Proyecto Vodafone COPS

Predicción de llamadas por motivos de facturación

Entregable del Hito 1

(Última modificación 26/03/2018)

Indice

Introducción

Descripción

El objetivo del proyecto Vodafone COPS es predecir a nivel de cliente, si dicho cliente va a realizar una llamada por motivos relacionados con su facturación (principalmente desacuerdos en la factura) al contact center de Vodafone durante el próximo ciclo de facturación. Esta situación supone un problema para Vodafone al producirse importantes picos de llamadas que sobrecargan la capacidad del contact center para poder dar servicio a todas de forma correcta.

Objetivos

Este proyecto tiene por objetivos:

- La mejora del modelo existente a nivel de poder predictivo y metodologías utilizadas.
- Mejorar el conocimiento de los motivos que desencadenan estas llamadas (interpretabilidad).
- El desarrollo de un modelo que permita a Vodafone obtener un público objetivo, tomar la iniciativa y llevar a cabo acciones preventivas, como llamar de forma proactiva a los clientes con mayor probabilidad de llamada para evitar así una llamada posterior.

Funcionalidades

Datamart analítico

El datamart analítico es un sistema de información que contiene todas las variables relevantes para la creación del modelo. Se genera a partir de la información disponible en la plataforma de Big Data (Big Data Platform o BDP) de Vodafone. El propio datamart se encuentra persistido también en la BDP y ha de ser actualizado periódicamente, permitiendo así obtener nuevas versiones del modelo que incorporen los últimos datos acerca de los clientes y la situación del mercado.

El datamart se crea a partir de información disponibilizada en la BDP:

- Log de interacciones: se trata de la tabla donde cada contacto del cliente con vodafone deja un registro. Por cada registro se tiene información del msisdn que ha realizado la llamada y una categorización de los motivos de la misma. Dicha categorización es realizada por el personal del contact center.
- Maestro de clasificación de interacciones: es una tabla de mapeo de las interacciones a unas clasificaciones propias del equipo de Vodafone COPS (Customer Operations).
- Fuente de datos de clientes de postpago: dicha fuente de datos contiene información mensualizada de los clientes de diverso tipo. Ciclo de facturación al que pertenece,

titular de la línea, tarifa de voz y datos, información socioeconómica de dicho cliente, etc.

La información incorporada en el datamart analítico se agrupa en diversas categorías:

- Identificación única de la observación: en base al msisdn, número de identificación y su ciclo de facturación.
- Interacciones: conteos de llamadas realizadas por el cliente por diversos motivos (facturación, churn, incidencias, etc.).
- Planes de voz y datos: información de los planes del cliente en cada ciclo de facturación y de si se ha producido un cambio en el mismo de forma reciente.
- Promociones: información de los códigos de promoción que el cliente tiene aplicados, así como de los meses restantes para la finalización de dicha promoción.
- Entorno social: información del lugar de procedencia del usuario de la línea, género, nacionalidad, edad, etc.
- Productos: información relacionada con los productos de vodafone, como el número de líneas de postpago y de prepago.

Nota importante: la puesta en producción de este modelo requiere que las tablas de origen utilizadas para la creación del datamart se encuentren actualizadas a principio de cada ciclo de facturación con un lag máximo de unas 48 o 72 horas.

Probabilidad de llamada por motivos de facturación

La predicción de la probabilidad de llamada por motivo de facturación evalúa la posibilidad de que un cliente realice una llamada por no conformidad con su factura en una ventana temporal de un ciclo de facturación.

Los clientes de Vodafone se encuentran segmentados en 4 ciclos de facturación diferentes en función del día del mes en el que se emite su factura. Dichos ciclos son:

- Día 1 de cada mes.
- Día 8 de cada mes.
- Día 15 de cada mes.
- Día 22 de cada mes.

Para cada cliente, se desea predecir la probabilidad de que realice al menos una llamada por motivos de facturación en el próximo ciclo de facturación, contando con información actualizada a cierre del ciclo de facturación actual.

El número de llamadas recibidas en el contact center de Vodafone según el ciclo de facturación se distribuye de la siguiente manera:

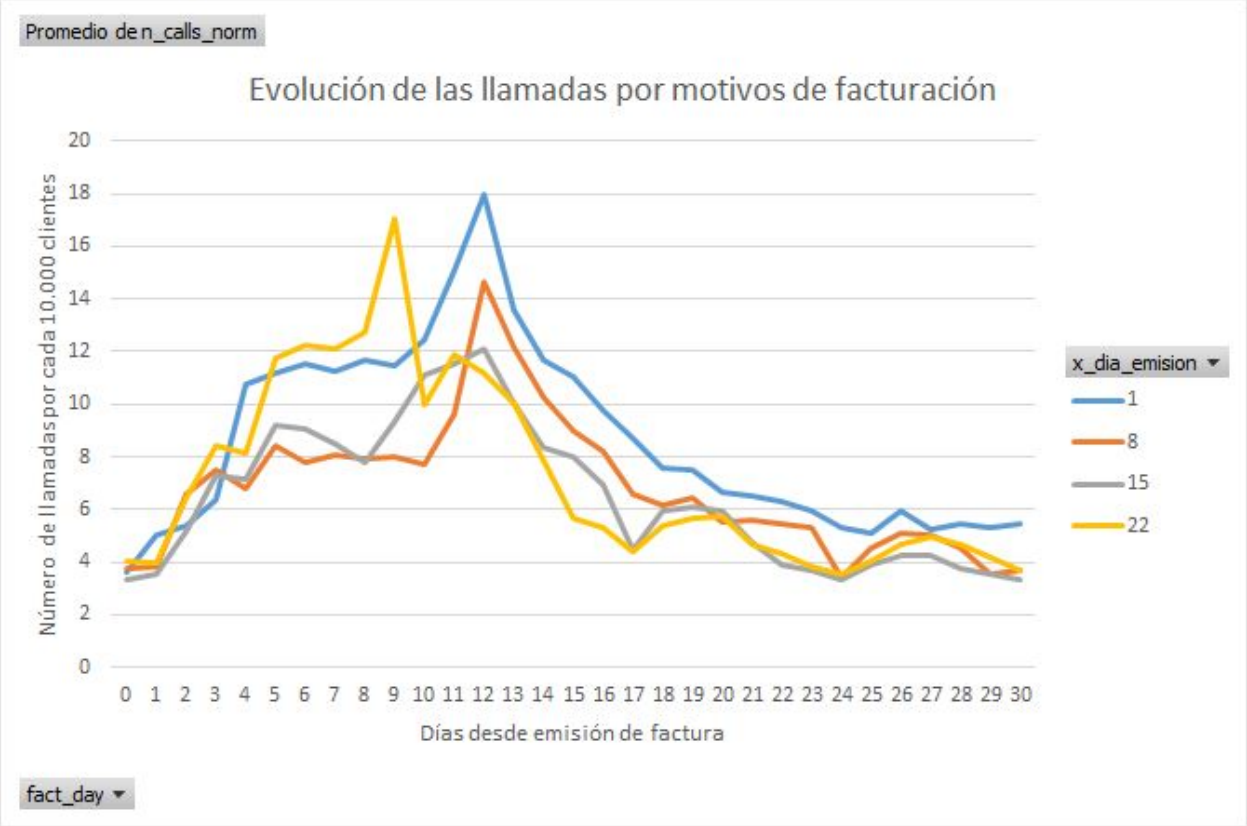

Figura 1. Evolución de las llamadas por motivos de facturación

Como se observa en la siguiente imagen, a partir del día de emisión de la factura, el número de llamadas aumenta hasta alcanzar un máximo alrededor del décimo día.

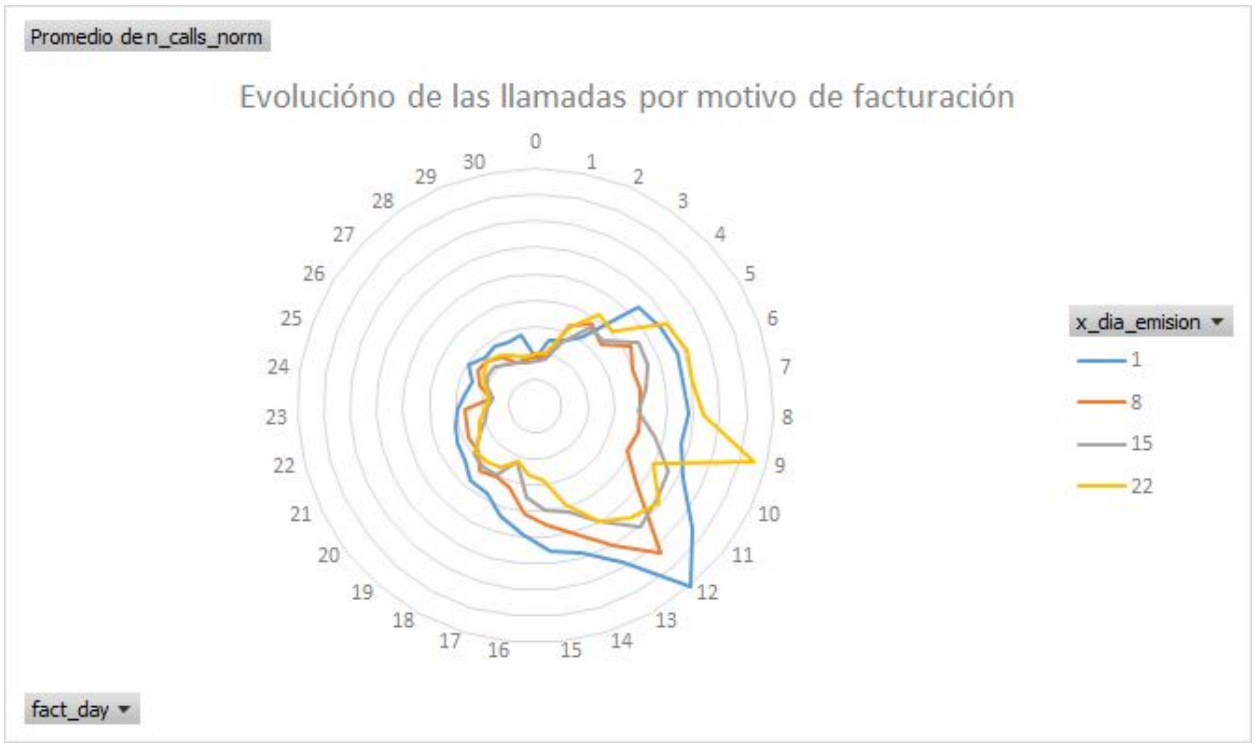

Figura 2. Evolución de las llamadas por motivos de facturación. Visualización radial.

Asimismo, se aprecia un mayor ratio de llamadas en los clientes pertenecientes al ciclo 1 de facturación frente al resto.
A nivel diario y durante todo 2017 y principio de 2018, la evolución de las llamadas se representa a continuación:

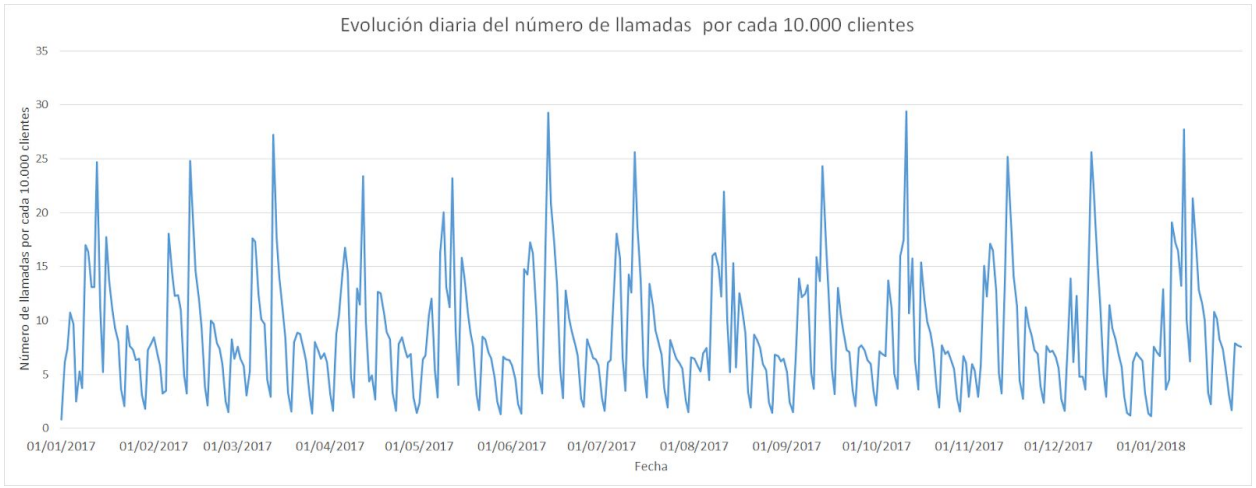

Figura 3. Evolución diaria de las llamadas por motivos de facturación

Destacan las caídas producidas en los fines de semana y los festivos a lo largo del año.

Metodología y desarrollo

Modelo de datos

Para el desarrollo del modelo se ha creado un datamart analítico cuyo eje principal son los ciclos de facturación. Cada cliente, en este caso identificado por su identificación (DNI, Tarjeta de Residente, etc.) y su msisdn, es considerado en cada ciclo de facturación una observación independiente.

Cada ciclo de facturación es identificado por un número de 8 cifras que indica:

- El año en el que se inicia.
- El mes en el que se inicia.
- El día del mes en el que se inicia.

Por ejemplo, el ciclo 20180215 es el que comienza el día 15 de Febrero de 2018 finalizando el 15 de Marzo. Se muestran a continuación los ciclos que se inician en el mes de enero y finalizan en el mes de febrero.

Figura 4. Ciclos de facturación iniciados en el mes de enero

Centrando el análisis en un cliente, para cada ciclo de facturación se contará con la información del ciclo actual (c) y de los n ciclos anteriores (c - 1 a c - n). Empleando la información conocida del cliente, se realizará la predicción de lo que sucederá en el siguiente ciclo de facturación.

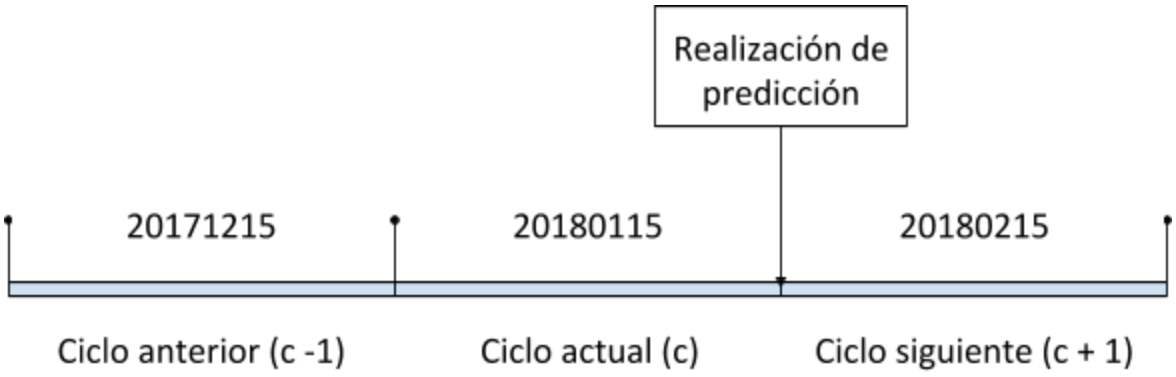

Figura 5. Ventanas temporales generadas durante la construcción del datamart

En el datamart, para cada cliente y ciclo de facturación se ha incorporado hasta el momento la siguiente información:

- Información de interacciones:
  - Llamadas por motivos de facturación en el ciclo actual.
  - Llamadas por motivos de facturación en el ciclo anterior.
  - Llamadas por motivo de cancelaciones en el ciclo actual.
  - Llamada por motivo de churn en el ciclo anterior.
  - Llamadas relacionadas con las tarifas y planes de precios.
  - Llamadas por incidencias con el servicio de adsl.
  - Llamadas por incidencias con el servicio móvil.
  - Llamadas por mejora del terminal.
  - Llamadas por reparación y envío de terminales.
  - Llamadas por nuevas altas.
  - Llamadas por gestión de productos y servicios.
- Información de planes de precios:
  - Plan de datos en el ciclo de facturación actual.
  - Cambio de plan de datos con respecto al ciclo anterior.
  - Plan de voz en el ciclo de facturación actual.
  - Cambio de plan de voz con respecto al ciclo anterior.
- Información de promociones:
  - Promociones aplicadas en el ciclo de facturación actual.
  - Meses restantes para finalizar la promoción.
- Información sobre productos y servicios:
  - Número de líneas de prepago.
  - Número de líneas de postpago.
  - Número total de líneas.
- Información del entorno social:

  - Código postal.
  - Región.
  - Género.
  - Tipo de documento de identidad.
  - Nacionalidad.
  - Edad.
- Variable objetivo a predecir:
  - Si el cliente realizará una o más llamadas por motivos de facturación en el siguiente ciclo de facturación.

Nota importante: actualmente no todas estas variables se han incorporado en el modelo que aquí se presenta. Las variables incorporadas en el datamart analítico han de servir de base para la realización de modelos más complejos y potentes, así como a la creación de otros modelos para la predicción de otras interacciones diferentes.

Un ejemplo de cómo se realiza la transformación desde el log de llamadas al datamart analítico organizado por ventanas según el ciclo de facturación se muestra a continuación para un único cliente:

msisdn	ident	call_date
621203977	23864123R	2018-01-30
621203977	23864123R	2018-01-25
621203977	23864123R	2018-01-05
621203977	23864123R	2017-12-27
621203977	23864123R	2017-12-24
621203977	23864123R	2017-12-05

Tabla 6. Log de llamadas

msisdn	ident	billing_cycle_id	n_calls_c -1	n_calls_c	n_calls_c+1
621203977	23864123R	20180122	3	2	?
621203977	23864123R	20171222	1	3	2
621203977	23864123R	20171122	?	1	3

Tabla 7. Datamart Analítico

Análisis descriptivo univariante

Una vez definidas las variables de interés se procede a realizar el análisis descriptivo univariante. Debido a la extensión del mismo se muestran aquí las variables con más capacidad de ordenar la variable objetivo de predicción. El resto del análisis univariante se incluirá en un anexo de este mismo documento.

