_REC_COLECTIVO_P_M_REC_P AS
SELECT
A.NUMERO_TELEFONO,
```

```
A.ID_DIA       ,
A.IMPORTE_RECARGA,
CASE
WHEN REC_PROMOCIONAL<RECARGAS THEN 0
ELSE 1 END REC_PROMOCIONAL
FROM
(SELECT
A.NUMERO_TELEFONO,
A.ID_DIA       ,
SUM(A.IMPORTE_RECARGA) IMPORTE_RECARGA,
SUM(-- Confidencial) REC_PROMOCIONAL, COUNT(*) RECARGAS
FROM  -- Confidencial
WHERE A.ID_ORIGEN_RECARGA=B.ID_ORIGEN_RECARGA
GROUP BY A.NUMERO_TELEFONO, A.ID_DIA) A
/
DROP TABLE MODELO_REC_REASIG_REC_PROM
/
CREATE TABLE MODELO_REC_REASIG_REC_PROM AS
SELECT A.NUMERO_TELEFONO, ID_DIA_DEST, SUM(IMPORTE_RECARGA) IMPORTE_RECARGA
FROM
(SELECT A.NUMERO_TELEFONO, MAX(A.ID_DIA) ID_DIA_DEST, B.ID_DIA ID_DIA_ORIG, B.IMPORTE_RECARGA
FROM MODELO_REC_COLECTIVO_P_M_REC_P A,
(SELECT NUMERO_TELEFONO, ID_DIA, IMPORTE_RECARGA
FROM MODELO_REC_COLECTIVO_P_M_REC_P
WHERE REC_PROMOCIONAL=1) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO
AND A.ID_DIA<B.ID_DIA
AND A.REC_PROMOCIONAL=0
GROUP BY A.NUMERO_TELEFONO,  B.ID_DIA, B.IMPORTE_RECARGA) A
GROUP BY A.NUMERO_TELEFONO, ID_DIA_DEST
/
DROP TABLE MODELO_REC_REASIG_REC_PROM_1
/
```

```
CREATE TABLE MODELO_REC_REASIG_REC_PROM_1 AS
SELECT
A.NUMERO_TELEFONO      ,
A.ID_DIA               ,
A.IMPORTE_RECARGA    IMPORTE_RECARGA_NO_PROMO  ,
NVL(B.IMPORTE_RECARGA,0) IMPORTE_RECARGA_PROMO,
A.IMPORTE_RECARGA + NVL(B.IMPORTE_RECARGA,0) IMPORTE_RECARGA
FROM MODELO_REC_COLECTIVO_P_M_REC_P A, MODELO_REC_REASIG_REC_PROM B
WHERE A.REC_PROMOCIONAL=0
AND A.NUMERO_TELEFONO=B.NUMERO_TELEFONO (+)
AND A.ID_DIA=B.ID_DIA_DEST (+)
/
/*************************************************************************************************************************/
/*************************************************ELIMINAMOS OUTLIERS*****************************************************/
/*************************************************************************************************************************/

DROP TABLE MODELO_REC_COLECTIVO_FIN
/
CREATE TABLE MODELO_REC_COLECTIVO_FIN AS
SELECT NUMERO_TELEFONO
FROM MODELO_REC_REASIG_REC_PROM_1 A
GROUP BY NUMERO_TELEFONO
HAVING COUNT(DISTINCT ID_DIA)<54
/
/*************************************************************************************************************************/
/************************************CALCULAMOS EL TIEMPO ENTRE RECARGAS**************************************************/
/*************************************************************************************************************************/
DROP TABLE MODELO_REC_COLECTIVO_P_M_REC_AG
/
CREATE TABLE MODELO_REC_COLECTIVO_P_M_REC_AG AS
SELECT B.NUMERO_TELEFONO, ID_DIA, SUM(IMPORTE_RECARGA) IMPORTE_RECARGA
FROM MODELO_REC_REASIG_REC_PROM_1, (SELECT NUMERO_TELEFONO, COUNT(DISTINCT ID_DIA) DIA FROM
MODELO_REC_REASIG_REC_PROM_1
```

```
                                        GROUP BY NUMERO_TELEFONO

                                        HAVING COUNT(DISTINCT ID_DIA)>=2) B

WHERE MODELO_REC_REASIG_REC_PROM_1.NUMERO_TELEFONO=B.NUMERO_TELEFONO

GROUP BY B.NUMERO_TELEFONO, ID_DIA

ORDER BY B.NUMERO_TELEFONO, ID_DIA DESC

/

-- SELECCIONAMIOS SOLO LAS LÍNEAS QUE TIENEN COMO MÍNIMO 2 RECARGAS EN EL PERÍDO DE ESTUDIO

-- AGRUPAMOS LOS IMPORTES RECARGADOS POR DÍA DE FORMA QUE CONSIDERAMOS COMO MÁXIMO UNA RECARGA POR
DÍA, SI SE DIERAN

-- MÁS AGRUPARIAMOS FUNDIMOS LAS RECARGAS SUMANDO LOS IMPORTES.

/**************CREAMOS UNA TABLA CON LA RECARGA N Y LA N+1 EN MISMA FILA PARA ANALIZAR TIEMPOS**************/

CREATE TABLE MODELO_REC_COLECTIVO_P_M_REC_AG2 AS

SELECT A.NUMERO_TELEFONO, A.ID_DIA ID_DIA_INI, B.ID_DIA ID_DIA_FIN, B.ID_DIA-A.ID_DIA DIF_DIAS, A.IMPORTE_RECARGA

FROM

(SELECT NUMERO_TELEFONO, ID_DIA, IMPORTE_RECARGA, ROWNUM COL

FROM MODELO_REC_COLECTIVO_P_M_REC_AG) A,

(SELECT NUMERO_TELEFONO, ID_DIA, IMPORTE_RECARGA, ROWNUM+1 COL

FROM MODELO_REC_COLECTIVO_P_M_REC_AG) B

WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO

AND A.COL=B.COL

/

DROP TABLE MODELO_REC_PARES_RECARGAS

/

-- IMPORTANTE, ESTAMOS CALCULANDO LA PENDIENTE CON LOS DATOS EN ORDEN INVERSO => SI LA PENDIENTE ES
NEGATIVA LA LÍNEA CADA VEZ

-- TARDA MÁS EN RECARGAR.

-- SOLO LO CALCULAMOS PARA LOS CLIENTES QUE TIENEN MÁS DE 5 RECATGAS EN EL PERÍDO DE ESTUDIO

CREATE TABLE MODELO_REC_PARES_RECARGAS AS

SELECT

NUMERO_TELEFONO, COUNT(*) PARES_RECARGAS, COVAR_POP(ROWNUM, DIF_DIAS)/(VARIANCE (ROWNUM))
PEND_DIF_DIAS

FROM MODELO_REC_COLECTIVO_P_M_REC_AG2

GROUP BY NUMERO_TELEFONO

HAVING COUNT(*)>=5;
```

```
/***********************************************VARIABLES X1 X180*****************************************************/

DROP TABLE MODELO_REC_COLECTIVO_P_M_REC_F180

/

CREATE TABLE MODELO_REC_COLECTIVO_P_M_REC_F180 AS

SELECT *

FROM MODELO_REC_REASIG_REC_PROM_1

WHERE ID_DIA<'19/11/12'