Este análisis tiene las siguientes características:

- Se ha realizado sobre la muestra balanceada de la población, por lo que la probabilidad media de realizar una llamada por motivos de facturación es 0.5 (la mitad de la población ha realizado una o más llamadas en el siguiente ciclo de facturación, mientras que la otra mitad no ha realizado ninguna llamada).
- Se ha incluido los ciclos de facturación comprendidos entre el primer ciclo iniciado en Enero de 2017 (20170101) y el último ciclo iniciado en Noviembre de 2017 (20171122).
- En el eje x del gráfico se muestran los diferentes valores que puede tomar la variable (feature).
    - En el caso de ser una variable puramente numérica (como la edad) se muestran los diferentes buckets o intervalos en los que se ha dividido la misma.
    - En el caso de ser una variable categórica se muestran las diferentes categorías. Si el número de categorías es muy elevado, se mostrarán únicamente aquellas que agrupen a una cantidad de población suficiente.
    - En el caso de las variables numéricas cuya población desciende bruscamente a partir de cierto valor, se ha realizado un truncado (clipping) de los valores incluyendo todos los restantes en las categorías primera y última.
- En el eje y izquierdo se muestra el número de observaciones que se incluyen en cada categoría del eje x.
- En el eje y derecho se muestra la probabilidad media de realizar una llamada por motivos de facturación de cada categoría. Para mejorar la visualización, este eje no comienza en cero, sino en el valor mínimo observado para dicha categoría. Esto deberá ser tenido en cuenta a la hora de interpretar los gráficos.

Productos y servicios contratados

Respecto al número de líneas que el cliente tiene contratadas se observa un efecto contrapuesto entre los casos de postpago y prepago. La contribución del número de líneas de postpago es reducida.

- Un mayor número de líneas de prepago contratadas aumenta la probabilidad de que un cliente realice una llamada por motivos de facturación:

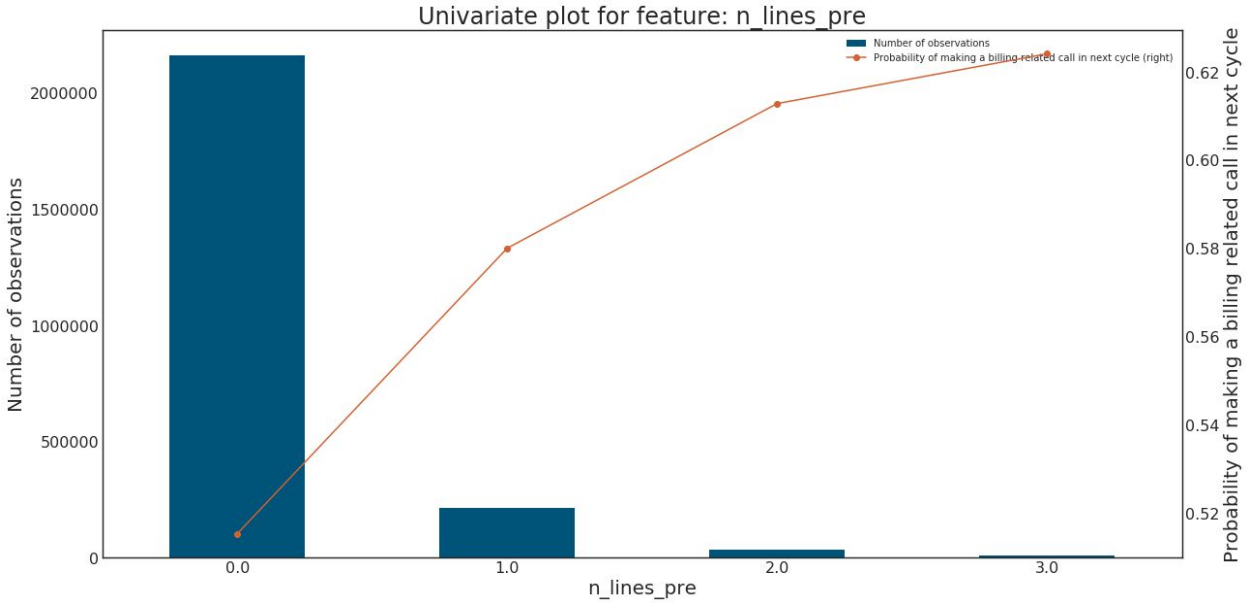

Figura 8. Gráfico univariante. Número de líneas de prepago

- Un mayor número de líneas de postpago no tiene un efecto claro sobre la probabilidad de realizar llamadas. La variación es reducida (de 0.49 a 0.54) y no se observa monotonicidad:

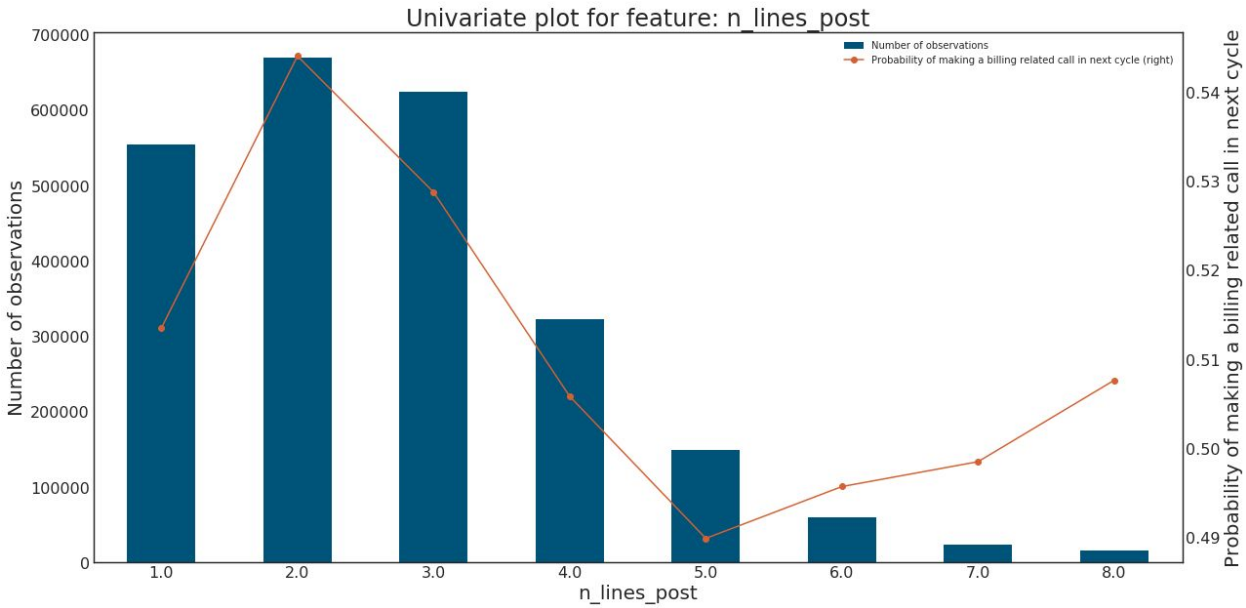

Figura 9. Gráfico univariante. Número de líneas de postpago

Códigos de promoción

En el caso de las promociones, estas se ofrecen con una duración de 24, 18, 12 o 6 meses. Se incluyen en este modelo inicial tanto los códigos concretos de promoción aplicados como el número de meses restantes para la finalización de las promociones.

- Se observan aumentos de la probabilidad para clientes con número de meses para finalizar la promoción cercanos a los meses de duración de las promociones. Esto puede ser provocado por clientes que acaban de incorporarse a la compañía y están recibiendo sus primeras facturas. El efecto de esta variable, que se mueve entre 0.6 y 0.8, es importante:

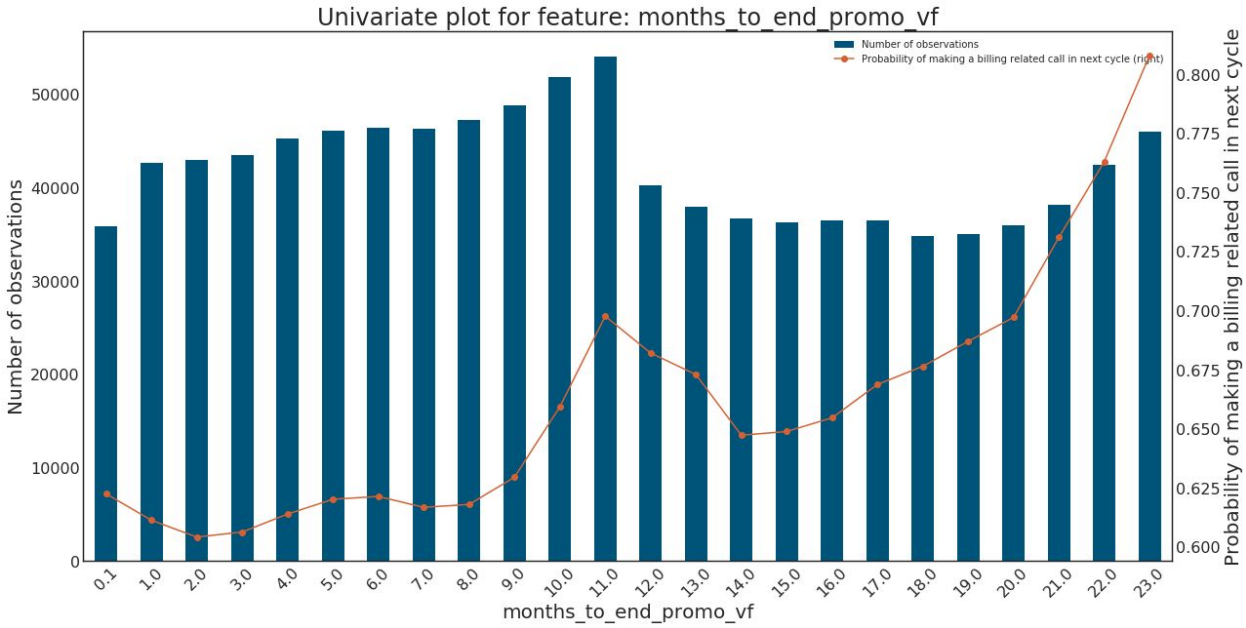

Figura 10. Gráfico univariante. Meses para finalizar promoción de tipo Vodafone

- En el caso de códigos concretos de promoción aplicados, se observa que hay ciertos códigos que tienen asociada una mayor o menor probabilidad de llamada, siendo este efecto también importante, en este caso variando desde prácticamente 0 a 0.8:

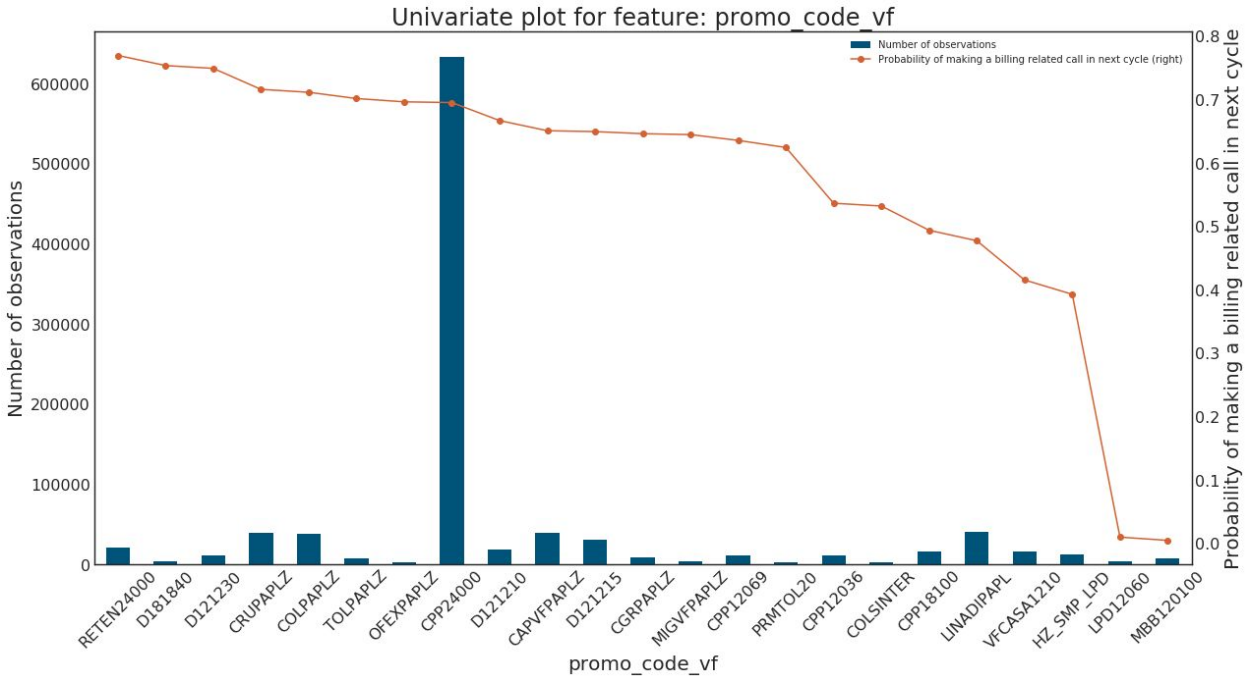

Figura 11. Gráfico univariante. Código de promoción de tipo Vodafone

Planes de precios

En el caso de los planes de voz, se observa un efecto tanto a nivel de plan concreto, como a nivel de cambio de estado.

- En el caso de un cambio de plan de voz o de datos en un ciclo de facturación, se observa la existencia de un aumento importante (de un 60%) de la probabilidad de realizar una llamada por parte del cliente, como sería de esperar. Un cambio de este tipo tiene asociada una incertidumbre inicial del cliente respecto a la factura y la posibilidad de desacuerdos con la misma al principio:

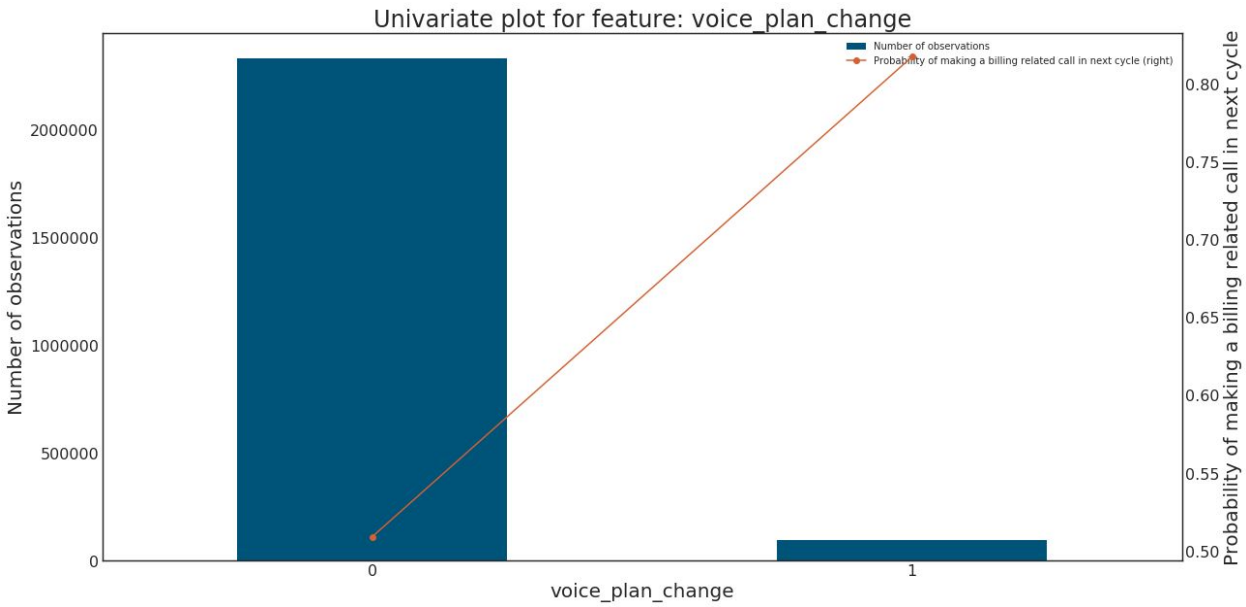

Figura 12. Gráfico univariante. Cambio en el plan de voz

- En el caso de planes concretos de voz y datos, se observa que ciertos planes estás asociados con una mayor probabilidad de llamada:

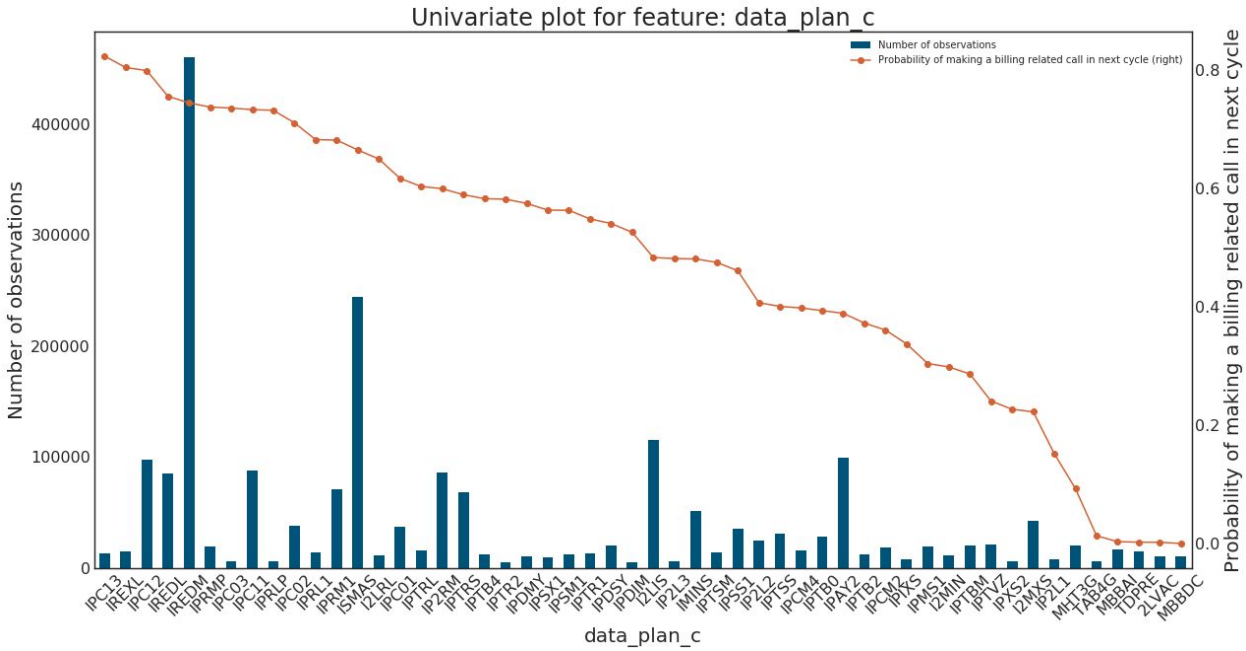

Figura 13. Gráfico univariante. Plan de voz

Llamadas realizadas en el ciclo actual y en ciclos anteriores

Para las llamadas realizadas en el ciclo actual y en los ciclos anteriores, se observa siempre el mismo efecto. Para aquellos clientes que realizan una o más llamadas, la probabilidad de realizar llamadas en el siguiente ciclo es mucho más alta que la media (un 80% más alta) llegando a una probabilidad cercana a 1.0 para clientes con varias llamadas realizadas. Es muy importante destacar que todas las llamadas independientemente de la tipología de la misma ofrecen un comportamiento muy similar.