AND ID_DIA>=TO_DATE('19/11/12','DD/MM/YY')-180

/

DROP TABLE MODELO_REC_DIAS_TELEF_180

/

CREATE TABLE MODELO_REC_DIAS_TELEF_180 AS

SELECT A.NUMERO_TELEFONO, A.ID_DIA, A.VAR_X, NVL(B.IMP_RECARGADO,0) IMP_RECARGADO, NVL(B.NUM_RECARGAS,0)
NUM_RECARGAS

FROM

(SELECT

NUMERO_TELEFONO, ID_DIA, MOD(ROWNUM,180) VAR_X

FROM

(SELECT ID_DIA

FROM  MODELO_REC_COLECTIVO_P_M_REC_F180 -- SACAMOS TODOS LOS DIAS POSIBLES CUIDADO, PUEDEN FALTAR

GROUP BY ID_DIA

ORDER BY ID_DIA) A,

(SELECT NUM_TEL NUMERO_TELEFONO

FROM MODELO_REC_COLECTIVO_P_MUES  -- SACAMOS TODAS LAS LÍNEAS POSIBLES

GROUP BY NUM_TEL)

ORDER BY NUMERO_TELEFONO, ID_DIA) A,

(SELECT NUMERO_TELEFONO, ID_DIA, TRUNC(SUM(IMPORTE_RECARGA)/1000,2) IMP_RECARGADO, COUNT(*)
NUM_RECARGAS -- SUMAMOS POR SI HAY MAS DE UNA RECARGA POR DIA

FROM MODELO_REC_COLECTIVO_P_M_REC_F180

GROUP BY NUMERO_TELEFONO, ID_DIA) B

WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO (+)

AND A.ID_DIA=B.ID_DIA (+)

/

UPDATE MODELO_REC_DIAS_TELEF_180
```

```
SET VAR_X=180 WHERE VAR_X=0
/
COMMIT
/
/*****************************************************************************************************************/
/*****************************TRANSPOSICIÓN DE LA MATRIZ  180 RECARGA CON IMP*************************************/
/*****************************************************************************************************************/
DROP TABLE MODELO_REC_TRANS_180
/
CREATE TABLE MODELO_REC_TRANS_180 AS
SELECT
NUMERO_TELEFONO,
SUM(DECODE (VAR_X,      1       ,       1       ,       0)*IMP_RECARGADO)               VAR_X1  ,
SUM(DECODE (VAR_X,      2       ,       1       ,       0)*IMP_RECARGADO)               VAR_X2  ,
SUM(DECODE (VAR_X,      3       ,       1       ,       0)*IMP_RECARGADO)               VAR_X3  ,
SUM(DECODE (VAR_X,      4       ,       1       ,       0)*IMP_RECARGADO)               VAR_X4  ,
SUM(DECODE (VAR_X,      5       ,       1       ,       0)*IMP_RECARGADO)               VAR_X5  ,
SUM(DECODE (VAR_X,      6       ,       1       ,       0)*IMP_RECARGADO)               VAR_X6  ,
……
SUM(DECODE (VAR_X,      177     ,       1       ,       0)*IMP_RECARGADO)               VAR_X177
        ,
SUM(DECODE (VAR_X,      178     ,       1       ,       0)*IMP_RECARGADO)               VAR_X178
        ,
SUM(DECODE (VAR_X,      179     ,       1       ,       0)*IMP_RECARGADO)               VAR_X179
        ,
SUM(DECODE (VAR_X,      180     ,       1       ,       0)*IMP_RECARGADO)               VAR_X180

FROM MODELO_REC_DIAS_TELEF_180
GROUP BY NUMERO_TELEFONO
/
/*****************************************************************************************************************/
/*************************************************VARIABLES Y1-Y28************************************************/
/*****************************************************************************************************************/
DROP TABLE MODELO_REC_VAR_Y_BRUTO
/
```

```
CREATE TABLE MODELO_REC_VAR_Y_BRUTO AS

SELECT *

FROM MODELO_REC_COLECTIVO_P_M_REC

WHERE ID_DIA>='19/11/12'

AND ID_DIA<TO_DATE('19/11/12','DD/MM/YY')+28

/

DROP TABLE MODELO_REC_VAR_Y

/

CREATE TABLE MODELO_REC_VAR_Y AS

SELECT A.NUMERO_TELEFONO, A.ID_DIA, A.VAR_Y, NVL(B.RECARGA,0) RECARGA

FROM

(SELECT NUMERO_TELEFONO, ID_DIA, ROWNUM, MOD(ROWNUM,28) VAR_Y

FROM

(SELECT

NUMERO_TELEFONO, ID_DIA

FROM

(SELECT ID_DIA

FROM  MODELO_REC_VAR_Y_BRUTO -- SACAMOS TODOS LOS DIAS POSIBLES CUIDADO, PUEDEN FALTAR

GROUP BY ID_DIA

ORDER BY ID_DIA) A,

(SELECT NUM_TEL NUMERO_TELEFONO

FROM MODELO_REC_COLECTIVO_P_MUES  -- SACAMOS TODAS LAS LÍNEAS POSIBLES

GROUP BY NUM_TEL)

ORDER BY NUMERO_TELEFONO, ID_DIA) A) A,

(SELECT A.NUMERO_TELEFONO, A.ID_DIA, SUM(IMPORTE_RECARGA)/1000 RECARGA -- SUMAMOS POR SI HAY MAS DE UNA
RECARGA POR DIA

FROM  MODELO_REC_VAR_Y_BRUTO A,

(SELECT NUMERO_TELEFONO, MIN(ID_DIA) ID_DIA -- SOLO NOS QUEDAMOS CON LA PRIMERA RECARGA PARA Y

FROM MODELO_REC_VAR_Y_BRUTO

GROUP BY NUMERO_TELEFONO) B

WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO

AND A.ID_DIA=B.ID_DIA

GROUP BY A.NUMERO_TELEFONO, A.ID_DIA) B
```

```
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO (+)
AND A.ID_DIA=B.ID_DIA (+)
/
UPDATE MODELO_REC_VAR_Y
SET VAR_Y=28 WHERE VAR_Y=0
/
COMMIT
/
DROP TABLE MODELO_REC_VAR_Y_TRANS_1_0
/
CREATE TABLE MODELO_REC_VAR_Y_TRANS_1_0 AS
SELECT NUMERO_TELEFONO,
SUM(DECODE (VAR_Y,      1      ,      1      ,      0)*RECARGA)          VAR_Y1   ,
SUM(DECODE (VAR_Y,      2      ,      1      ,      0)*RECARGA)          VAR_Y2   ,
SUM(DECODE (VAR_Y,      3      ,      1      ,      0)*RECARGA)          VAR_Y3   ,
……
SUM(DECODE (VAR_Y,      25     ,      1      ,      0)*RECARGA)          VAR_Y25  ,
SUM(DECODE (VAR_Y,      26     ,      1      ,      0)*RECARGA)          VAR_Y26  ,
SUM(DECODE (VAR_Y,      27     ,      1      ,      0)*RECARGA)          VAR_Y27  ,
SUM(DECODE (VAR_Y,      28     ,      1      ,      0)*RECARGA)          VAR_Y28  ,
SUM(RECARGA) RECARGA
FROM
(SELECT
NUMERO_TELEFONO, ID_DIA, VAR_Y,
CASE
WHEN RECARGA>0 THEN 1 ELSE 0 END RECARGA
FROM MODELO_REC_VAR_Y) A
GROUP BY NUMERO_TELEFONO
/
/*****************************************************************************************************************/
/*************************************UNIMOS LA SALIDA DE Y EN UNA SOLA VARIABLE*************************************/
/*************************************UNA VARIABLE CON LOS 28 DÍAS + 29 NO RECARGA*************************************/
/*****************************************************************************************************************/
```