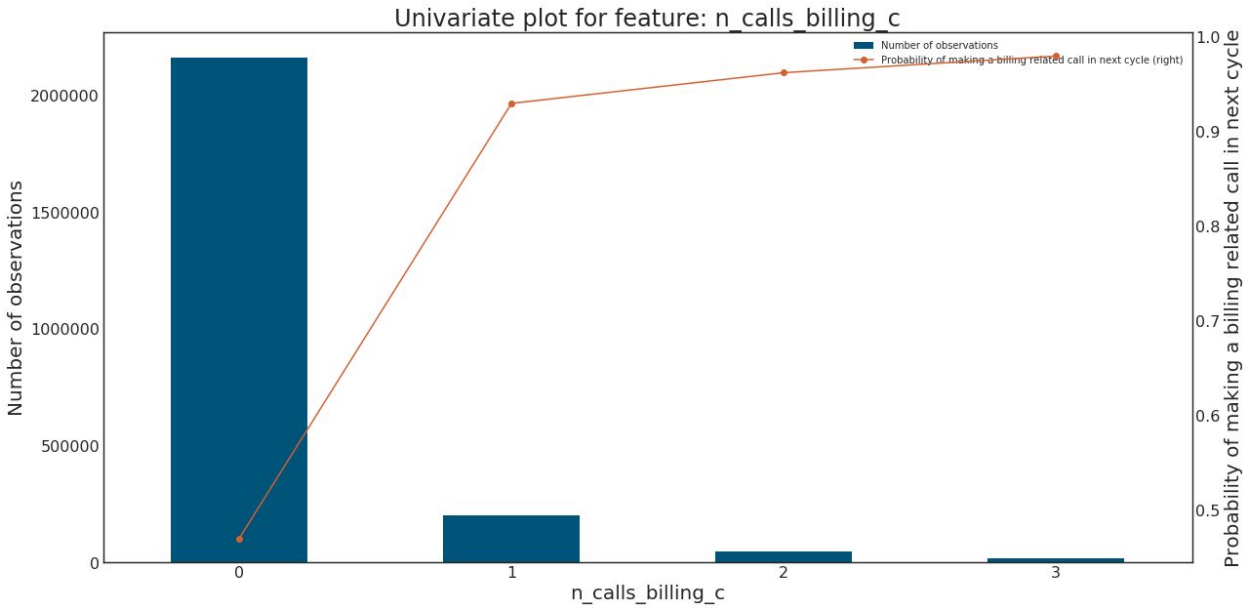

Figura 14. Gráfico univariante. Plan de voz

Entorno social

Respecto al entorno social, se han evaluado las contribuciones de la edad, la nacionalidad y el código postal del lugar de residencia del cliente.

- En el caso del código postal se observan diferencias importantes en la probabilidad de llamada en función del código postal concreto. Se observan variaciones de hasta el 30% en la probabilidad según el código postal para aquellos códigos postales con mayor número de muestras:

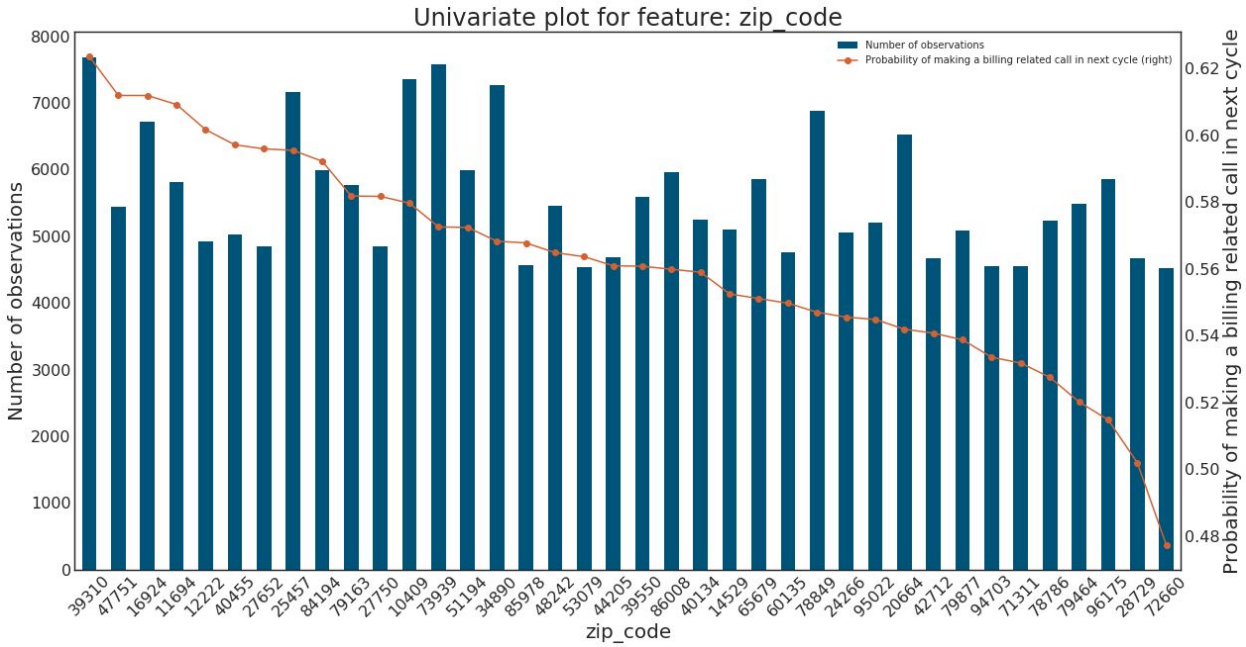

Figura 15. Gráfico univariante. Código postal

- En el caso de la nacionalidad del cliente también se observan diferencias importantes. Aquellas nacionalidades tradicionalmente asociadas a un menor poder adquisitivo (de América Latina y África) tienen probabilidades de realizar llamadas superiores (hasta el doble) a las nacionalidades de Europa Occidental:

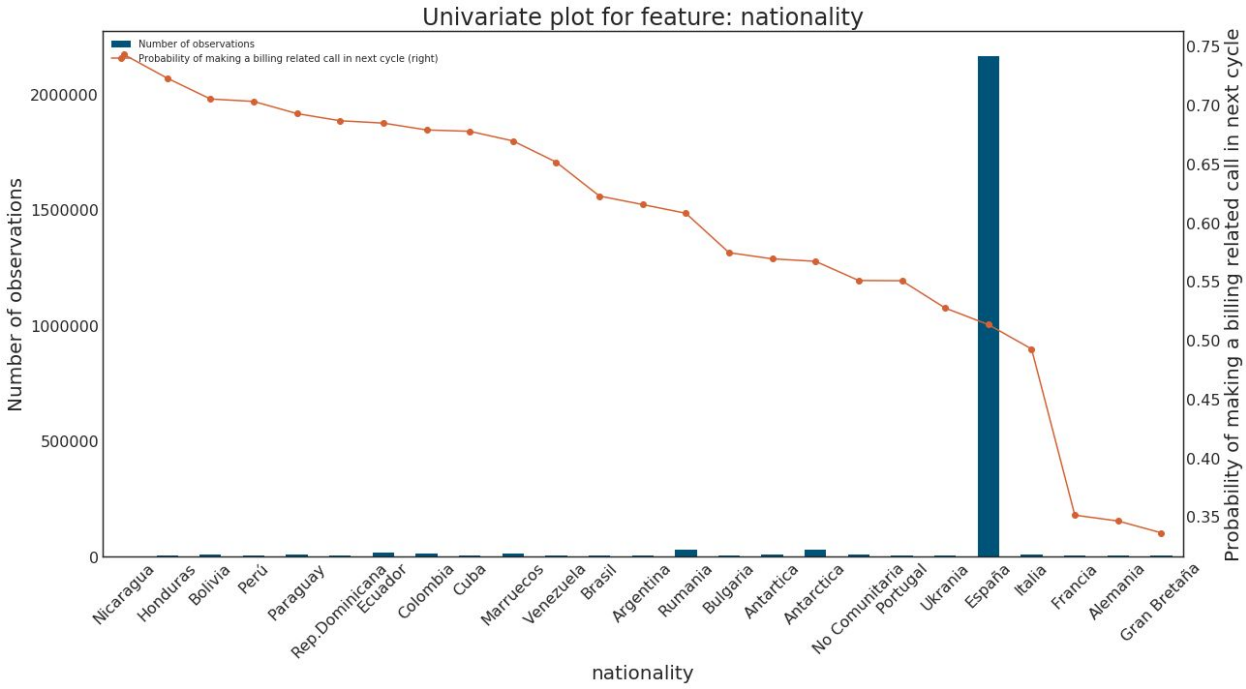

Figura 16. Gráfico univariante. Nacionalidad

- En el caso de la edad se observa un efecto claro. Aquellos clientes con más edad tienden a realizar menos llamadas. También aquellos clientes que no pasan los 13 años tienen una probabilidad de llamada reducida. Los clientes en torno a los 20 años tienden a tener una probabilidad de más del doble que la de aquellos mayores de 80:

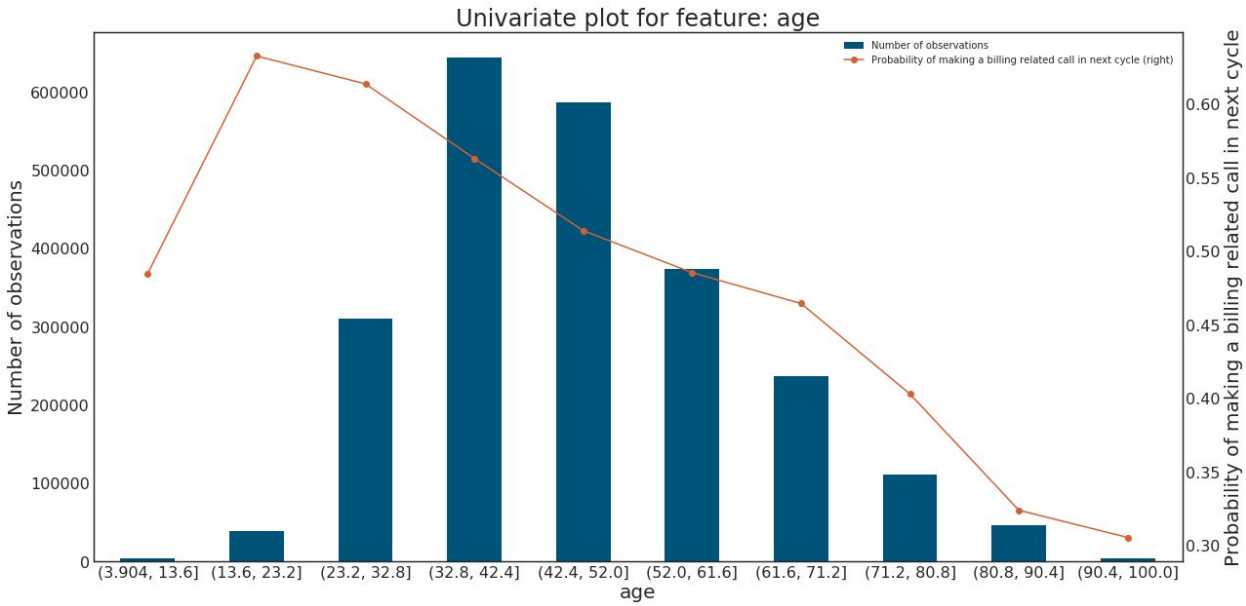

Figura 17. Gráfico univariante. Edad

Modelo de predicción de llamadas

Se ha optado por una regresión logística como algoritmo. Dada la similitud de la casuística de este proyecto con los modelos de predicción de fugas, donde el uso de la regresión logística está muy extendido, es de esperar un comportamiento aceptable de la misma

El modelo se ha realizado utilizando:

- scikit-learn para el prototipado del modelo (obtención de parámetros óptimos y cálculo de métricas y curvas de rendimiento).
- Spark ML para su implementación final pensando en una futura puesta en producción.

No se descarta que en futuras etapas del proyecto se amplíe el catálogo de librerías utilizadas.

Preparación de los datos

La información almacenada en el datamart analítico se ha procesado previamente pasando por las siguientes etapas de preprocesamiento:

- Se eliminan los valores nulos de las features categóricas sustituyendolas por una categoría que agrupa todos los nulos.
- Se crean unos diccionarios de valores 'vistos' para las variables categóricas. En el testeo, cualquier valor nuevo que no haya estado presente durante el entrenamiento del modelo, será omitido como categoría para poder así realizar el scoring de cualquier nueva observación
- Se sustituyen los valores nulos de las features numéricas por su media.
- Se sustituyen algunos valores nulos de ciertas features por cero (número de llamadas).

Se ha realizado también un proceso de feature engineering y selección basado en:

- Eliminar variables poco representativas del modelo lineal basándose en los valores del estadístico p.
- Eliminar variables muy correladas entre sí.
- Reducción de categorías de ciertas variables cuando existía mucha diferencia de población entre las diferentes categorías, agrupando varias categorías en otras nuevas.
- Eliminación de categorías en las que el número de categorías es muy grande.

Las variables resultantes de este proceso son las siguientes:

- Llamadas por motivos de facturación.
- Llamadas por motivo de cancelaciones.
- Llamada por motivo de churn.
- Llamadas relacionadas con las tarifas y planes de precios.
- Llamadas por incidencias con el servicio de adsl.
- Llamadas por incidencias con el servicio móvil.
- Llamadas por mejora del terminal.
- Llamadas por reparación y envío de terminales.

- Llamadas por nuevas altas.
- Llamadas por gestión de productos y servicios.
- Ciclo de facturación.
- Género.
- Nacionalidad (Español / Extranjero).
- Número de líneas de postpago.
- Número de líneas de prepago.
- Edad.
- Meses para la finalización de código de promoción de tipo Vodafone.
- Cambio en el plan de voz.

Creación de conjuntos de entrenamiento y de testeo

Se ha seguido la siguiente estrategia para la división del total de las observaciones en diferentes conjuntos para seleccionar el mejor modelo, entrenarlo y probarlo: tomar como datos de entrenamiento aquellas observaciones anteriores a cierto momento y como datos de prueba las observaciones posteriores al mismo. Esto permite evaluar el poder predictivo del modelo a futuro, es decir, en condiciones de utilización real. Cabe esperar que el poder predictivo del modelo se diluya a medida que los datos se alejan en el tiempo del periodo de entrenamiento. En este proyecto al trabajar con ciclos de facturación, se han aislado un número de ciclos de facturación correspondientes a dos meses de datos.