```
DROP TABLE MODELO_REC_VAR_Y_TRANS_1_0_UV
/
CREATE TABLE MODELO_REC_VAR_Y_TRANS_1_0_UV AS
SELECT B.NUMERO_TELEFONO, NVL(VAR_Y,29) VAR_Y_DIA
FROM
(SELECT NUMERO_TELEFONO, VAR_Y
FROM MODELO_REC_VAR_Y
WHERE RECARGA>0) A, MODELO_REC_COLECTIVO_FIN B
WHERE A.NUMERO_TELEFONO(+)=B.NUMERO_TELEFONO
/
/************************************UNA VARIABLE CON LAS 4 SEMANAS + 5 NO RECARGA*********************************/
DROP TABLE MODELO_REC_VAR_Y_TRANS_1_0_SEM
/
CREATE TABLE MODELO_REC_VAR_Y_TRANS_1_0_SEM AS
SELECT B.NUMERO_TELEFONO, NVL(VAR_Y,5) VAR_Y_SEM
FROM
(SELECT NUMERO_TELEFONO,
CASE
WHEN VAR_Y BETWEEN 1 AND 7 THEN 1
WHEN VAR_Y BETWEEN 8 AND 14 THEN 2
WHEN VAR_Y BETWEEN 15 AND 21 THEN 3
WHEN VAR_Y BETWEEN 22 AND 28 THEN 4
ELSE 5 END VAR_Y
FROM MODELO_REC_VAR_Y
WHERE RECARGA>0) A, MODELO_REC_COLECTIVO_FIN B
WHERE A.NUMERO_TELEFONO(+)=B.NUMERO_TELEFONO
/
/************************UNA VARIABLE CON LOS 2 BLOQUES DE 2 SEMANAS + 3 NO RECARGA**************************/
DROP TABLE MODELO_REC_VAR_Y_TRANS_1_0_QN
/
CREATE TABLE MODELO_REC_VAR_Y_TRANS_1_0_QN AS
SELECT B.NUMERO_TELEFONO, NVL(VAR_Y,3) VAR_Y_QN
FROM
```

```
(SELECT NUMERO_TELEFONO,
CASE
WHEN VAR_Y BETWEEN 1 AND 14 THEN 1
WHEN VAR_Y BETWEEN 15 AND 28 THEN 2
ELSE 3 END VAR_Y
FROM MODELO_REC_VAR_Y
WHERE RECARGA>0) A, MODELO_REC_COLECTIVO_FIN B
WHERE A.NUMERO_TELEFONO(+)=B.NUMERO_TELEFONO
/
/**********UNA VARIABLE PARA EL MES (4 SEM) + 0 NO RECARGA**********/
DROP TABLE MODELO_REC_VAR_Y_TRANS_1_0_MES
/
CREATE TABLE MODELO_REC_VAR_Y_TRANS_1_0_MES AS
SELECT B.NUMERO_TELEFONO, NVL(VAR_Y,0) VAR_Y_MES
FROM
(SELECT NUMERO_TELEFONO,
CASE
WHEN VAR_Y BETWEEN 1 AND 28 THEN 1
ELSE 0 END VAR_Y
FROM MODELO_REC_VAR_Y
WHERE RECARGA>0) A, MODELO_REC_COLECTIVO_FIN B
WHERE A.NUMERO_TELEFONO(+)=B.NUMERO_TELEFONO
/
/**********/
/**********DIVIDIMOS EL FICHERO EN ENTRENAMIENTO VALIDACIÓN Y TEST**********/
/**********/
DROP TABLE MODELO_REC_COLECTIVO_EVT
/
CREATE TABLE MODELO_REC_COLECTIVO_EVT AS
SELECT
NUMERO_TELEFONO,
CASE
WHEN NUM_FILA <=12000 THEN 3   --> TEST
```

```
WHEN NUM_FILA <=24000 THEN 2  --> VALIDACIÓN

ELSE 1 END ENT_VAL_TES       --> ENTRENAMIENTO

FROM

(SELECT NUMERO_TELEFONO, ROWNUM NUM_FILA

FROM

(SELECT NUMERO_TELEFONO

FROM MODELO_REC_COLECTIVO_FIN

ORDER BY DBMS_RANDOM.VALUE) A)A

/

DROP TABLE MODELO_REC_VAR_Y_TRANS_1_0_EVT

/

CREATE TABLE MODELO_REC_VAR_Y_TRANS_1_0_EVT AS

SELECT

A.ENT_VAL_TES,

VAR_X1                  ,

VAR_X2                  ,

……..

VAR_X178                ,

VAR_X179                ,

VAR_X180                ,

VAR_Y_DIA               ,

VAR_Y_SEM                   ,

VAR_Y_QN                    ,

VAR_Y_MES

FROM MODELO_REC_COLECTIVO_EVT A, MODELO_REC_VAR_Y_TRANS_1_0_UV B, MODELO_REC_TRANS_180 C,
MODELO_REC_VAR_Y_TRANS_1_0_SEM D,

MODELO_REC_VAR_Y_TRANS_1_0_QN E, MODELO_REC_VAR_Y_TRANS_1_0_MES F

WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO

AND A.NUMERO_TELEFONO=C.NUMERO_TELEFONO

AND A.NUMERO_TELEFONO=D.NUMERO_TELEFONO

AND A.NUMERO_TELEFONO=E.NUMERO_TELEFONO

AND A.NUMERO_TELEFONO=F.NUMERO_TELEFONO

/
```

```
/*************************************************************************************************************************************/
/*************************************************XPORT FICHEROS**********************************************************************/
/*************************************************************************************************************************************/
CREATE TABLE MODELO_REC_SPOOL_VQUIN_TOCADA
AS
SELECT * FROM MODELO_REC_VAR_Y_TRANS_1_0_EVT;
/
DELETE MODELO_REC_SPOOL_VQUIN_TOCADA
WHERE ENT_VAL_TES=1 AND VAR_Y_QN=2
/
COMMIT
/
UPDATE MODELO_REC_SPOOL_VQUIN_TOCADA
SET VAR_Y_QN=0 WHERE VAR_Y_QN IN (2,3)
/
DROP TABLE MODELO_REC_MUESTRA_SALDO
/
CREATE INDEX INDX_MOD_REC_SALDO ON MODELO_REC_MUESTRA_SALDO(NUMERO_TELEFONO)
/
ANALYZE TABLE MODELO_REC_MUESTRA_SALDO ESTIMATE STATISTICS
/
/*************************************************************************************************************************************/
/*****************************************UNIFICACION INFORMACIÓN DE SALDOS***********************************************************/
/*************************************************************************************************************************************/
-- Confidencial
/*************************************************************************************************************************************/
/*************************************************MONTAMOS LOS SALDOS DIARIOS*********************************************************/
/*************************************************************************************************************************************/
DROP TABLE MODELO_REC_TRAF_AGRUP
/
CREATE TABLE MODELO_REC_TRAF_AGRUP
-- Confidencial
/
```