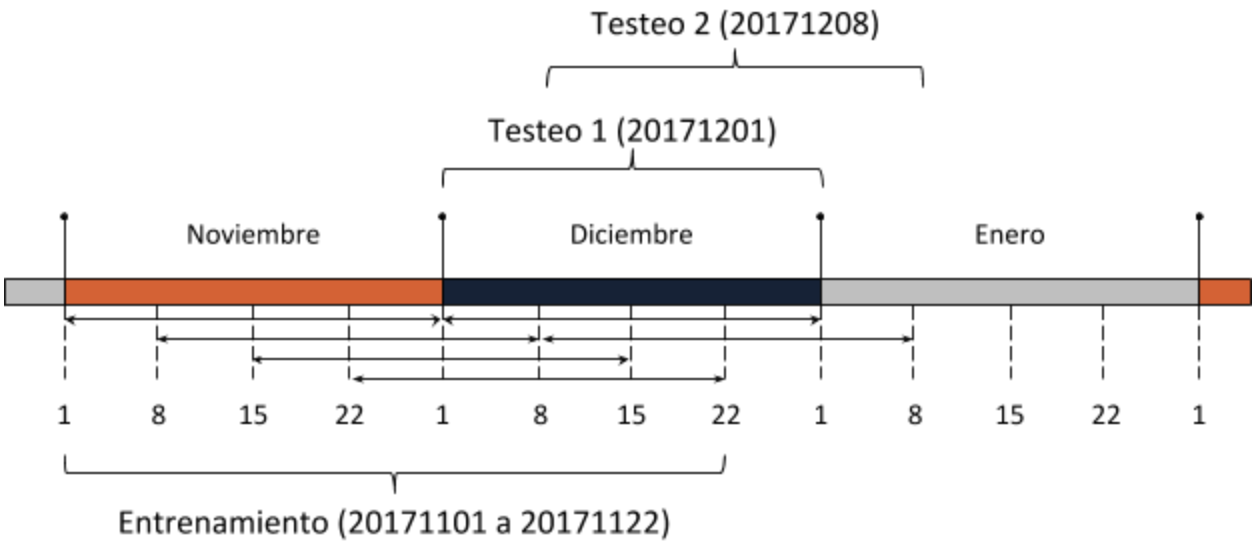

Figura 18. División de las observaciones en conjuntos de entrenamiento y testeo

Fruto de la aplicación de estas estrategias, se obtienen tres conjuntos de datos:

- Conjunto de entrenamiento: los datos empleados para el entrenamiento del modelo. Correspondientes a los datos de los ciclos de facturación desde el 1 de Noviembre de 2017 hasta el 30 de Noviembre de 2017. Se han realizado pruebas utilizando tanto balanceo de la clase minoritaria como utilizando la proporción real. Al tratarse de un modelo lineal, los mejores resultados se han obtenido utilizando las proporciones reales.
- Conjunto de validación: se ha empleado el sistema de validación cruzada utilizando muestreo estratificado con 10 segmentos, preservando así la proporción de cada clase. Se realiza por tanto el entrenamiento 10 veces, tomando como set de validación cada vez un 10% distinto del conjunto de entrenamiento.
- Conjuntos de testeo: datos empleados para medir el potencial predictivo del modelo. Siguiendo la estrategia descrita, se van a definir dos conjuntos de testeo de la siguiente forma:
    - Conjunto de testeo 1: todas las observaciones (reales, sin balancear) correspondientes al ciclo de facturación 20171201.
    - Conjunto de testeo 2: todas las observaciones (reales, sin balancear) correspondientes al ciclo de facturación 20171208.

Parámetros del modelo

Los parámetros del modelo de regresión lineal son los siguientes:

- Parámetro de regularización: es un parámetro que penaliza altos valores de los coeficientes en la regresión lineal, evitando así que se produzca sobreajuste bajo ciertas condiciones.
- Ordenada en el origen: si incluir u omitir el parámetro de ordenada en el origen (intercept).

Entrenamiento del modelo

Metodología de entrenamiento

Para el entrenamiento del modelo se realiza una búsqueda sistemática (no aleatoria) a través de una grid de parámetros, probando en el proceso todas las combinaciones posibles:

- Parámetro de regularización: se prueban valores que varían de 10^{-4} a 10^{4} distribuidos logarítmicamente en 10 muestras.
- Ordenada en el origen: se prueba tanto a incorporar ordenada en el origen como a omitirla.

Esto supone realizar el entrenamiento del modelo 20 veces (10 * 2) para realizar la selección del modelo con los mejores parámetros posibles.

La métrica objetivo para realizar la selección del modelo es el área bajo la curva roc (AUC). Se selecciona por tanto el modelo con mejor AUC.

Resultado del entrenamiento y mejor modelo obtenido

El mejor modelo obtenido tras el entrenamiento es el siguiente:

- Parámetro de regularización C: 0.0001
- Ordenada en el origen: Con ordenada en el origen
- AUC en conjunto de entrenamiento: 0.78

Variables más importantes e interpretabilidad

Las variables más importantes del modelo se muestran a continuación. Aquellas variables con valor positivo son las que hacen aumentar la probabilidad de que un cliente realice una llamada por motivos de facturación en el siguiente ciclo de facturación. Aquellas variables con valor negativo son aquellas que hacen disminuir la probabilidad de que un cliente realice una llamada en el siguiente ciclo de facturación.

Nombre coeficiente	valor	p_value	t_value
n_calls_billing_c	0.2178	~0	4389.18
n_calls_billing_c_minus_1	0.1637	~0	3318.82
months_to_end_promo_vf	0.0958	~0	2027.39
n_calls_churn_c	0.0902	~0	1830.92
voice_plan_change	0.0868	~0	1800.6
n_calls_ser_man_c_minus_1	0.0787	~0	1600.28
n_calls_dsl_inc_c	0.0768	~0	1566.36
n_calls_ser_man_c	0.0608	~0	1253.94
n_calls_churn_c_minus_1	0.0506	~0	1023.4
n_calls_dsl_inc_c_minus_1	0.0483	~0	972.743
n_calls_tariff_c	0.0461	~0	966.143
billing_cycle_id_01	0.0425	255.33	~0
n_calls_new_adds_c	0.0374	~0	768.5
n_calls_mobile_inc_c	0.0348	~0	724.873
n_calls_tariff_c_minus_1	0.0339	~0	711.312

n_lines_pre	0.033	~0	698.416
n_calls_new_adds_c_minus_1	0.0275	~0	556.496
n_calls_mobile_inc_c_minus_1	0.0266	~0	553.775
n_calls_device_upgr_c	0.0215	~0	437.206
nationality_Extranjero	0.0189	-	-
n_calls_device_del_rep_c	0.0159	~0	330.532
gender_Mujer	0.0154	-	-
n_calls_device_upgr_c_minus_1	0.012	~0	243.999
n_calls_device_del_rep_c_minus_1	0.0109	~0	226.404
billing_cycle_id_22	0.0048	232.393	~0
gender_Varón	-0.0154	-	-
nationality_Español	-0.0189	-	-
billing_cycle_id_15	-0.0255	226.851	~0
billing_cycle_id_08	-0.0263	234.774	~0
n_lines_post	-0.0671	~0	-1416.89
age	-0.155	~0	-3241.81

Tabla 19. Importancia de los coeficientes del modelo

Destacan positivamente aquellas variables relacionadas con el número de llamadas realizadas en el ciclo actual y el anterior, y negativamente principalmente la edad. También es destacable la contribución del ciclo de facturación del cliente, teniendo aquellos del cicl 1 una mayor probabilidad de llamada que aquellos de los ciclos 15 y 8.
Algunas variables como la nacionalidad o el género no tienen valores para los estadísticos p y t al ser binarias (el coeficiente de una contrarresta perfectamente el de su complementaria).

Desempeño del modelo en conjuntos de test

Se muestran a continuación una serie de métricas de desempeño del modelo.

Lift

Las curvas de lift para los conjuntos de train y test son muy similares y tienen el siguiente aspecto:

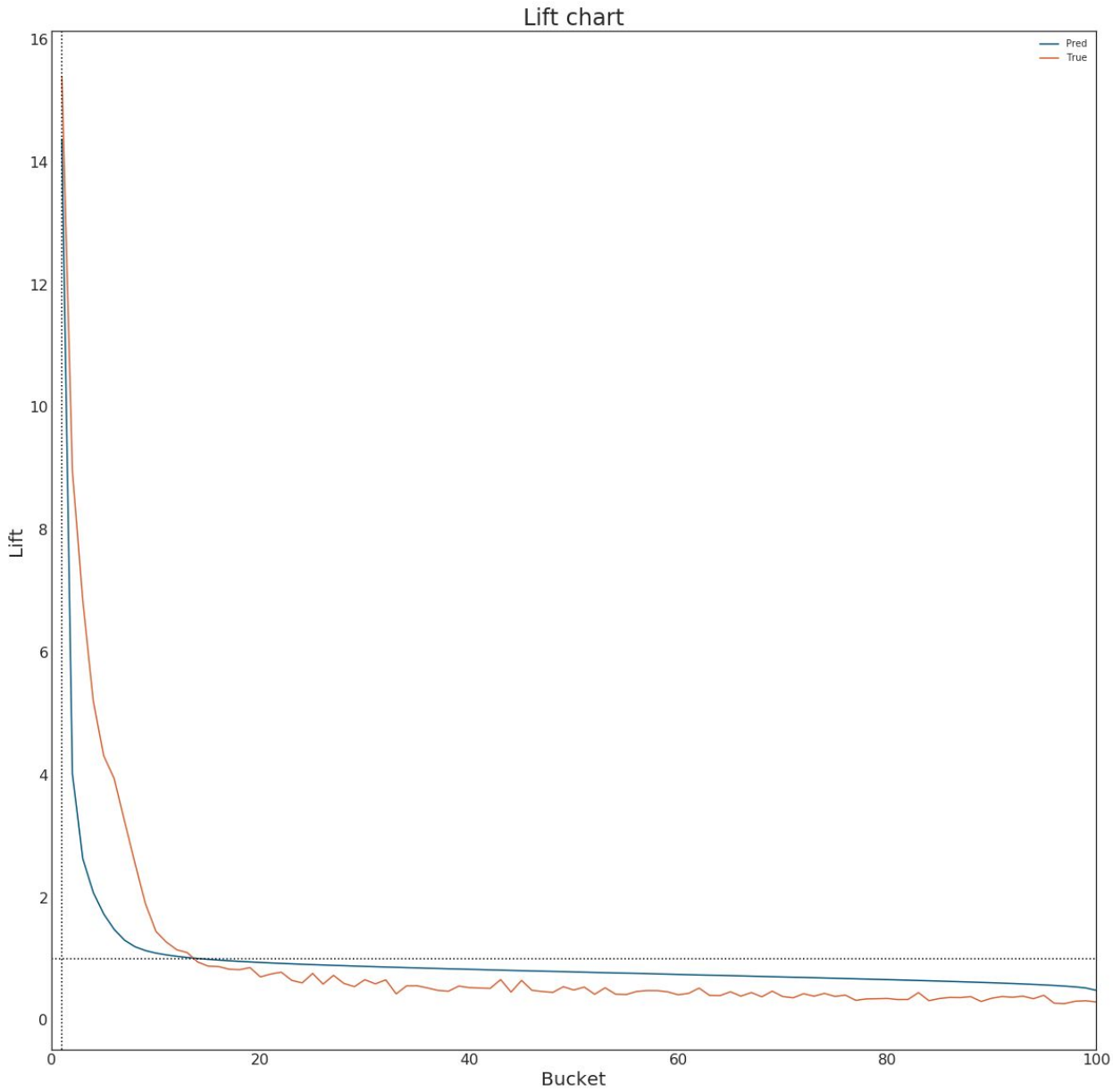

Figura 20. Gráfica de Lift. Conjunto de Test 2

Se aprecia en la gráfica que la capacidad predictiva del modelo es muy alta en los primeros percentiles dada la pendiente muy inclinada, reduciéndose mucho a partir del percentil 20. Esto indica que el modelo permite discriminar rápidamente a un alto % de las observaciones positivas con únicamente un 10% de la población, pero su poder predictivo se reduce una vez

descuenta el efecto de las variables más importantes (las llamadas en el ciclo actual y los anteriores).

Respuesta capturada

Las curvas de respuesta capturada para todos los conjuntos se muestran a continuación:

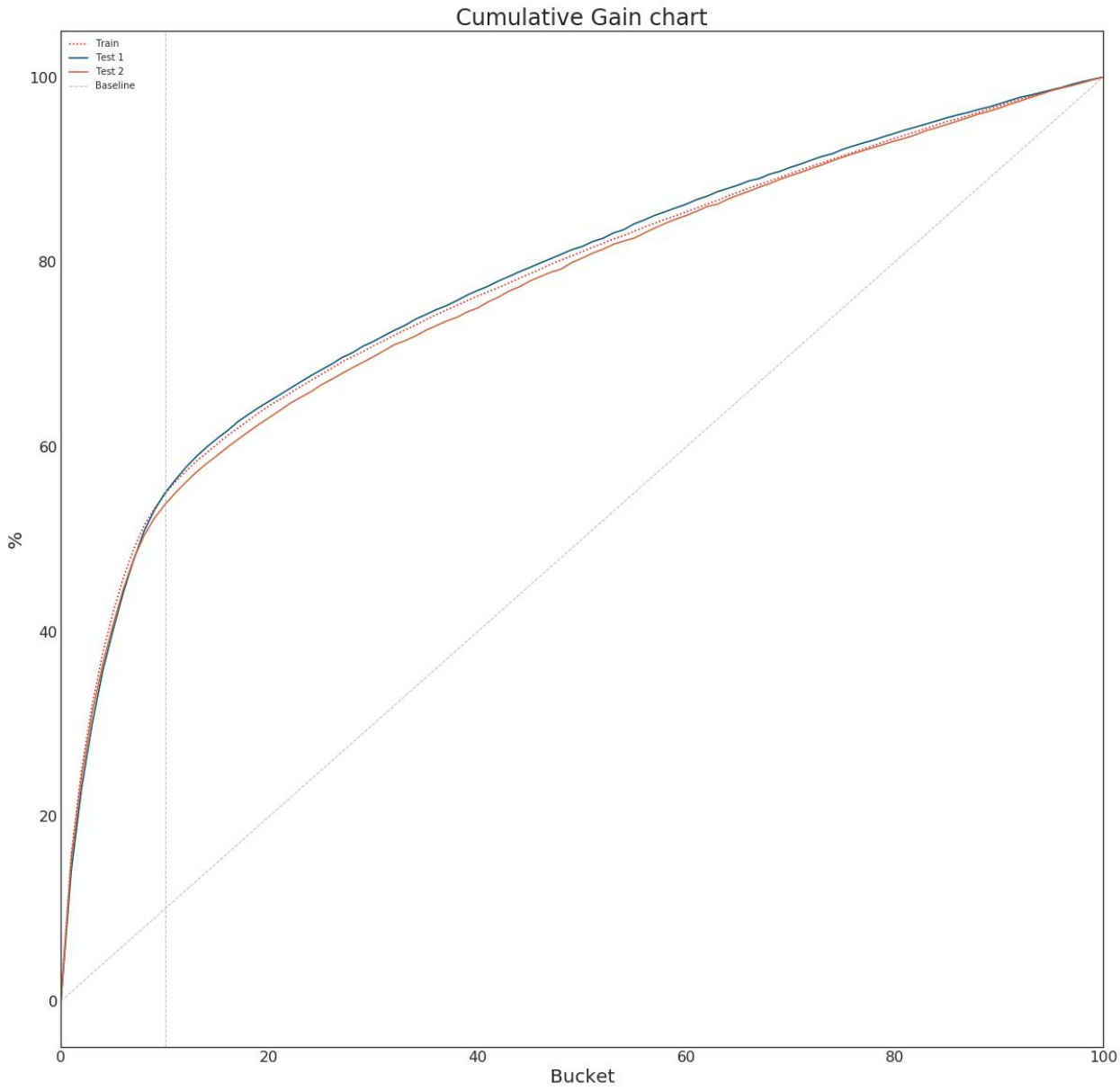

Figura 21: Gráfica de Ganancia acumulada. Conjunto de Test 2

De forma análoga al lift, se aprecia que la respuesta capturada es muy similar para los diferentes conjuntos. Se aprecia claramente que el modelo es capaz de separar algo menos del 60% de las observaciones positivas con únicamente un 10% de la población.

Curva ROC & AUC

Los valores de las AUC para los conjuntos de test 1 y dos son los siguientes:

- Conjunto de test 1: 0.78
- Conjunto de test 2: 0.77

Se muestra a continuación la curva ROC para el conjunto de test 2, siendo ésta muy similar a la del conjunto de entrenamiento y test 1.

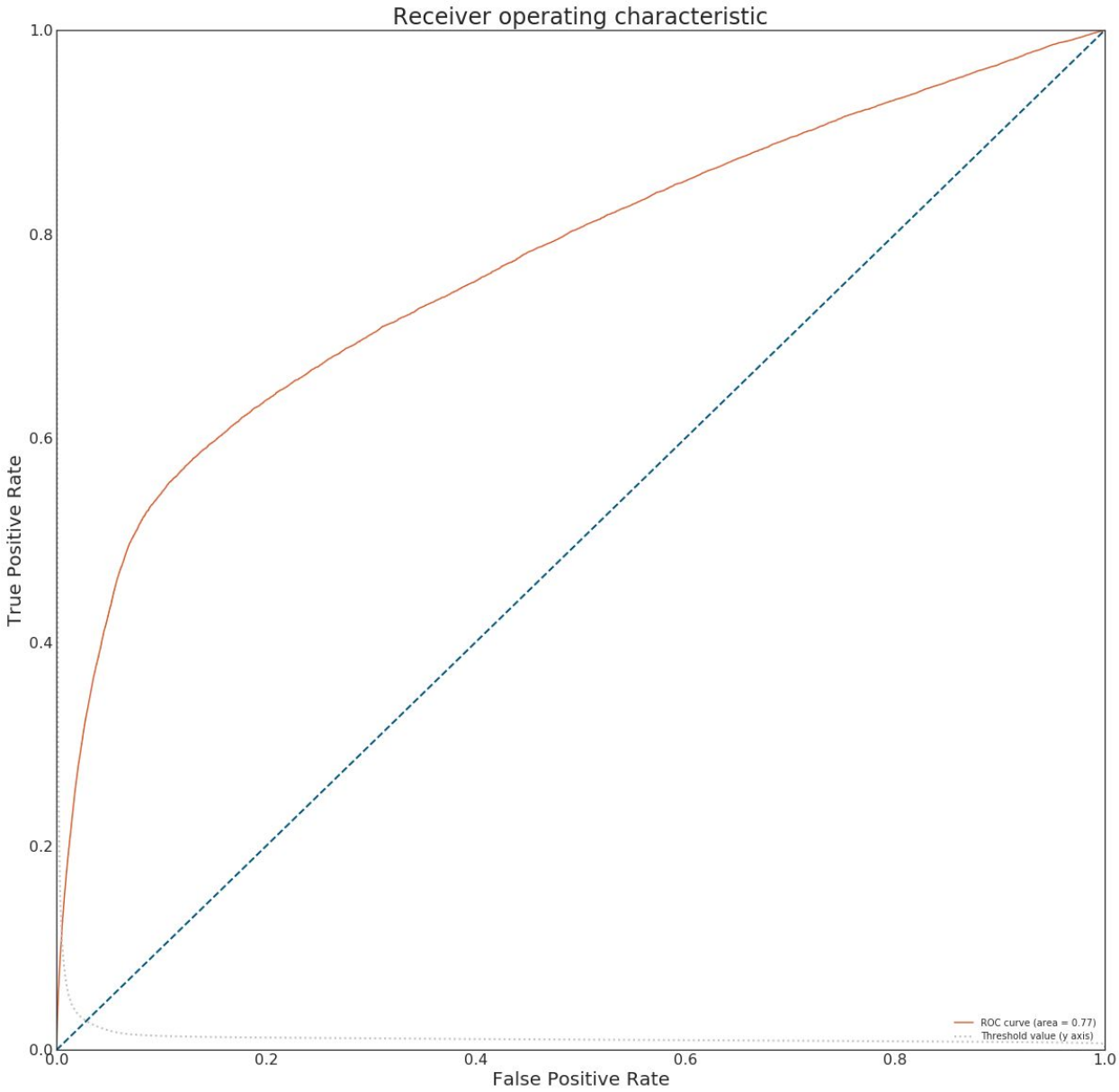

Figura 22: Curva ROC. Conjunto de Test 2

Otras métricas de rendimiento del modelo

Al ser un modelo cuyo output deseado es una lista de clientes ordenada por probabilidad de llamada, es de menor interés proporcionar métricas de clasificación que dependen de un umbral particular de decisión, como:

- Matriz de confusión.
- Precision.
- Recall.