```
DROP TABLE MODELO_REC_SALDO
/
CREATE TABLE MODELO_REC_SALDO
AS
SELECT A.NUMERO_TELEFONO, A.ID_DIA, B.SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, MAX(FECHA_REGISTRO) FECHA_REGISTRO
FROM  MODELO_REC_TRAF_AGRUP
GROUP BY NUMERO_TELEFONO, ID_DIA) A, MODELO_REC_TRAF_AGRUP B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO
AND A.FECHA_REGISTRO=B.FECHA_REGISTRO
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
/*****************************************************************************************************************************/
/****************************************COMPLETAMOS LOS DÍAS QUE FALTA*******************************************************/
/*****************************************************************************************************************************/

DROP TABLE MODELO_REC_SALDO_DIA
/
CREATE TABLE MODELO_REC_SALDO_DIA
AS
SELECT A.NUMERO_TELEFONO, A.ID_DIA, B.SALDO
FROM
(SELECT B.NUMERO_TELEFONO, A.ID_DIA
FROM
(SELECT ID_DIA
FROM DW_CO_DIAS
WHERE ID_DIA BETWEEN --) A,
MODELO_REC_MUESTRA_SALDO B) A,
MODELO_REC_SALDO B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO (+)
AND A.ID_DIA=B.ID_DIA (+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
```

```
/
/*********************************PARA LOS DÍAS SIN SALDO LES PEGAMOS EL DEL ANTERIOR*****************************/
DROP TABLE MODELO_REC_SALDO_DIA_C1
/
CREATE TABLE MODELO_REC_SALDO_DIA_C1
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C2
/
CREATE TABLE MODELO_REC_SALDO_DIA_C2
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C1) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
```

```
FROM MODELO_REC_SALDO_DIA_C1) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C1
/
DROP TABLE MODELO_REC_SALDO_DIA_C3
/
CREATE TABLE MODELO_REC_SALDO_DIA_C3
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C2) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C2) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C2
/
DROP TABLE MODELO_REC_SALDO_DIA_C4
/
CREATE TABLE MODELO_REC_SALDO_DIA_C4
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
```

```
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C3) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C3) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C3
/
DROP TABLE MODELO_REC_SALDO_DIA_C5
/
CREATE TABLE MODELO_REC_SALDO_DIA_C5
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C4) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C4) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C4
/
DROP TABLE MODELO_REC_SALDO_DIA_C6
/
```

```
CREATE TABLE MODELO_REC_SALDO_DIA_C6
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C5) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C5) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C5
/
DROP TABLE MODELO_REC_SALDO_DIA_C7
/
CREATE TABLE MODELO_REC_SALDO_DIA_C7
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C6) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C6) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
```

```
/
DROP TABLE MODELO_REC_SALDO_DIA_C6
/
DROP TABLE MODELO_REC_SALDO_DIA_C8
/
CREATE TABLE MODELO_REC_SALDO_DIA_C8
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C7) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C7) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C7
/
DROP TABLE MODELO_REC_SALDO_DIA_C9
/
CREATE TABLE MODELO_REC_SALDO_DIA_C9
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C8) A,
```

```
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C8) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C8
/
DROP TABLE MODELO_REC_SALDO_DIA_C10
/
CREATE TABLE MODELO_REC_SALDO_DIA_C10
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C9) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C9) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C9
/
DROP TABLE MODELO_REC_SALDO_DIA_C11
/
CREATE TABLE MODELO_REC_SALDO_DIA_C11
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
```

```
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C10) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C10) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C10
/
DROP TABLE MODELO_REC_SALDO_DIA_C12
/
CREATE TABLE MODELO_REC_SALDO_DIA_C12
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C11) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C11) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C11
/
DROP TABLE MODELO_REC_SALDO_DIA_C13
```

```
/
CREATE TABLE MODELO_REC_SALDO_DIA_C13
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C12) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C12) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C12
/
DROP TABLE MODELO_REC_SALDO_DIA_C14
/
CREATE TABLE MODELO_REC_SALDO_DIA_C14
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C13) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C13) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
```

```
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C13
/
DROP TABLE MODELO_REC_SALDO_DIA_C15
/
CREATE TABLE MODELO_REC_SALDO_DIA_C15
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C14) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C14) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
DROP TABLE MODELO_REC_SALDO_DIA_C14
/
DROP TABLE MODELO_REC_SALDO_DIA_C16
/
CREATE TABLE MODELO_REC_SALDO_DIA_C16
AS
SELECT A.NUMERO_TELEFONO,
A.ID_DIA,
CASE WHEN A.SALDO IS NULL THEN B.SALDO
ELSE A.SALDO END SALDO
FROM
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
```

```
FROM MODELO_REC_SALDO_DIA_C15) A,
(SELECT NUMERO_TELEFONO, ID_DIA, SALDO, ROWNUM ID_COL
FROM MODELO_REC_SALDO_DIA_C15) B
WHERE A.NUMERO_TELEFONO=B.NUMERO_TELEFONO(+)
AND A.ID_COL-1=B.ID_COL(+)
ORDER BY A.NUMERO_TELEFONO, A.ID_DIA
/
/*****************************************************************************************************************/
/*****************************************************VARIABLES Z1-Z90*****************************************************/
/*****************************************************************************************************************/
DROP TABLE MODELO_REC_DIAS_VAR_Z
/
CREATE TABLE MODELO_REC_DIAS_VAR_Z
AS
SELECT NUMERO_TELEFONO, ID_DIA, NVL(SALDO,0) SALDO, MOD(ROWNUM,90) VAR_Z
FROM ( SELECT NUMERO_TELEFONO, ID_DIA, MIN(SALDO) SALDO
FROM MODELO_REC_SALDO_DIA_C16
WHERE ID_DIA BETWEEN TO_DATE('18/11/12','DD/MM/YY')-90+1 AND TO_DATE('18/11/12','DD/MM/YY')
GROUP BY NUMERO_TELEFONO, ID_DIA
ORDER BY NUMERO_TELEFONO, ID_DIA)
/
UPDATE MODELO_REC_DIAS_VAR_Z
SET VAR_Z=90 WHERE VAR_Z=0
/
COMMIT
/
DROP TABLE MODELO_REC_DIAS_VAR_Z90
/
CREATE TABLE MODELO_REC_DIAS_VAR_Z90 AS
SELECT
NUMERO_TELEFONO,
SUM(DECODE (VAR_Z,      1      ,      1      ,      0)*SALDO)              VAR_Z1   ,
SUM(DECODE (VAR_Z,      2      ,      1      ,      0)*SALDO)              VAR_Z2   ,
```

```
SUM(DECODE (VAR_Z,      3      ,      1      ,      0)*SALDO)              VAR_Z3   ,
SUM(DECODE (VAR_Z,      4      ,      1      ,      0)*SALDO)              VAR_Z4   ,
……
SUM(DECODE (VAR_Z,      88     ,      1      ,      0)*SALDO)              VAR_Z88  ,
SUM(DECODE (VAR_Z,      89     ,      1      ,      0)*SALDO)              VAR_Z89  ,
SUM(DECODE (VAR_Z,      90     ,      1      ,      0)*SALDO)              VAR_Z90
FROM MODELO_REC_DIAS_VAR_Z
GROUP BY NUMERO_TELEFONO
/
DELETE MODELO_REC_DIAS_VAR_Z90
WHERE NUMERO_TELEFONO IN (SELECT NUMERO_TELEFONO
                          FROM MODELO_REC_DIAS_VAR_Z
                          WHERE SALDO>=100)
/
```

A7.- Código Octave, lanzador modelos RL

```
function [F1_Ms,v_prob_05,pred_05,all_theta] = RC_LG_LANZADOR(Matriz_trabajo, ...
                                                        B, ...
                                                        T, ...
                                                        var_X, ...
                                                        var_y, ...
                                                        lambda,MaxIter)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%
%           REGRESIÓN LOGÍSTICA CON FUNCIÓN OPT. fmincg.m
%
%             ENE-2013
%
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%         ESTE PROGRAMA NECESITA LAS FUNCIONES
%         ------------------------------------------------------
%
%   lrCostFunction.m (logistic regression cost function)
%   oneVsAll.m
%   predictOneVsAll.m
%   predict.m
%   RC_F1_score.m
%   RC_MATTHEWS_score.m
%   rocPMTK.m -> función plor ROC
%   RC_Balanceo_matriz_tm -> función para balancear la muestra
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%            ENTRADAS
%
```

```
%   El proceso recibe una matriz (Matriz_trabajo) que debe contener:
%
%   1 columna: 1,2,3 correspondiente a entrenamiento, validación y test
%   n columanas con los datos de entrada X
%   1 columna final con el vector del target, con valores posibles de 1 a n
%
%   B: Se le pasa el % de unos que queremos que tenga la matriz de entrenamien
%     si el campo lleva el valor 0, no ejecutamos valanceo
%   T: en caso de que sea multitarget sobre que target trabajamos
%     si T es igual a 0 trabajamos sobre todos a la vez, pero en ese caso no
%     aplicamos balanceo.
%
%   var_X: vector con las columnas de las variables de entrada
%   var_Y: columna con la variable target
%   lambada: parámetro de regularización
%   MaxIter: número máximo de iteraciones del algoritmo de busqueda de min.
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%                       SALIDAS
%
%   F1_Ms: resumen de las principales métricas del modelo ajustado teniendo
%        en cuenta diferente probabilidades de corte.
%   v_prob: vector con la probabilidad de cada caso de pertenece a la clase 1,
%                           para problemas con target binario, para problemas multitarget
%                                   vector con la clasificación de pertenencia a los targets.
%   salida_reg_log_XXXXXXX.text: fichero con información de la ejecución
%                           por defecto se guarda en el directorio
%                                                                   Octave\3.2.4_gcc-4.4.0\bin          %
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%                                           EJEMPLO DE USO
```

```
%
% RC_LG_LANZADOR(Matriz_trabajo,
%                    B,            -> % del balanceo de los ceros del target
%                    T,            -> valor del target 1,2,3....
%                    var_X,   -> matriz col var independientes [2:181]
%      var_y,    -> columna var dependiente  ej: 185
%      lambda,    -> parámetro de regularización ej: 1
%      MaxIter)   -> num max iteracciones fmincg ej:50
%
%  Ejecución para target {0,1} sin balancear
%  [F1_Ms,v_prob,pred]=RC_LG_LANZADOR(Recargas,0,0,[2:181],185,1,50);
%  Ejecución para target {0,1} balanceado al 55% de ceros
%  [F1_Ms,v_prob,pred]=RC_LG_LANZADOR(Recargas,0.55,0,[2:181],185,1,50);
%  Ejecución para target {1,2,3} con 3=0 sin balanceo
%  [F1_Ms,v_prob,pred]=RC_LG_LANZADOR(Recargas,0,3,[2:181],184,1,50);
%  Ejecución para target {1,2,3} con 3=0 balanceado al 30% de ceros
%  [F1_Ms,v_prob,pred]=RC_LG_LANZADOR(Recargas,0.3,3,[2:181],184,1,50);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clc;
close;
fprintf('COMENZANDO EN AJUSTE DE LA REGRESION LOGISTICA...\n')
%%%%%%%%%%%%%%%%%%%%BALANCEO DE LA MUESTRA%%%%%%%%%%%%%%%%%%%%%%%%%%%
Matriz_trabajo_sin_B=Matriz_trabajo;
if B==0
fprintf('NO SE BALANCEA LA MUESTRA\n');
else
fprintf('EJECUTAMOS EL BALANCEO DE LA MUESTRA...\n');
y_ent_sin_B=Matriz_trabajo(Matriz_trabajo(:,1)==1,var_y);
Matriz_trabajo= RC_Balanceo_matriz_tm(B,T,Matriz_trabajo,var_y);
endif;

%%%%%%%%%%%%%%%%%EXTRAEMOS LOS DATOS DE ENTRENAMIENTO%%%%%%%%%%%%%%%%%%%%
X=Matriz_trabajo(Matriz_trabajo(:,1)==1,var_X);
```

```
y=Matriz_trabajo(Matriz_trabajo(:,1)==1,var_y);
%%%%%%%%%%%%%%%%%%AJUSTAMOS LA REGRESION LOGISTICA%%%%%%%%%%%%%%%%%%%%%%%%
fprintf('\nLANZAMOS EL OPTIMIZADOR DE LA REGRESION...\n')
if T==0
num_labels=size(unique(y),1)-1;
else
num_labels=size(unique(y),1);
endif;
[all_theta] = oneVsAll(X, y, num_labels, lambda,MaxIter);
%%%%%%%%%%%%%%%%%%EXTRAEMOS LOS DATOS DE TEST%%%%%%%%%%%%%%%%%%%%%%%%
X_test=Matriz_trabajo(Matriz_trabajo(:,1)==3,var_X);
y_test=Matriz_trabajo(Matriz_trabajo(:,1)==3,var_y);
%%%%%%%%%%%%%%%%CALCULAMOS LA PREVISIÓN CONP=0.5%%%%%%%%%%%%%%%%%%%%%%%%
[v_prob_05,pred_05] = predictOneVsAll(all_theta,X_test, 0.5);
%%%%%%%%%%%%%%%%CALCULAMOS LA PROBABILIDAD A POSTERIORI%%%%%%%%%%%%%%%%%%
if B!=0 & num_labels==1
fprintf('AJUSTAMOS LA PROBABILIDAD A POSTERIORI DEBIDO AL BALANCEO...\n')
v_prob_sin_modificar=v_prob_05;
v_prob=RC_real_post_prob(v_prob_05,y_ent_sin_B,y);
endif;
%%%%%%%%%%%%%ANALIZAMOS EL ACIERTO EN FUNCIÓN DE LA PROB DE CORTE%%%%%%%%%%%%%%%%
if num_labels==1
vector_prob=[0.5,0.6,0.7,0.8,0.9];
for i=1 : size(vector_prob,2)
fprintf('CALCULANDO LA CORRELACION DE MATHEWS PARA %10.2f...\n',vector_prob(i));
[v_prob,pred] = predictOneVsAll(all_theta, X_test,vector_prob(i));
[Ms, accuracy,precision, recall]=RC_MATTHEWS_score(y_test,pred);
F1_Ms(i,:)=[vector_prob(i) Ms accuracy precision recall size(pred,1) sum(pred==1)...
 (sum(pred==1)./size(pred,1)).*100 ];
%%%%%%%%%%%%%CURVAS ROC PARA LOS CORTE EN FUNCION PROBABILIDAD%%%%%%%%%%%%%%%%
hold on;
RC_ROC_mt(v_prob,pred, y_test,T,i,vector_prob(i));
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
endfor;
hold off;
fprintf('ESPERANDO PULSE ENTER...\n');
pause;
else
fprintf('CALCULANDO EL F1 SCORE...\n');
[v_prob,pred] = predictOneVsAll(all_theta, X_test,0);
[F1, accuracy,precision, recall]=RC_F1_score(y_test,pred);
label_0=max(unique(y_test)).*ones(size(y_test,1),1);
F1_Ms=[0 F1 accuracy precision recall size(pred,1) sum(pred!=label_0)...
 (sum(pred!=label_0)./size(pred,1)).*100 ];
endif;
%%%%%%%%%%%%CURVA ROC DEL MODELO GENERICO P=0.5%%%%%%%%%%%%%%%%%%%%%%%%%
close;
hold on;
RC_ROC_mt(v_prob_05,pred_05, y_test,T,1,0.5);
hold off;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
filename=char(strftime ("salida_reg_log_%d_%m_%H_%M_%S.txt", localtime (time ())));
fid=fopen(filename,"w");
fprintf(fid,"                    EJECUCION AJUSTE REGRESION LOGISTICA\n");
fprintf(fid,"                    -------------------------------------\n");
fecha_hora=char(strftime ("%R:%S %A %e %B %Y", localtime (time ())));
fprintf(fid,"REALIZADO A LAS: '%s' \n",fecha_hora);
if num_labels==1
y_ent_prop_1=(sum(Matriz_trabajo_sin_B(Matriz_trabajo_sin_B(:,1)==3,var_y)==1)./...
        size(Matriz_trabajo_sin_B(Matriz_trabajo_sin_B(:,1)==3,var_y),1)).*100;
fprintf(fid,"LA PROPORCION DE POSITIVOS EN LA MUESTRA DE TEST: %10.2f%%\n",y_ent_prop_1);
else
[u, i, j] = unique (Matriz_trabajo_sin_B(Matriz_trabajo_sin_B(:,1)==3,var_y));
M_salida=[u accumarray(j, 1) (accumarray(j, 1)./size(Matriz_trabajo_sin_B(...
Matriz_trabajo_sin_B(:,1)==3,var_y),1)).*100];
fprintf(fid,"\n LA PROPORCION EN LA MUESTRA DE TEST: \n");
```

```
fprintf(fid,' %10.0f %10.0f %10.2f%% \n', M_salida');
endif;
if B==0
fprintf(fid,"NO SE BALANCEA LA MUESTRA\n");
else
[u, i, j] = unique (Matriz_trabajo(Matriz_trabajo(:,1)==1,var_y));
M_salida=[u accumarray(j, 1) (accumarray(j, 1)./size(Matriz_trabajo(...
Matriz_trabajo(:,1)==1,var_y),1)).*100];
fprintf(fid,"\n LA PROPORCION EN LA MUESTRA DE ENTRENAMIENTO: \n");
fprintf(fid,' %10.0f %10.0f %10.2f%% \n', M_salida');
endif;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if num_labels==1
fprintf(fid,'\n  P. CORTE   MATTHEWS C     ACCURACY        PRECISION          RECALL              N   SELECCION  P SEL TOTAL\n');
else
fprintf(fid,'\n  P. CORTE          F1     ACCURACY        PRECISION          RECALL              N   SELECCION  P SEL TOTAL\n');
endif;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
printf("\nGUARDANDO INFORMACION EN FICHERO: '%s' \n", filename)
fprintf(fid,"%10.2f\t %10.2f\t %10.2f%%\t %10.2f%%\t %10.2f%%\t %10.0f\t %10.0f\t %10.2f%%\t \n",F1_Ms');
fprintf(fid,"\n MATRIZ DE CONFUSION:\n");
fprintf(fid,"\n PRED FIL \ REAL COL \n");
Mat_confusion=RC_Matriz_Confusion(y_test,pred_05);
dlmwrite(filename,Mat_confusion,'delimiter','\t','-append');
fclose(fid);
all_theta
end;
```

A8.- Código Octave, F1 Score

```
function [F1,accuracy,precision,recall] = RC_F1_score(yt,pred)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%                  RC_F1_SCORE
%
%  Esta función calcula el F1_score de ajuste del modelo partiendo de dos
%  vectores, el de datos reales y el de datos estimados.
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%                  ENTRADAS
%
%   El proceso recibe dos vectores columna:
%
%   yt: vector con los datos de target de la muestra de test, la función
%      no realiza el filtrado del conjunto de de test, hay que pasar el
%      vector filtrado
%
%
%   pred: vector con la clasificación de las predicciones del modelo para
%       la muestra de test, hay que pasar el vetor columna filtrado
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%                  SALIDAS
%
%  F1: valor del F1_score
%  precision: % de acierto sobre clasificación de 1
%  recall: % del total de 1 de la muestra cuantos estamos clasificando bien
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
%
%                                        EJEMPLO DE USO
%
%  RC_F1_score(yt,pred)
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

accuracy=mean(double(pred == yt)).*100;

if length(unique(yt))<=2
precision=(sum((pred+yt)==2)/(sum((pred+yt)==2)+sum(((pred==1)+(yt==0))==2))).*100;
recall=(sum((pred+yt)==2)/(sum((pred+yt)==2)+sum(((pred==0)+(yt==1))==2))).*100;
F1=2*(precision.*recall)/(precision.+recall);
else
label_0=max(unique(yt)).*ones(size(yt,1),1);
precision=(sum(pred==yt & yt!=label_0)./sum(pred!=label_0)).*100;
recall   =sum(pred==yt & yt!=label_0)./(sum(yt!=label_0)).*100;
F1=2*(precision.*recall)/(precision.+recall);
End
```

A9.- Código Octave, Matthews Score

```
function [Ms accuracy precision recall] = RC_MATTHEWS_score(yt,pred)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%            MATTHEWS CORRELATION COEFFICIENT
%            --------------------------------
%  Esta función calcula el Ms (MCC) coeficiente de correlación de Matthews.
%  Se usa en machine learning para medir la calidad de clasificacoder bianrios.
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%                 ENTRADAS
%                 --------
%   El proceso recibe dos vectores columna:
%
%   yt: vector con los datos de target de la muestra de test, la función no realiza el filtrado del conjunto de de test, hay que pasar el
%  vector filtrado
%   pred: vector con la clasificación de las predicciones del modelo para la muestra de test, hay que pasar el vetor columnafiltrado
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                 SALIDAS
%  Ms: valor de la correlación de Matthews
%  precision: % de acierto sobre clasificación de 1
%  recall: % del total de 1 de la muestra cuantos estamos clasificando bien
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%                                 EJEMPLO DE USO
%  RC_MATTHEWS_score(yt,pred)
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
accuracy=mean(double(pred == yt)).*100;

precision=(sum((pred+yt)==2)/(sum((pred+yt)==2)+sum(((pred==1)+(yt==0))==2))).*100;

recall=(sum((pred+yt)==2)/(sum((pred+yt)==2)+sum(((pred==0)+(yt==1))==2))).*100;

TP=sum((pred+yt)==2);

TN=sum((pred+yt)==0);

FP=sum(((pred==1)+(yt==0))==2);

FN=sum(((pred==0)+(yt==1))==2);

Ms=((TP.*TN)-(FP.*FN))./sqrt((TP+FP).*(TP+FN).*(TN+FP).*(TN+FN));

End
```

A10.- Código Octave, lanzador MLP

```
function [F1_Ms,v_prob ] = RC_MLP_LANZADOR(Matriz_trabajo, ...
                                            B, ...
                                            T, ...
                                             estandarizar,...
                                            var_X,...
                                            var_y,...
                                           hidden_layer_size,...
                                            lambda,
                                            MaxIter)


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%               PERCEPTRON MULTICAPA CON BACK PROPAGATION y FUNCIÓN OPT. fmincg.m
%                       ENE-2013
%
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%                ESTE PROGRAMA NECESITA LAS FUNCIONES
%
%    nnCostFunction.m -> función de coste de la red neurnal
%    randInitializeWeights.m -> inicializa las matrices de pesos de forma  aleatoria.
%    RC_Balanceo_matriz_tm -> función para balancear la muestra
%    predict.m
%    RC_F1_score.m
%    RC_MATTHEWS_score.m
%    rocPMTK.m -> función plor ROC
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%                       ENTRADAS
```

```
%   El proceso recibe una matriz (Matriz_trabajo) que debe contener:
%
%   1 columna: 1,2,3 correspondiente a entrenamiento, validación y test
%   n columnas con los datos de entrada X
%   1 columna final con el vector del target, con valores posibles de 1 a n
%
%   B: Se le pasa el % de unos que queremos que tenga la matriz de entrenamiento
%     si se le pasa 0 no realiza balanceo.
%   T: en caso de que sea multitarget sobre que target trabajamos
%     si T es igual a 0 trabajamos sobre todos a la vez, pero en ese caso no
%     aplicamos balanceo
%
%   hidden_layer_size: neuronas en la capa oculta
%   lambada: parámetro de regularización
%   MaxIter: número máximo de iteraciones del algoritmo de busqueda de min.
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                       SALIDAS
%   F1_Ms: resumen de las principales métricas del modelo ajustado teniendo
%        en cuenta diferente probabilidades de corte.
%   v_prob: vector con la probabilidad de cada caso de pertenece a la clase 1,
%                para problemas con target binario, para problemas multitarget
%                vector con la clasificación de pertenencia a los targets.
%   salida_reg_log_XXXXXXX.text: fichero con información de la ejecución por defecto se guarda en el directorio
%              Octave\3.2.4_gcc-4.4.0\bin
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%                                                   EJEMPLO DE USO
% RC_MLP_LANZADOR(Matriz_trabajo,
%              B, -> % del balanceo del target, 0 no balanceo
%              T, -> valor del target si es 0 multiclase
%              var_X,      -> matriz col var independientes [2:181]
%             var_y,      -> columna var dependiente  ej: 185
```

```
%           hidden_layer_size -> neuronas en capa oculta
%           lambda,     -> parámetro de regularización ej: 1
%           MaxIter)    -> num max iteracciones fmincg ej:50
%
%  [F1_Ms,v_prob]=RC_MLP_LANZADOR(Recargas,0,0,'NO',2:181],185,25,1,2)
%  [F1_Ms,v_prob]=RC_MLP_LANZADOR(Recargas,0,0,'SI',2:181],185,25,1,2)
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tic;
clc;
close;
fprintf('COMENZANDO EN AJUSTE DEL PERCEPTRON MULTICAPA...\n')
%%%%%%%%%%%%%%%%%%%%%%%%%BALANCEO DE LA MUESTRA%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Matriz_trabajo_sin_B=Matriz_trabajo;
if B==0
fprintf('NO SE BALANCEA LA MUESTRA\n');
else
fprintf('EJECUTAMOS EL BALANCEO DE LA MUESTRA...\n');
y_ent_sin_B=Matriz_trabajo(Matriz_trabajo(:,1)==1,var_y);
Matriz_trabajo= RC_Balanceo_matriz_tm(B,T,Matriz_trabajo,var_y);
endif;
%%%%%%%%%%%%%%%%%%%%%%EXTRAEMOS LOS DATOS DE ENTRENAMIENTO%%%%%%%%%%%%%%%%%%%%
input_layer_size=length(var_X);
X=Matriz_trabajo(Matriz_trabajo(:,1)==1,var_X);
y=Matriz_trabajo(Matriz_trabajo(:,1)==1,var_y);
%%%%%%%%%%%%%%%%%%ESTANDARIZAMOS LOS DATOS DE ENTRENAMIENTO%%%%%%%%%%%%%%%%%%%%
if estandarizar=='SI'
fprintf('\nESTANDARIZANDO LOS DATOS DE ENTRANEMIENTO A NIVEL DE REGISTRO...\n');
X=RC_Estandar_Reg(X);
else
fprintf('\nCUIDADO LOS DATOS DE ENTRADA NO ESTÁN ESTANDARIZADOS...\n');
endif;
%%%%%%%%%%%%%%%INICIALIZAMOS LAS MATRICES DE PESOS DE FORMA ALEATORIA%%%%%%%%%%%%%
```

```
fprintf('\nLANZAMOS EL OPTIMIZADOR DE LA NN...\n')
if T==0
num_labels=size(unique(y),1)-1;
else
num_labels=size(unique(y),1);
endif;
initial_Theta1 = randInitializeWeights(input_layer_size, hidden_layer_size);
initial_Theta2 = randInitializeWeights(hidden_layer_size, num_labels);
initial_nn_params = [initial_Theta1(:) ; initial_Theta2(:)];
%%%%%%%%%%%%%%%%%%%%%%%%%%EJECUTAMOS EL MLP%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


options = optimset('MaxIter', MaxIter);
costFunction = @(p) nnCostFunction(p, ...
                        input_layer_size, ...
                        hidden_layer_size, ...
                        num_labels, X, y, lambda);
[nn_params, cost] = fmincg(costFunction, initial_nn_params, options);
%%%%%%%%%%%%%%%%%%%%%%%%%%VISUALIZAMOS EL COSTE DE LA NN%%%%%%%%%%%%%%%%%%%%%%%%%%
figure(1);
hold on;
plot(cost);
title('EVOLUCION DE LA FUNCION DE COSTE');
hold off;
%%%%%%%%%%%%%%%%%%%%OBTENEMOS LAS MATRICES DE PESOS DEL ENTREANMIENTO%%%%%%%%%%%%
Theta1 = reshape(nn_params(1:hidden_layer_size * (input_layer_size + 1)), ...
            hidden_layer_size, (input_layer_size + 1));


Theta2 = reshape(nn_params((1 + (hidden_layer_size * (input_layer_size + 1))):end), ...
            num_labels, (hidden_layer_size + 1));
%%%%%%%%%%%%%%%%%%%%%%%%%%EXTRAEMOS LOS DATOS DE TEST%%%%%%%%%%%%%%%%%%%%%%%%%%
X_test=Matriz_trabajo(Matriz_trabajo(:,1)==3,var_X);
y_test=Matriz_trabajo(Matriz_trabajo(:,1)==3,var_y);
%%%%%%%%%%%%%%%%%ESTANDARIZAMOS LOS DATOS DE TEST%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
if estandarizar=='SI'
fprintf('ªnESTANDARIZANDO LOS DATOS DE TEST A NIVEL DE REGISTRO...\n');
X_test=RC_Estandar_Reg(X_test);
else
fprintf('\nNO ESTANDARIZAMOS LOS DATOS DE TEST...\n');
endif;
%%%%%%%%%%%%%%%%%%%%CALCULAMOS LA PREVISIÓN PARA TEST%%%%%%%%%%%%%%%%%%%%
[pred_05,v_prob_05] = predict(Theta1, Theta2, X_test,0.5);
%%%%%%%%%%%%%%%%%%%%CALCULAMOS LA PROBABILIDAD A POSTERIORI%%%%%%%%%%%%%%%%%%%%
if B!=0 & num_labels==1
fprintf('AJUSTAMOS LA PROBABILIDAD A POSTERIORI DEBIDO AL BALANCEO...\n')
v_prob_sin_modificar=v_prob_05;
v_prob=RC_real_post_prob(v_prob_05,y_ent_sin_B,y);
endif;
%%%%%%%%%%%%ANALIZAMOS EL ACIERTO EN FUNCIÓN DE LA PROB DE CORTE%%%%%%%%%%%%%%%%
if num_labels==1
vector_prob=[0.5,0.6,0.7,0.8,0.9];
for i=1 : size(vector_prob,2)
fprintf('CALCULANDO LA CORRELACION DE MATHEWS PARA %1.2f...\n',vector_prob(i));
[pred,v_prob] = predict(Theta1,Theta2, X_test,vector_prob(i));
[Ms, accuracy,precision, recall]=RC_MATTHEWS_score(y_test,pred);
F1_Ms(i,:)=[vector_prob(i) Ms accuracy precision recall size(pred,1) sum(pred==1)...
 (sum(pred==1)./size(pred,1)).*100 ];
%%%%%%%%%%%%%%%%%%CURVAS ROC PARA LOS CORTE EN FUNCION PROBABILIDAD%%%%%%%%%%%%%%%%
figure(2);
hold on;
RC_ROC_mt(v_prob,pred, y_test,T,i,vector_prob(i));
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
endfor;
hold off;
fprintf('ESPERANDO PULSE ENTER...\n');
pause;
else
```

```
fprintf('CALCULANDO EL F1 SCORE...\n');
[pred,v_prob] = predict(Theta1,Theta2, X_test,0);
[F1, accuracy,precision, recall]=RC_F1_score(y_test,pred_05);
label_0=max(unique(y_test)).*ones(size(y_test,1),1);
F1_Ms=[0 F1 accuracy precision recall size(pred,1) sum(pred_05!=label_0)...
 (sum(pred_05!=label_0)./size(pred,1)).*100 ];
endif;
%%%%%%%%%%%%%%%%%%CURVA ROC DEL MODELO GENERICO P=0.5%%%%%%%%%%%%%%%%%%%%
figure(3);
hold on;
RC_ROC_mt(v_prob_05,pred_05, y_test,T,1,0.5);
hold off;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
filename=char(strftime ("salida_nn_mlp_%d_%m_%H_%M_%S.txt", localtime (time ())));
fid=fopen(filename,"w");
fprintf(fid,"                    EJECUCION AJUSTE RED NEURONAL MLP\n");
fprintf(fid,"                    ----------------------------------\n");
fecha_hora=char(strftime ("%R:%S %A %e %B %Y", localtime (time ())));
fprintf(fid,"REALIZADO A LAS: '%s' \n",fecha_hora);
%%%%%%%%%%%%%%%%%%%%%%%%INFORMACIÓN SOBRE LA EJECUCIÓN%%%%%%%%%%%%%%%%%%%%%
tiempo_eje=toc;
fprintf(fid,"\nINFORMACIÓN SOBRE LA EJECUCIÓN: \n");
fprintf(fid,"\nTIEMPO DE EJECUCIÓN: %5.2f minutos\n", tiempo_eje/60);
fprintf(fid,"\nPARAMETROS DEL MODELO:\n");
fprintf(fid,"RC_LG_LANZADOR(Matriz,%1.2f,%1.0f,%1.0f,var_X,var_y,%2.0f,%1.3f,%3.0f)\n",...
B,T, estandarizar,hidden_layer_size,lambda,MaxIter);
fprintf(fid,"\nVARIABLES EMPLEADAS:\n");
fprintf(fid,"VARIABLE INDEPENDIENTE: %i\n", var_X);
fprintf(fid,"\nVARIABLE DEPENDIENTE:\n");
fprintf(fid,"VARIABLE INDEPENDIENTE: %i\n", var_y);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if num_labels==1
y_ent_prop_1=(sum(Matriz_trabajo_sin_B(Matriz_trabajo_sin_B(:,1)==3,var_y)==1)./...
```

```
    size(Matriz_trabajo_sin_B(Matriz_trabajo_sin_B(:,1)==3,var_y),1)).*100;
fprintf(fid,"\nLA PROPORCION DE POSITIVOS EN LA MUESTRA DE TEST: %10.2f%%\n",y_ent_prop_1);
else
[u, i, j] = unique (Matriz_trabajo_sin_B(Matriz_trabajo_sin_B(:,1)==3,var_y));
M_salida=[u accumarray(j, 1) (accumarray(j, 1)./size(Matriz_trabajo_sin_B(...
Matriz_trabajo_sin_B(:,1)==3,var_y),1)).*100];
fprintf(fid,"\nLA PROPORCION EN LA MUESTRA DE TEST: \n");
fprintf(fid,' %10.0f %10.0f %10.2f%% \n', M_salida');
endif;
if B==0
fprintf(fid,"\nNO SE BALANCEA LA MUESTRA\n");
else
[u, i, j] = unique (Matriz_trabajo(Matriz_trabajo(:,1)==1,var_y));
M_salida=[u accumarray(j, 1) (accumarray(j, 1)./size(Matriz_trabajo(...
Matriz_trabajo(:,1)==1,var_y),1)).*100];
fprintf(fid,"\n LA PROPORCION EN LA MUESTRA DE ENTRENAMIENTO: \n");
fprintf(fid,' %10.0f %10.0f %10.2f%% \n', M_salida');
endif;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if num_labels==1
fprintf(fid,'\n  P. CORTE   MATTHEWS C     ACCURACY        PRECISION          RECALL              N   SELECCION  P SEL TOTAL\n');
else
fprintf(fid,'\n  P. CORTE          F1     ACCURACY        PRECISION          RECALL              N   SELECCION  P SEL TOTAL\n');
endif;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
printf("\nGUARDANDO INFORMACION EN FICHERO: '%s' \n", filename)
fprintf(fid,"%10.2f\t %10.2f\t %10.2f%%\t %10.2f%%\t %10.2f%%\t %10.0f\t %10.0f\t %10.2f%%\t \n",F1_Ms');
fprintf(fid,"\n MATRIZ DE CONFUSION:\n");
fprintf(fid,"\n PRED FIL \ REAL COL \n");
Mat_confusion=RC_Matriz_Confusion(y_test,pred_05);
dlmwrite(filename,Mat_confusion,'delimiter','\t','-append');
fclose(fid);
end;
```

Referencias y Bibliografía

[1] DeGroot, Morris H. : Probabilidad y Estadística, Addison-Wesley Iberoamericana, 565-567.

[2] van Rijsbergen, C.J.(1979): Information Retrieval

[3] Powers, David M W (2007/2011): Evaluation: From Precision, Recall and F-Factor to ROC, Informedness Markedness & Correlation.

[4] Kattamuri S. Sarma: (4.7.2 Adjusting the Predicted Probabilities for Over-sampling). Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications.

[5] Hernández Orallo. J y otros: Introducción a la Minería de Datos.

[6] Cuadras C. M.: Métodos de Análisis Multivariante.

[7] DTW(febrero 2003): http://en.wikipedia.org/wiki/Dynamic_time_warping

[8] Bishop, M.: Pattern Recognition and Machine Learning, Springer.