- Etc.

Por interés, dichas métricas pueden consultarse en el anexo de este documento para un umbral estándar de 0.5.

Anexos

Anexo 1. Gráficos Univariantes de las variables del modelo

- Número total de líneas contratadas por el cliente (prepago + postpago):

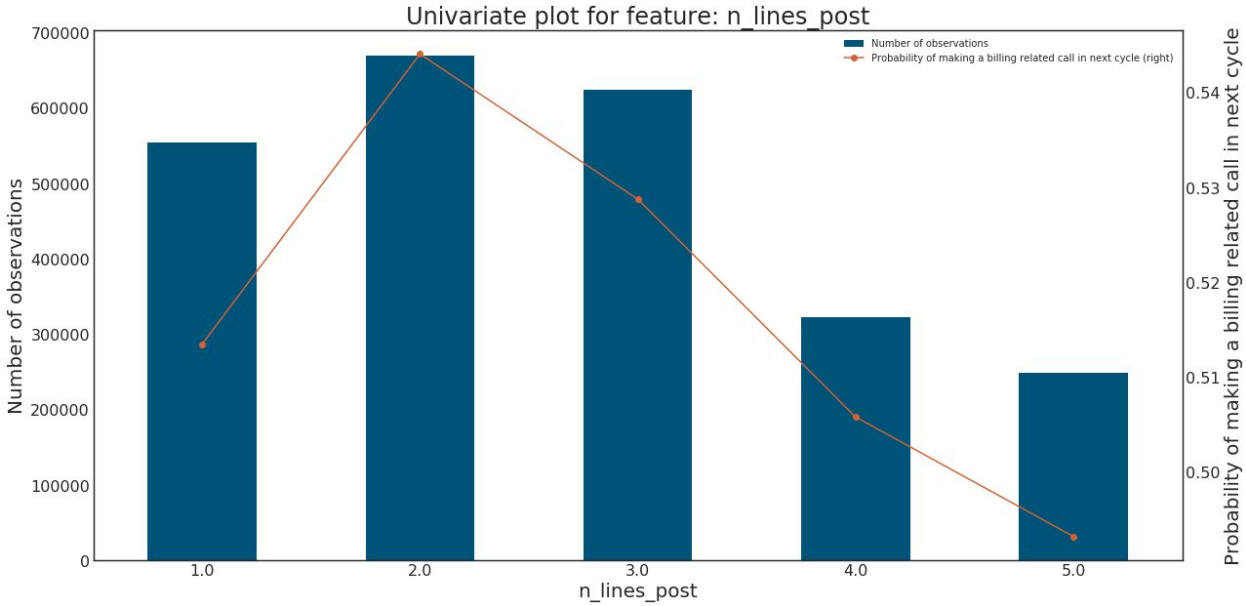

Figura 23. Gráfico univariante. Número total de líneas

- Número total de líneas de prepago:

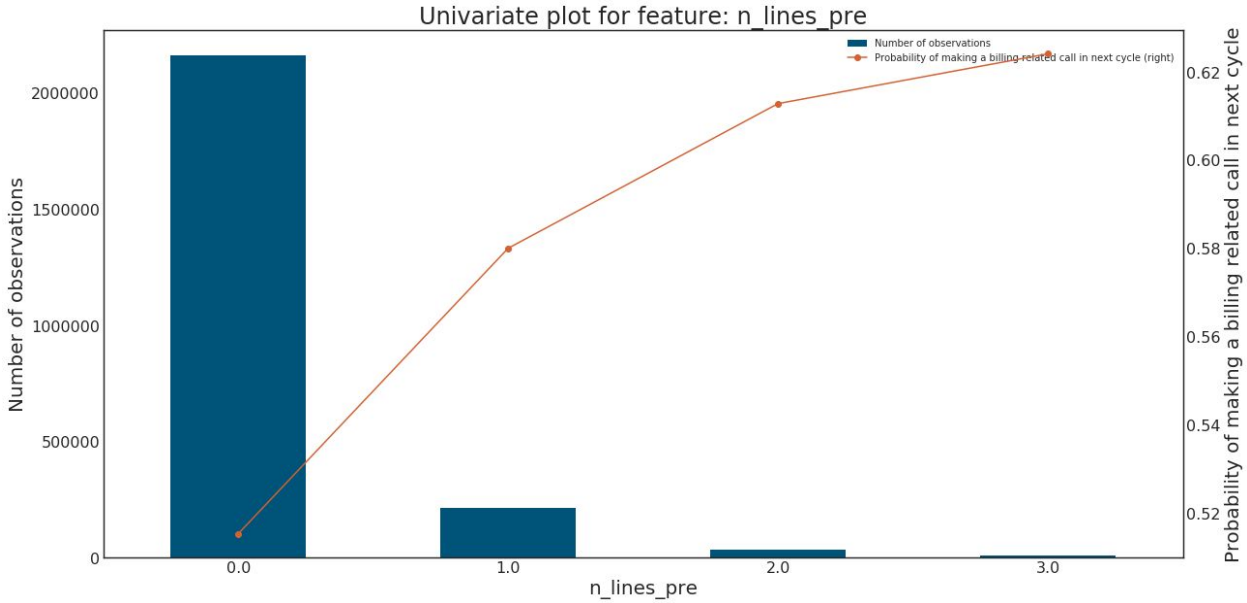

Figura 24. Gráfico univariante. Número total de líneas de prepago

- Número total de líneas de postpago:

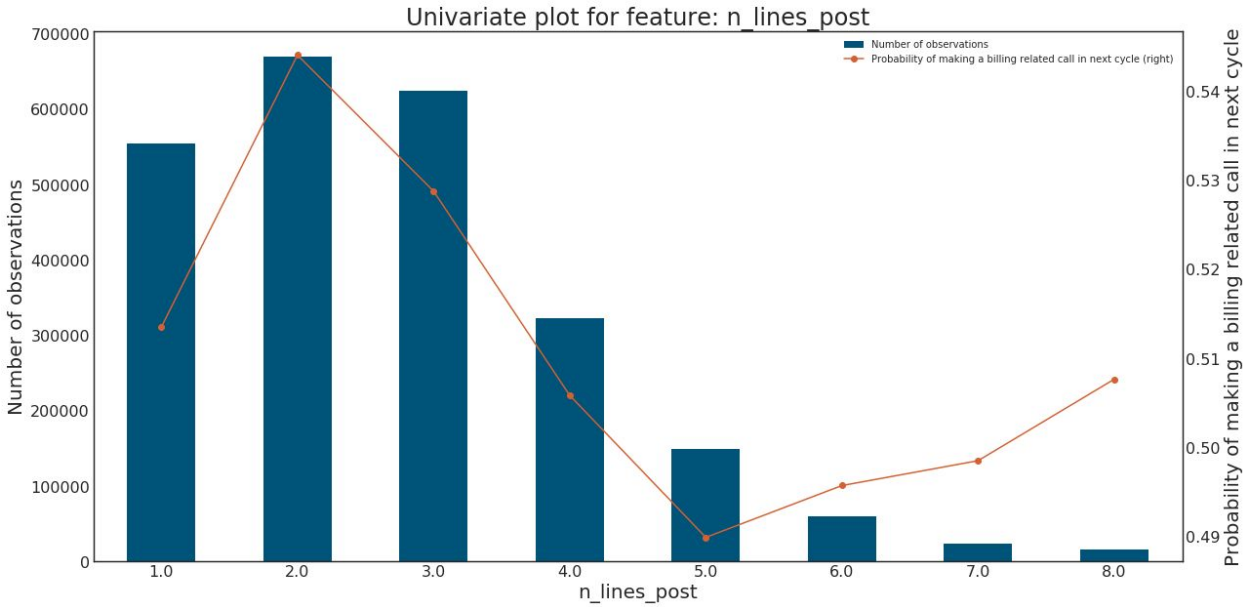

Figura 25. Gráfico univariante. Número total de líneas de postpago

- Número de meses para terminar promoción tipo tarifa:

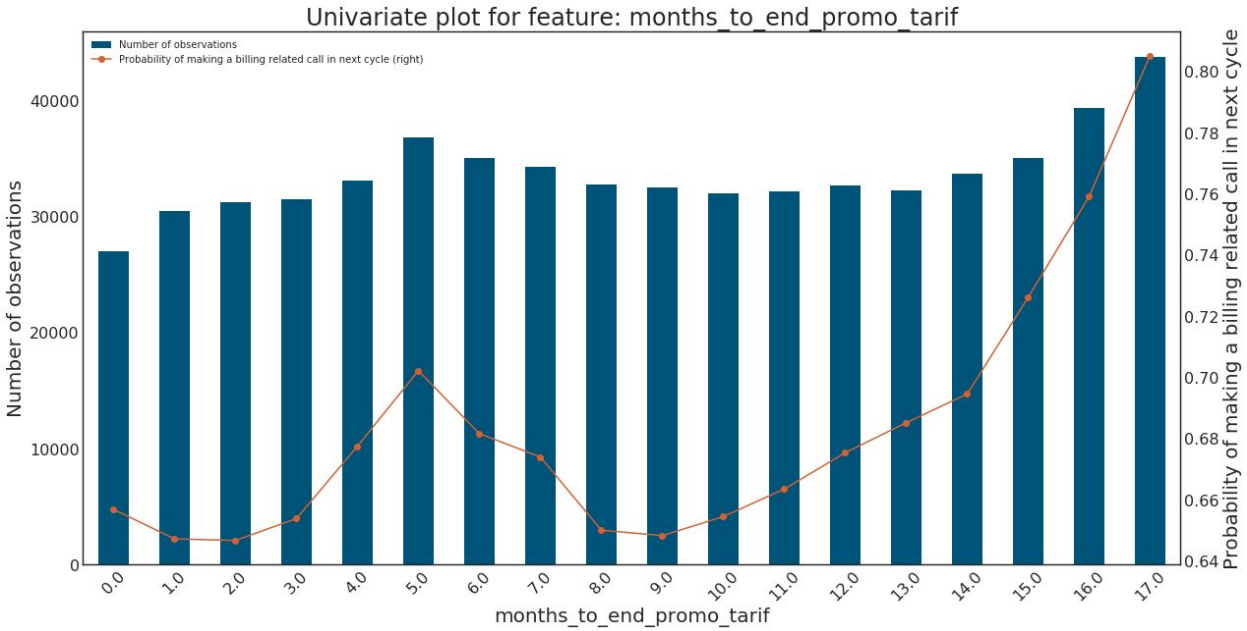

Figura 26. Gráfico univariante. Número de meses para finalizar promoción tipo tarifa

- Número de meses para terminar promoción tipo Vodafone:

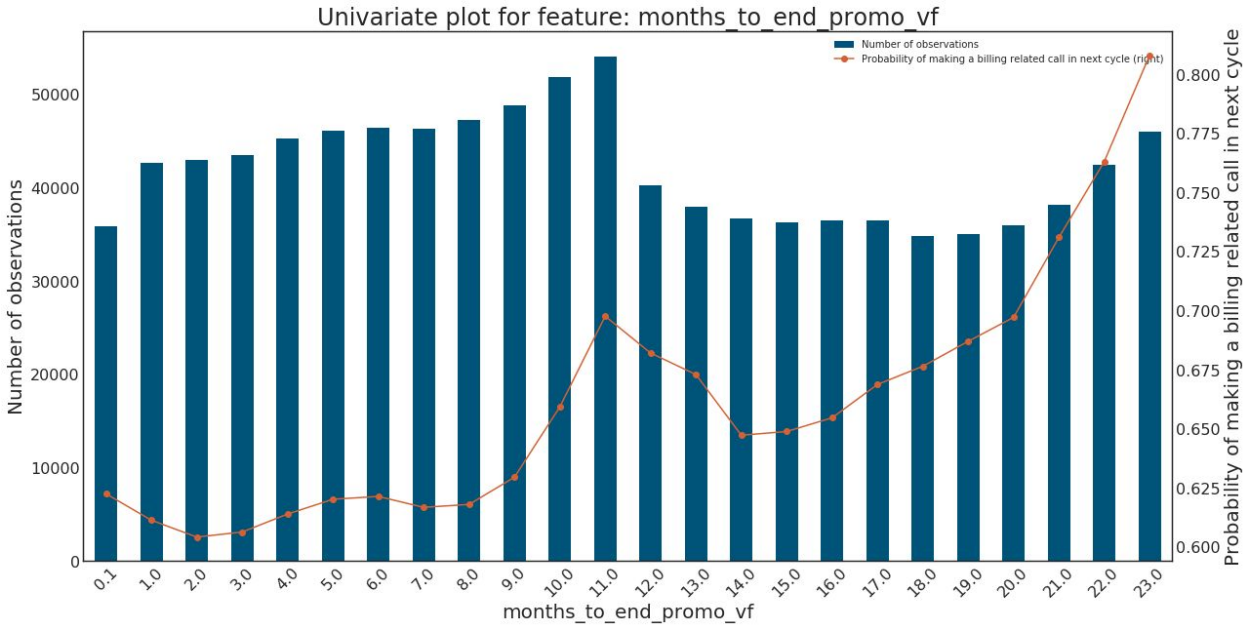

Figura 27. Gráfico univariante. Número de meses para finalizar promoción tipo Vodafone

- Cambio de tarifa de voz:

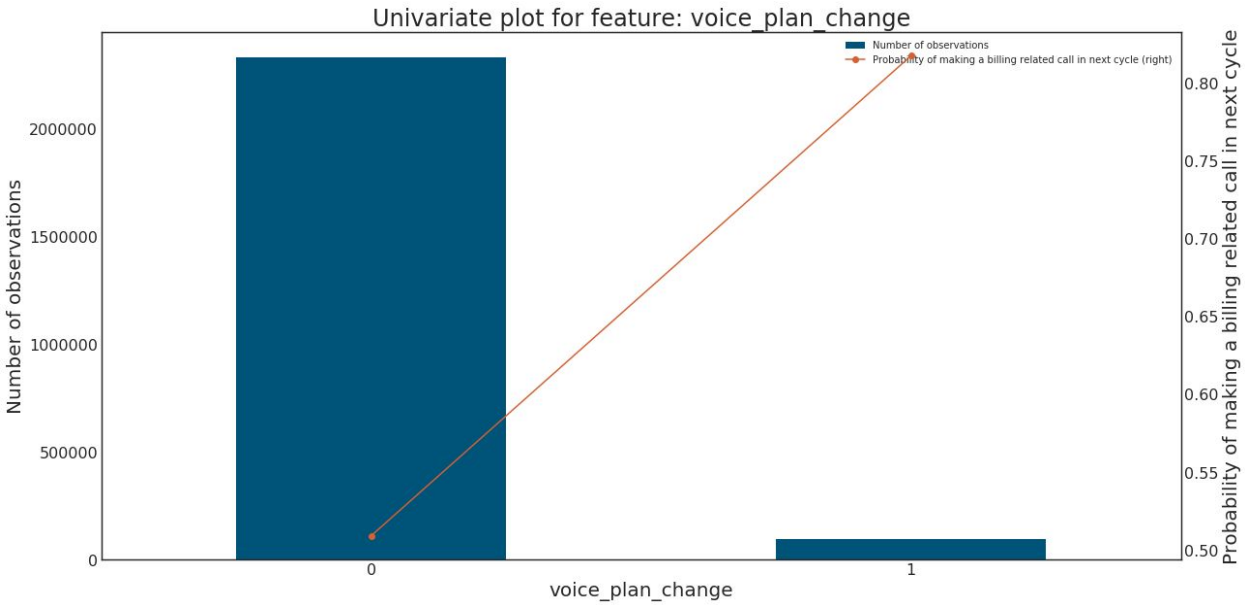

Figura 28. Gráfico univariante. Cambio de tarifa de voz

- Cambio de tarifa de datos:

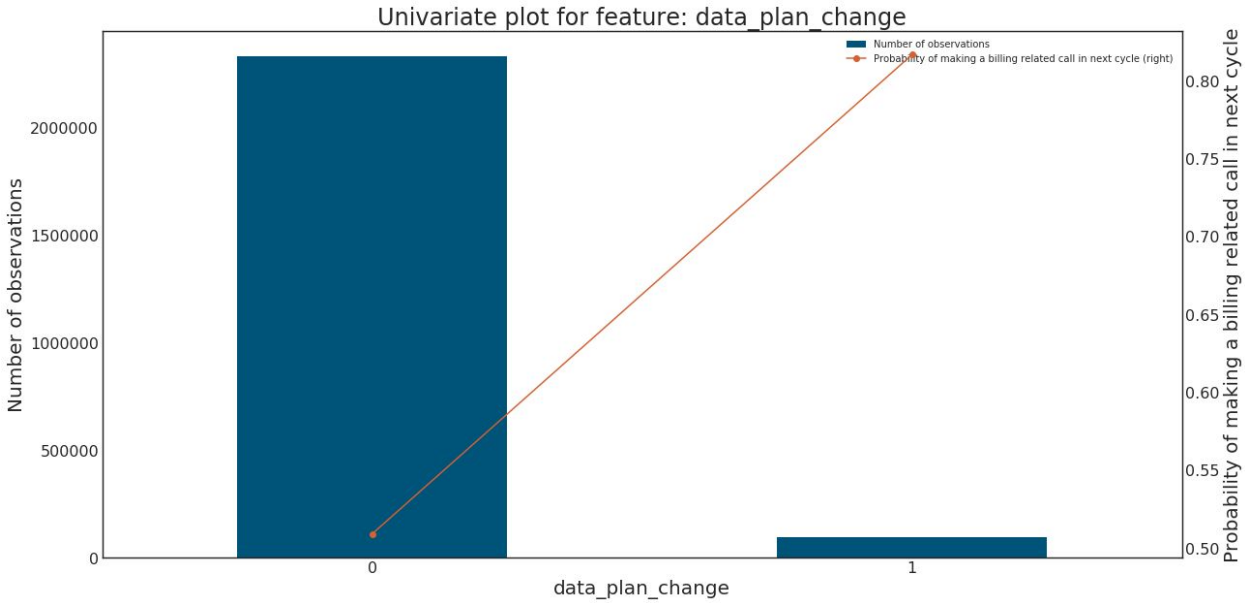

Figura 29. Gráfico univariante. Cambio de tarifa de datos

- Número de llamadas por motivos de facturación realizadas en el ciclo actual:

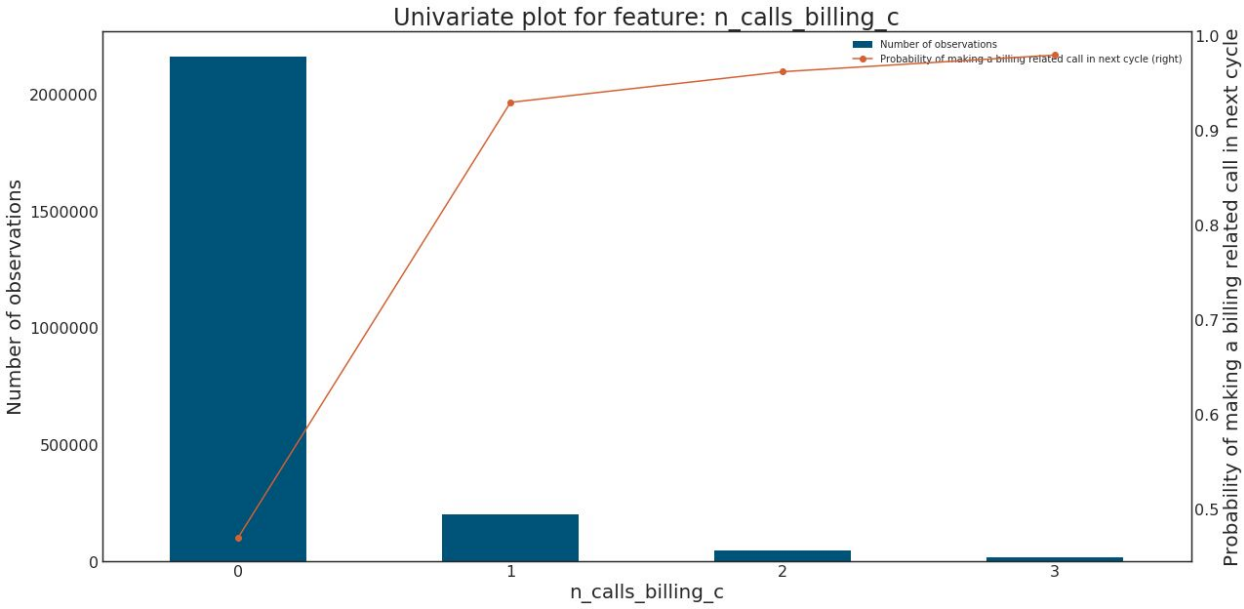

Figura 30. Gráfico univariante. Número de llamadas por motivos de facturación en ciclo actual

- Número de llamadas por motivos de facturación en ciclo anterior:

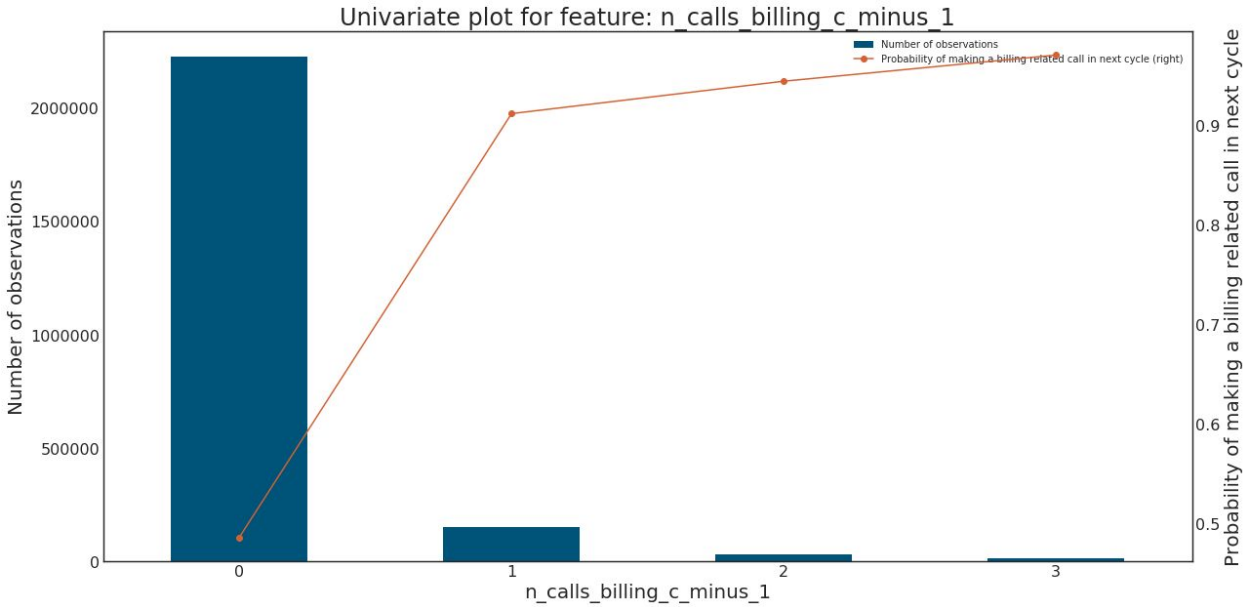

Figura 31. Gráfico univariante. Número de llamadas por motivos de facturación en el ciclo anterior

- Número de llamadas por motivos de fuga en el ciclo actual:

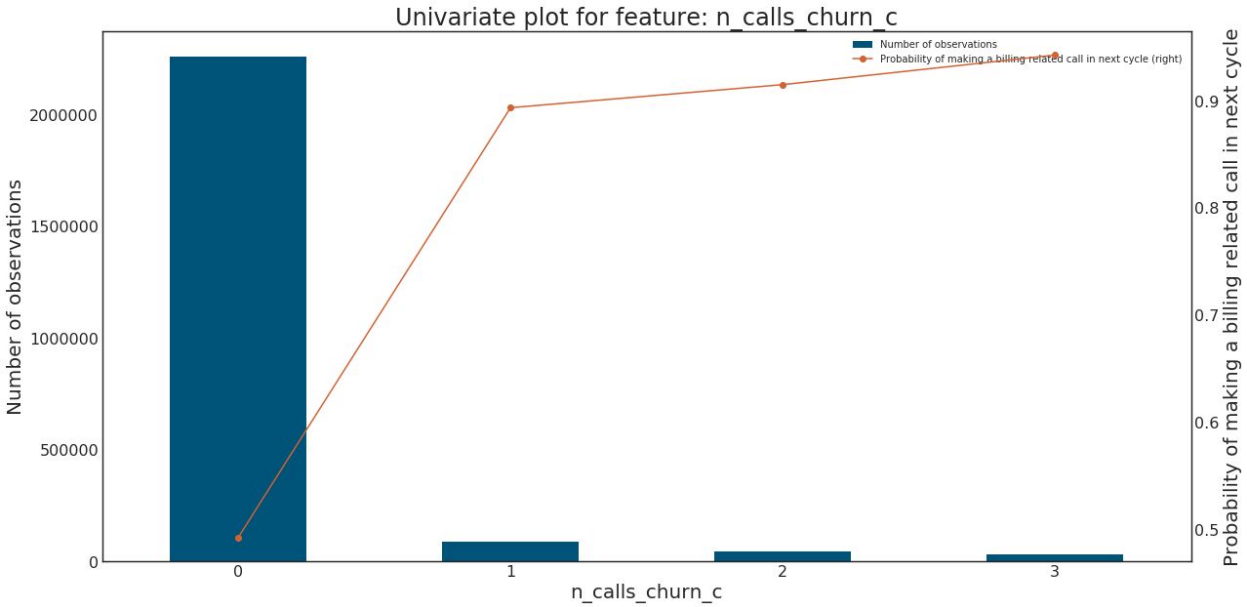

Figura 32. Gráfico univariante. Número de llamadas por motivos de fuga en el ciclo actual

- Número de llamadas por motivos de fuga en el ciclo anterior:

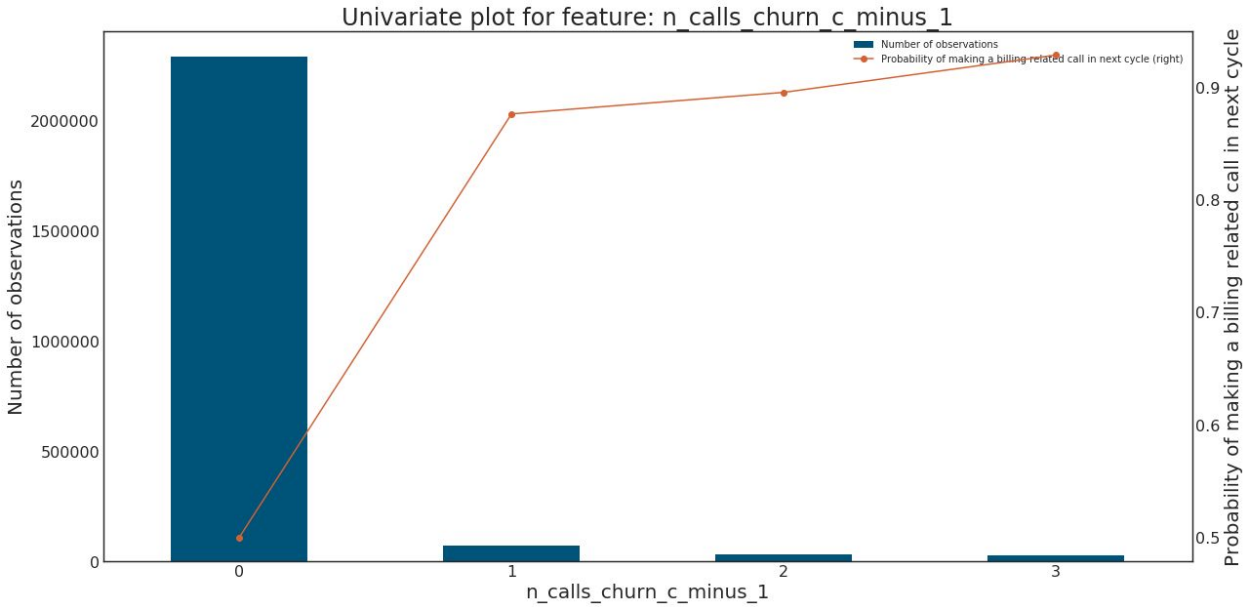

Figura 33. Gráfico univariante. Número de llamadas por motivos de fuga en el ciclo anterior

- Número de llamadas relacionadas con tarifas en el ciclo actual:

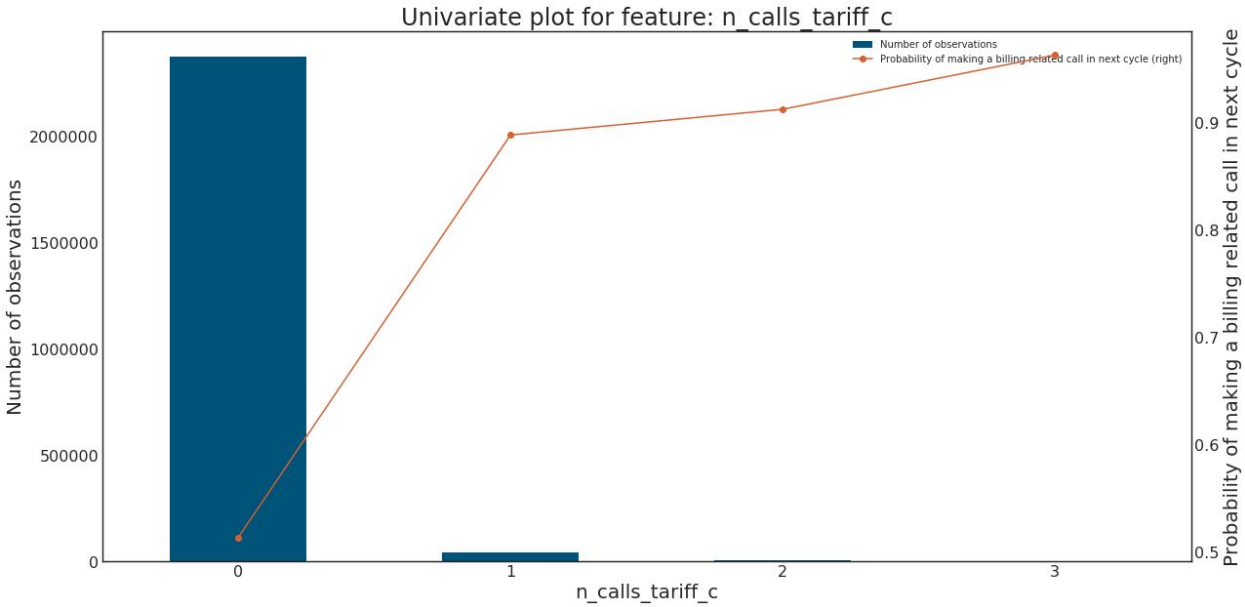

Figura 34. Gráfico univariante. Número de llamadas relacionadas con tarifas en el ciclo actual

- Número de llamadas relacionadas con tarifas en el ciclo anterior:

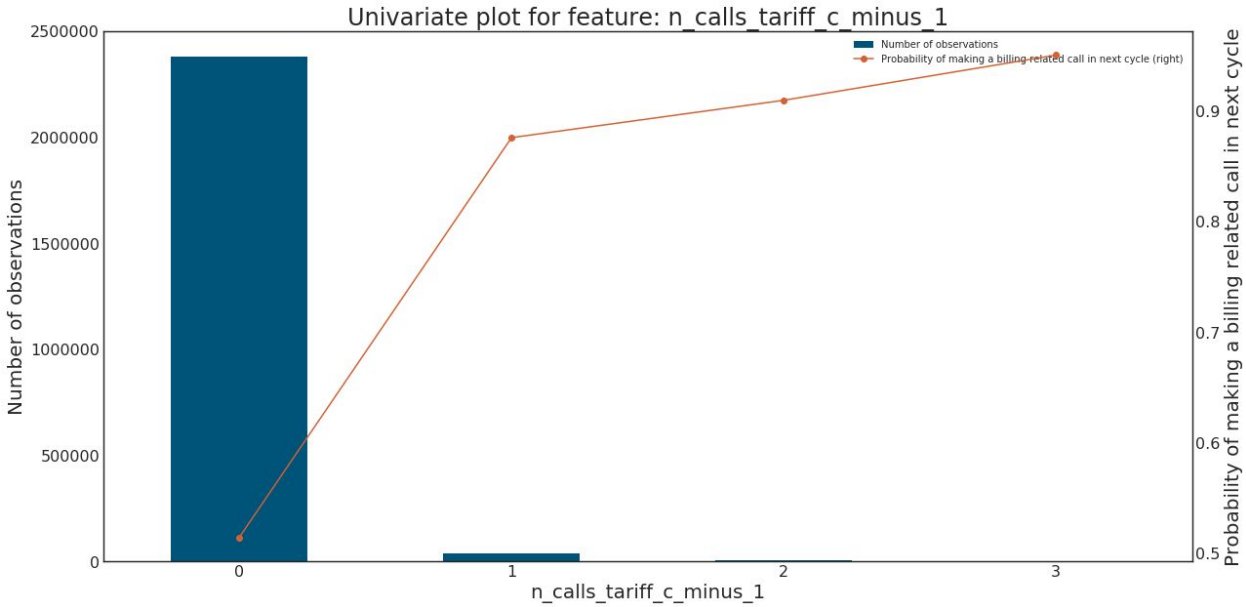

Figura 35. Gráfico univariante. Número de llamadas relacionadas con tarifas en el ciclo anterior

- Número de llamadas relacionadas con incidencias en ADSL en ciclo actual:

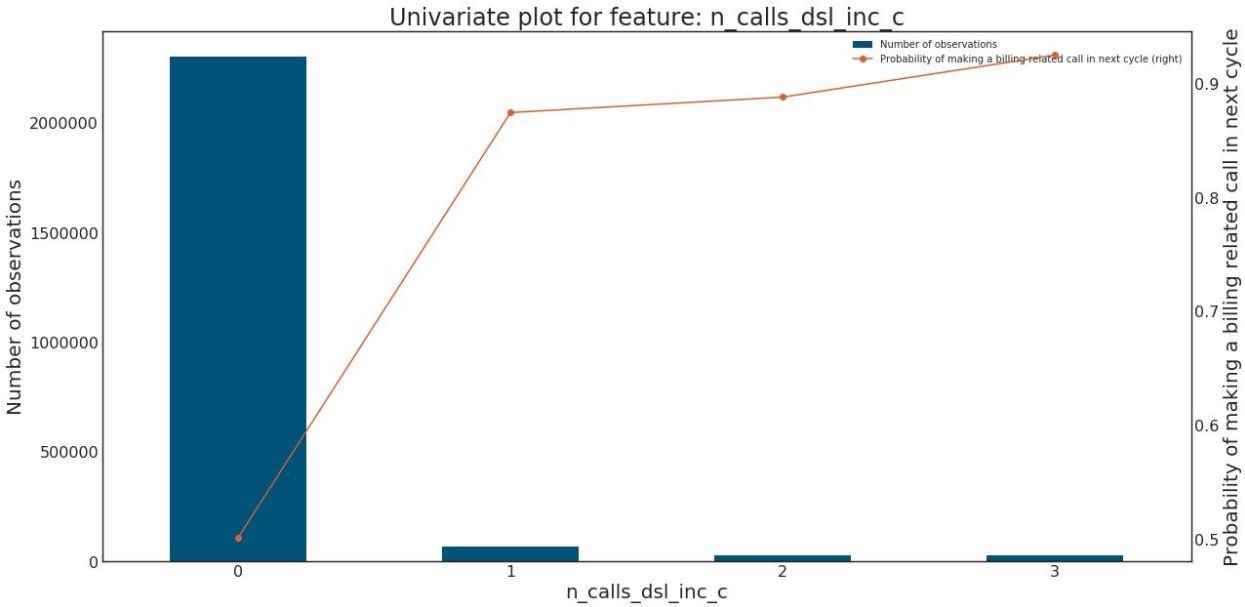

Figura 36. Gráfico univariante. Número de llamadas realizadas por incidencias en ADSL en ciclo actual

- Número de llamadas realizadas por incidencias en ADSL en ciclo anterior:

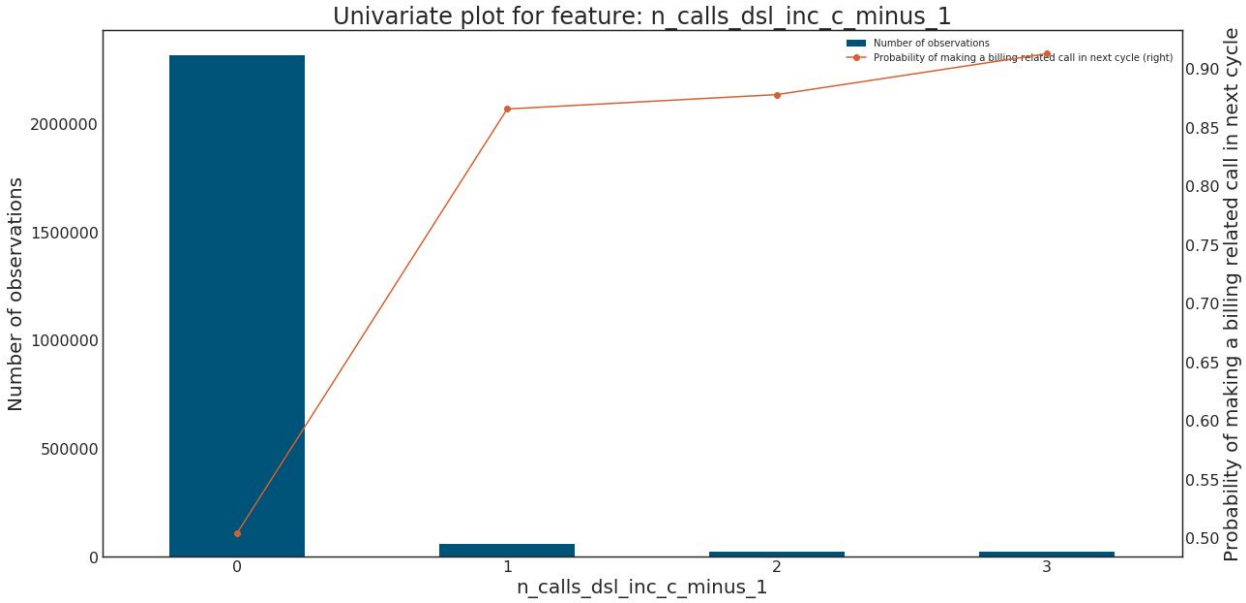

Figura 37. Gráfico univariante. Número de llamadas realizadas por incidencias en ADSL en ciclo anterior

- Número de llamadas relacionadas con incidencias en servicio móvil en el ciclo actual:

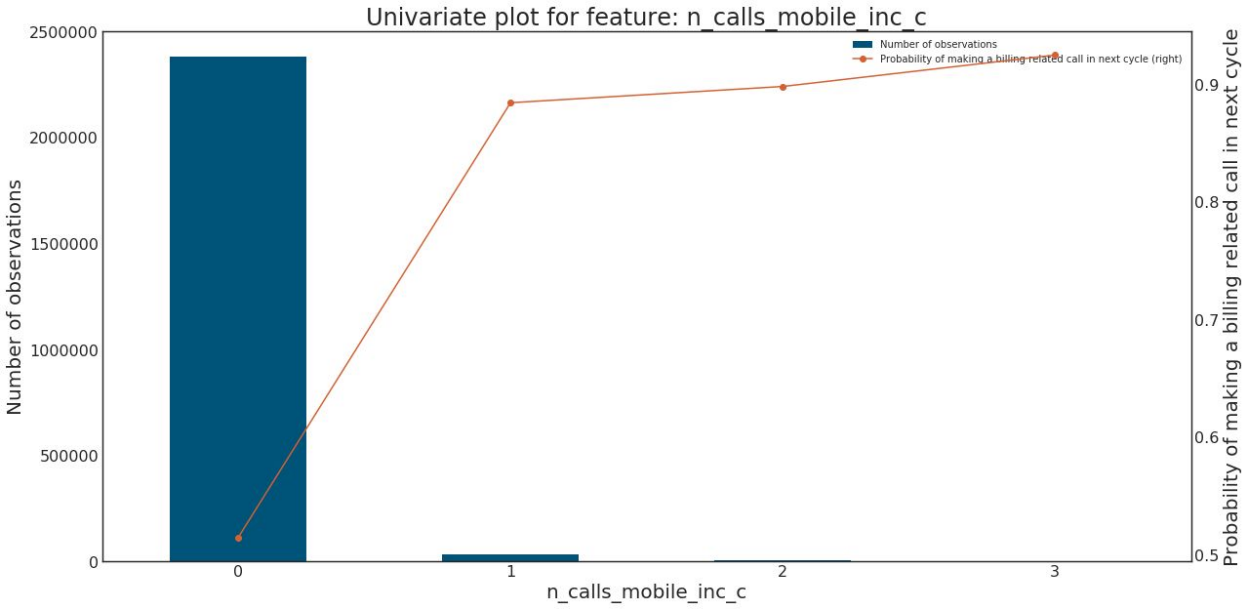

Figura 38. Gráfico univariante. Número de llamadas realizadas por incidencias en el servicio móvil, en ciclo actual

- Número de llamadas realizadas por incidencias en el servicio móvil, en el ciclo anterior:

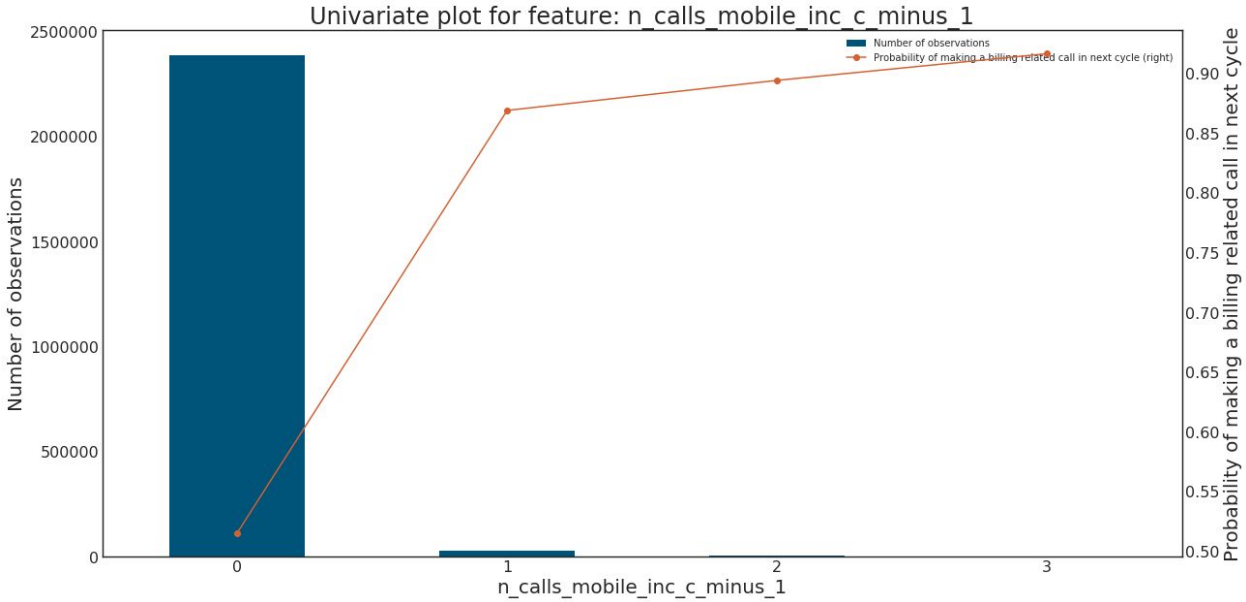

Figura 39. Gráfico univariante. Número de llamadas realizadas por incidencias en el servicio móvil, en ciclo anterior

- Número de llamadas realizadas por adquisición o mejora del terminal en el ciclo actual:

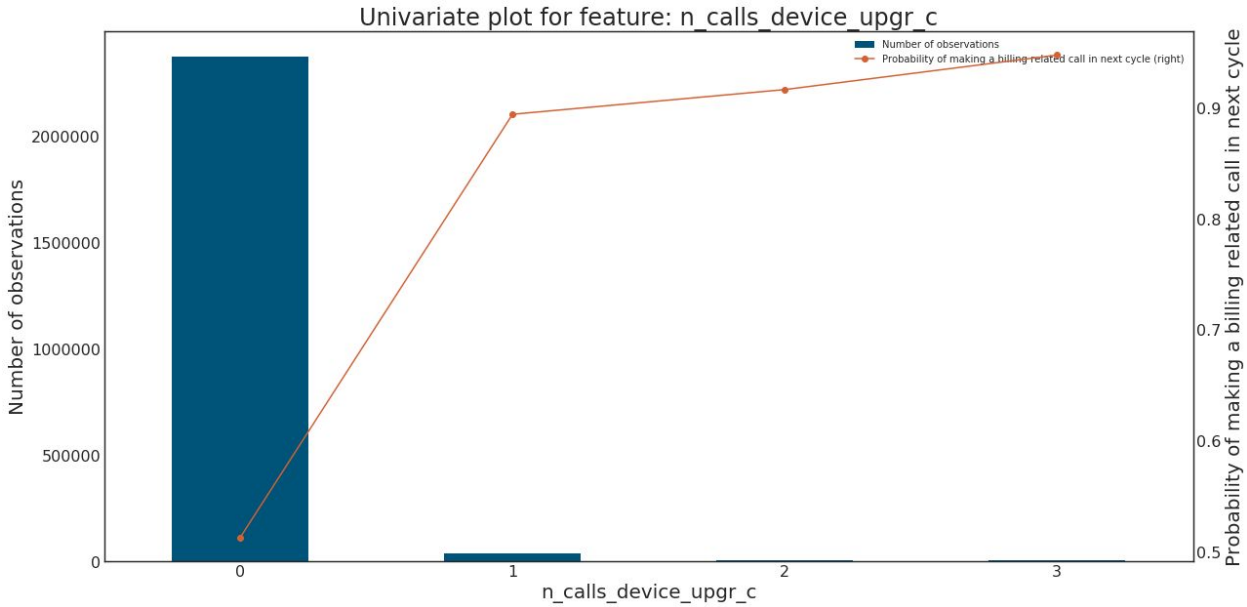

Figura 40. Gráfico univariante. Número de llamadas realizadas por por adquisición o mejora del terminal en el ciclo actual

- Número de llamadas realizadas por adquisición o mejora del terminal en el ciclo anterior:

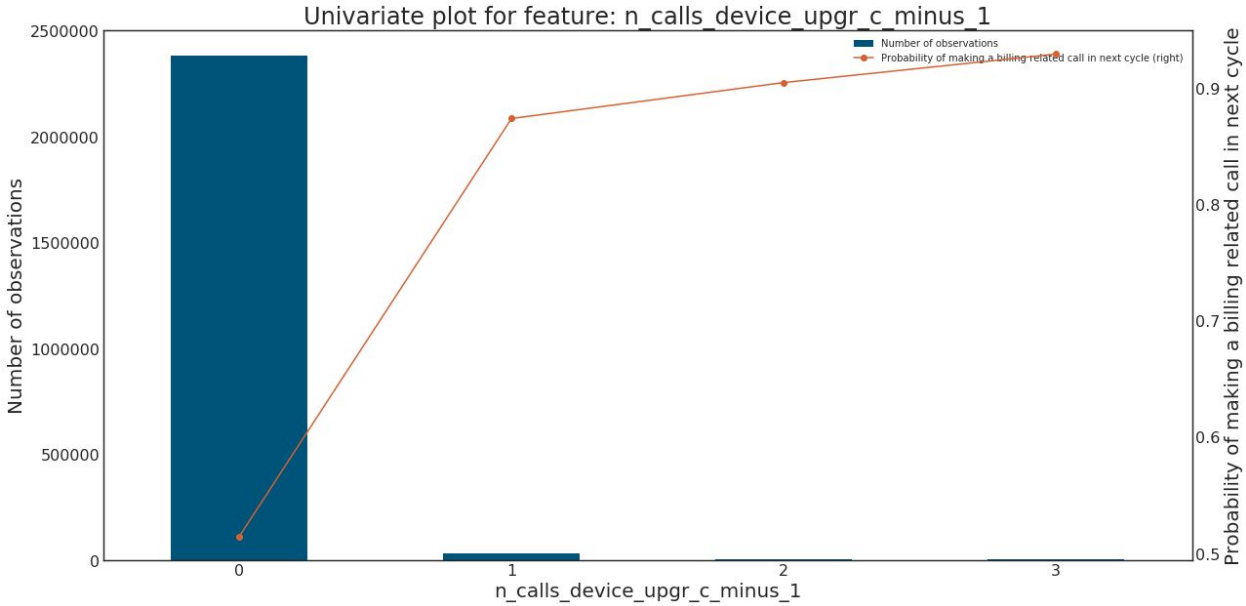

Figura 41. Gráfico univariante. Número de llamadas realizadas por por adquisición o mejora del terminal en el ciclo anterior

- Número de llamadas realizadas por motivos de reparación o entrega de terminal en el ciclo actual:

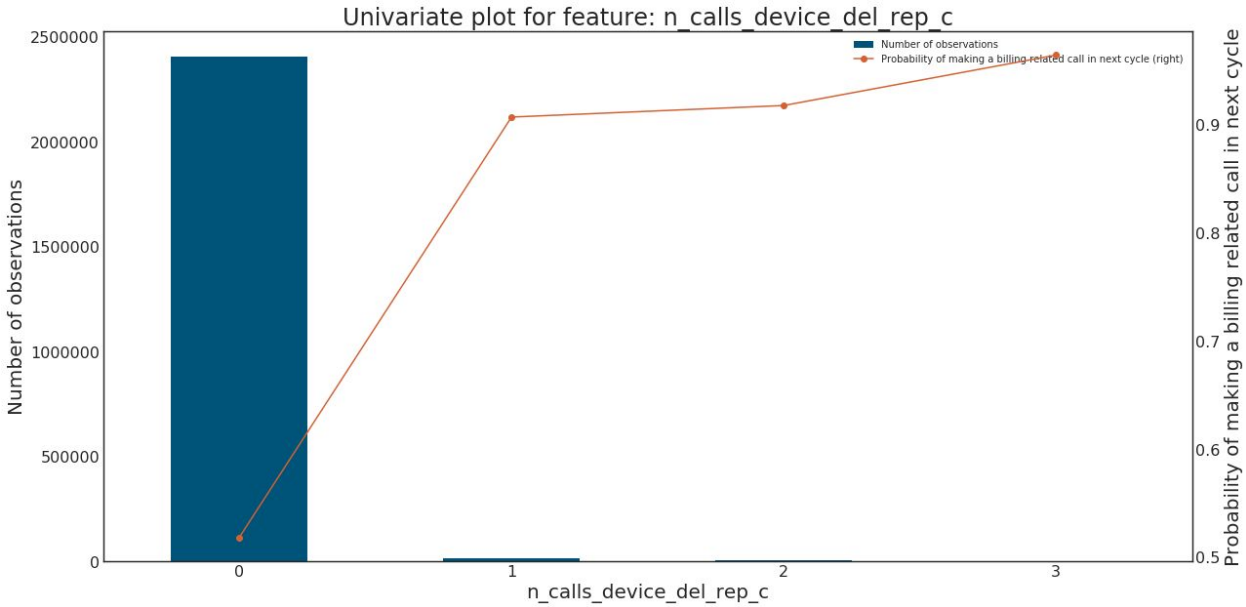

Figura 42. Gráfico univariante. Número de llamadas realizadas por por reparación o entrega de terminal en el ciclo actual

- Número de llamadas realizadas por motivos de reparación o entrega de terminal en el ciclo anterior:

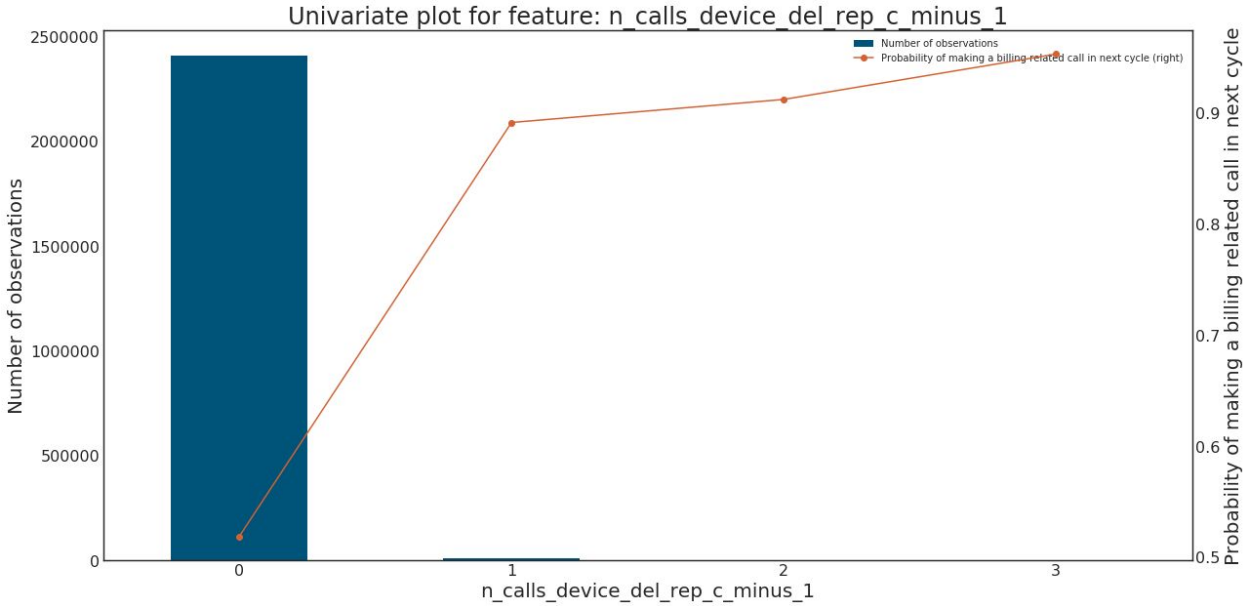

Figura 43. Gráfico univariante. Número de llamadas realizadas por por reparación o entrega de terminal en el ciclo anterior

- Número de llamadas realizadas por nueva alta en el ciclo actual:

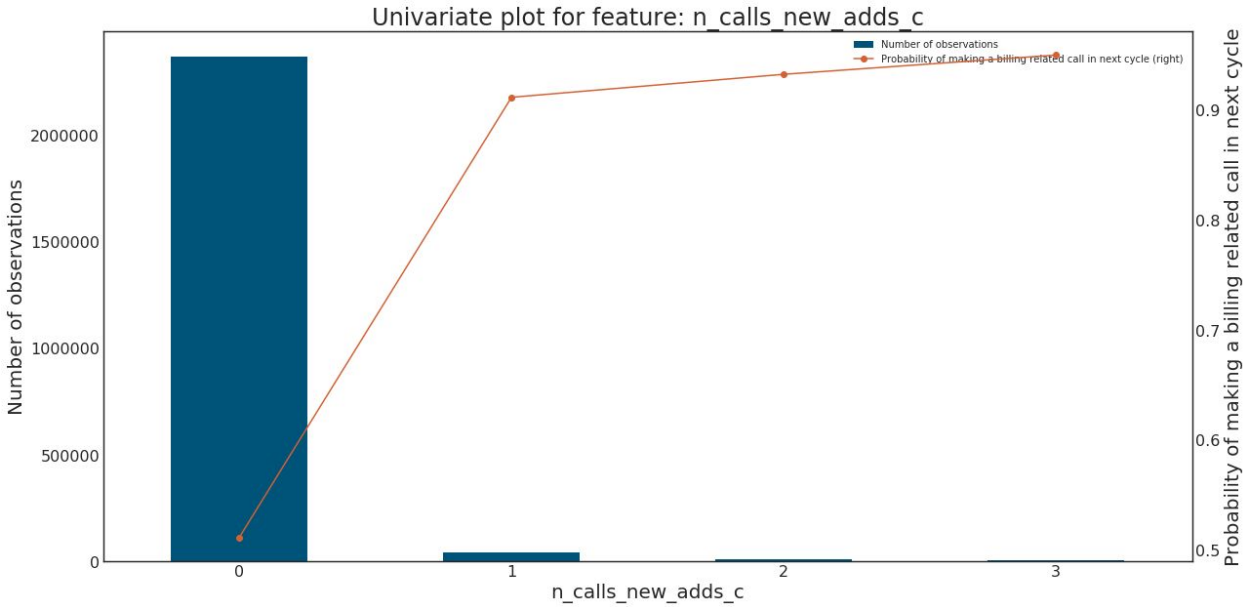

Figura 44. Gráfico univariante. Número de llamadas realizadas por nueva alta en el ciclo actual

- Número de llamadas realizadas por nueva alta en el ciclo anterior:

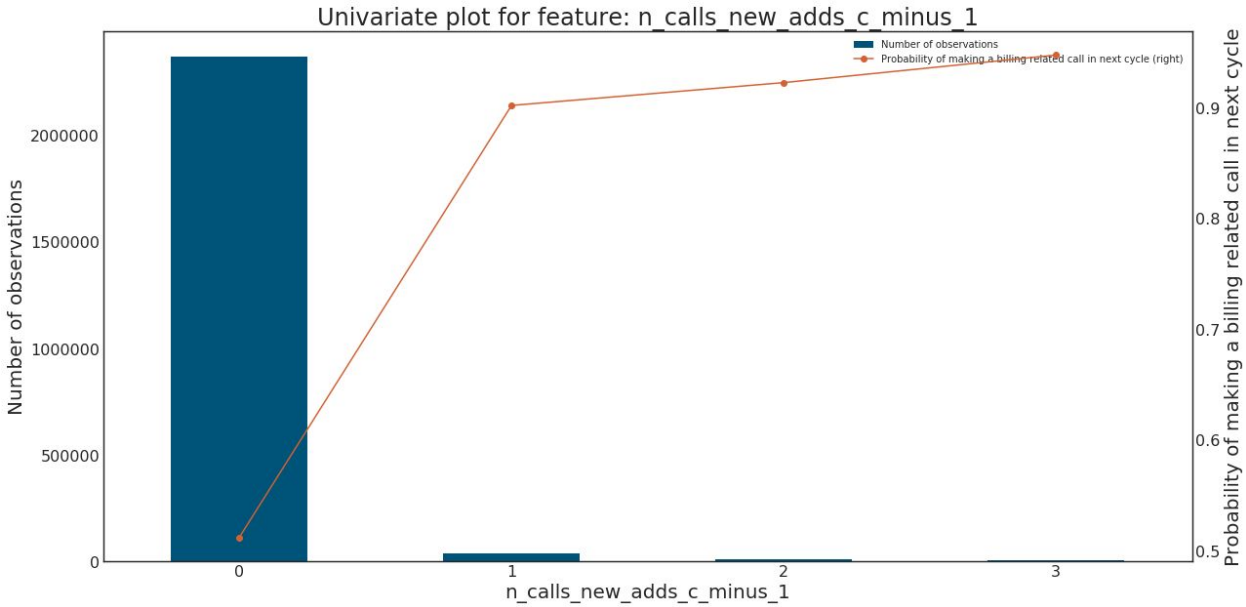

Figura 45. Gráfico univariante. Número de llamadas realizadas por nueva alta en el ciclo anterior

- Número de llamadas realizadas por servicio de mantenimiento en el ciclo actual:

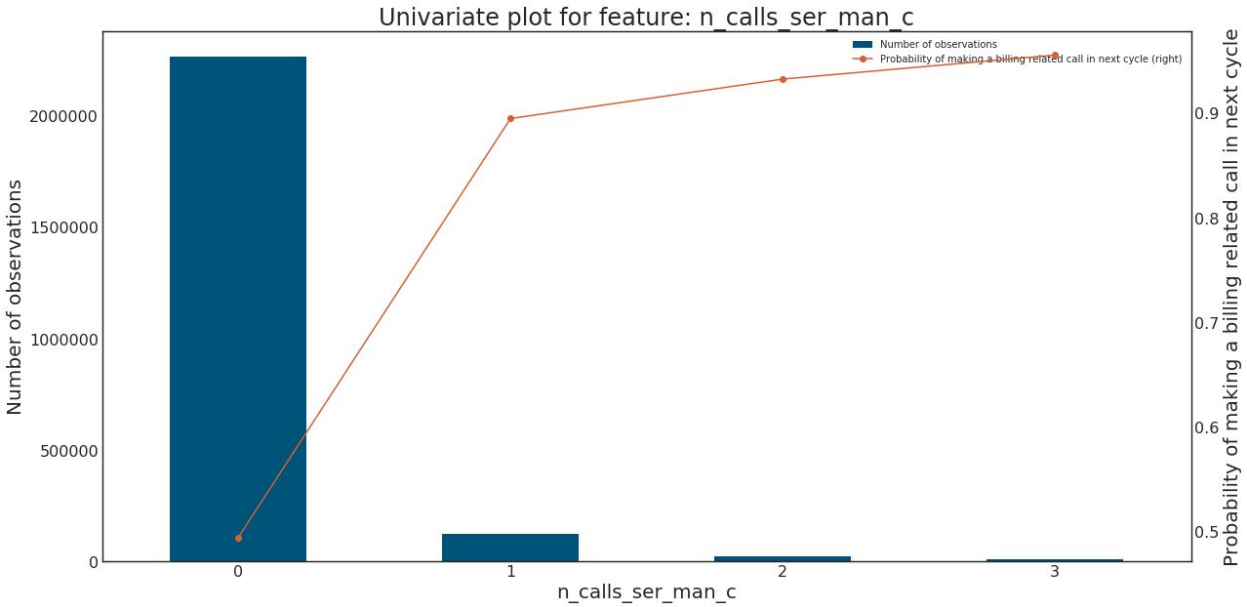

Figura 46. Gráfico univariante. Número de llamadas realizadas por servicio de mantenimiento en el ciclo actual

- Número de llamadas realizadas por servicio de mantenimiento en el ciclo anterior:

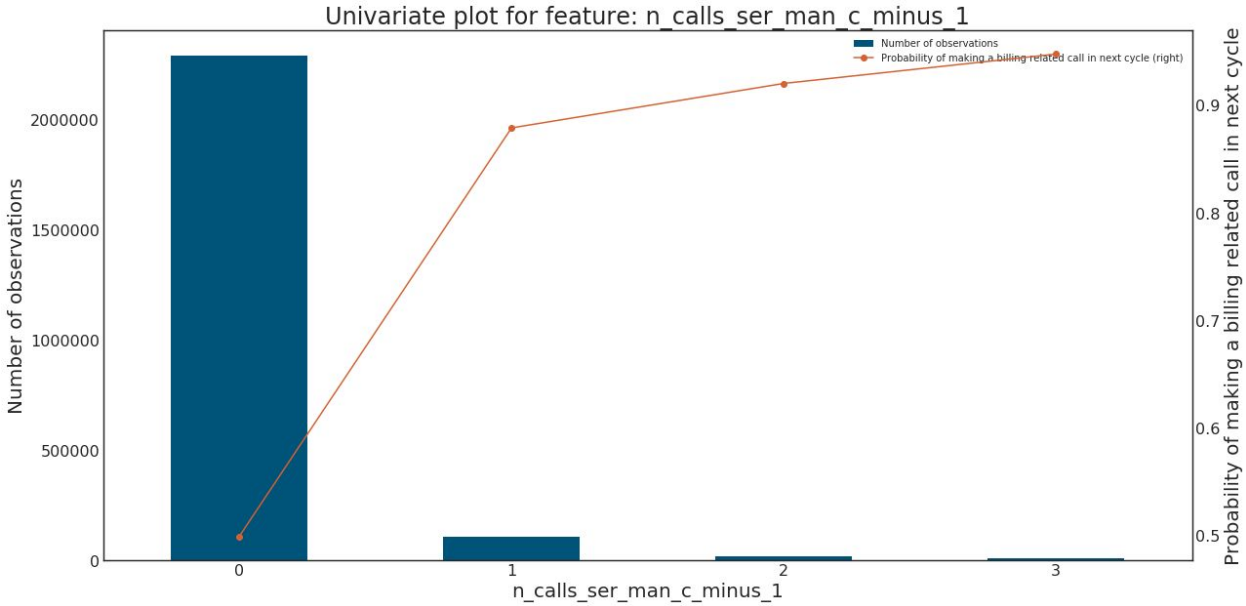

Figura 47. Gráfico univariante. Número de llamadas realizadas por servicio de mantenimiento en el ciclo anterior

- Plan de datos:

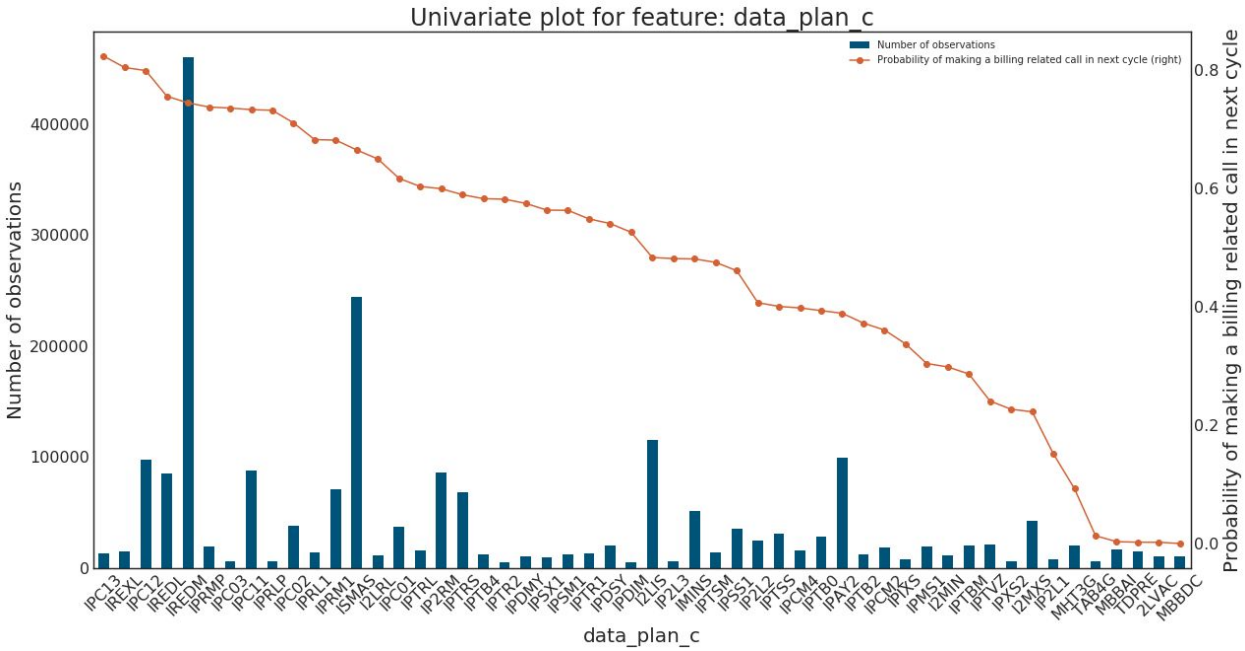

Figura 48. Gráfico univariante. Plan de datos

- Plan de voz:

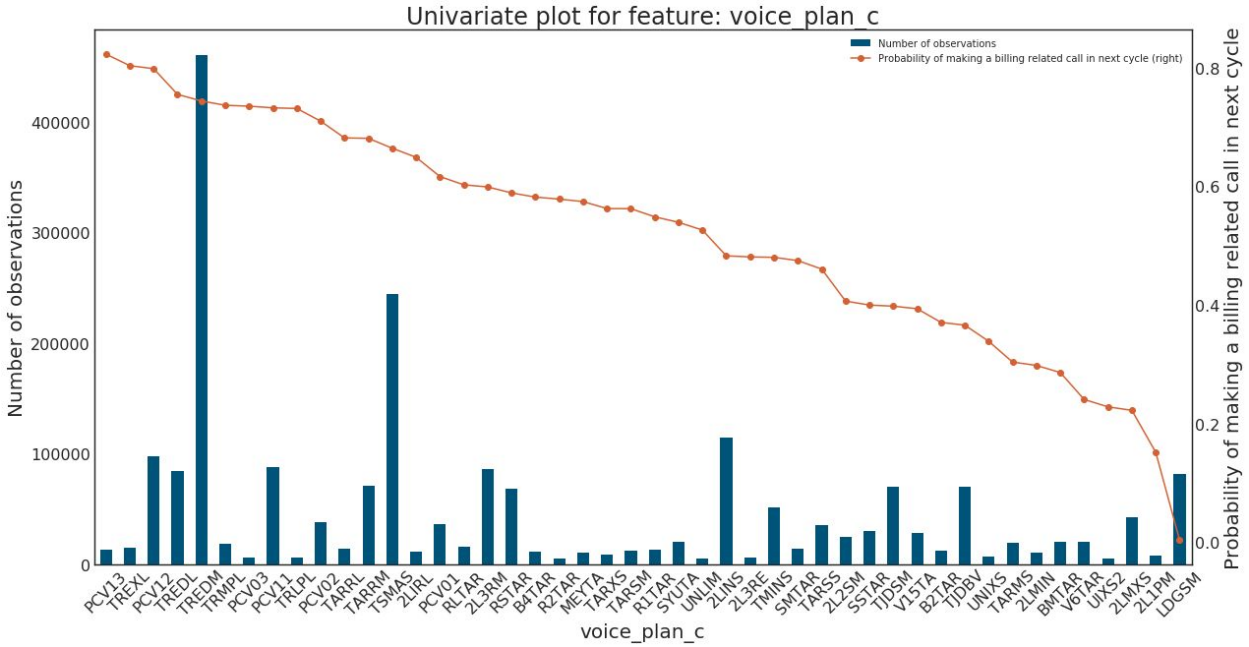

Figura 49. Gráfico univariante. Plan de voz

- Código de promoción de tipo Vodafone:

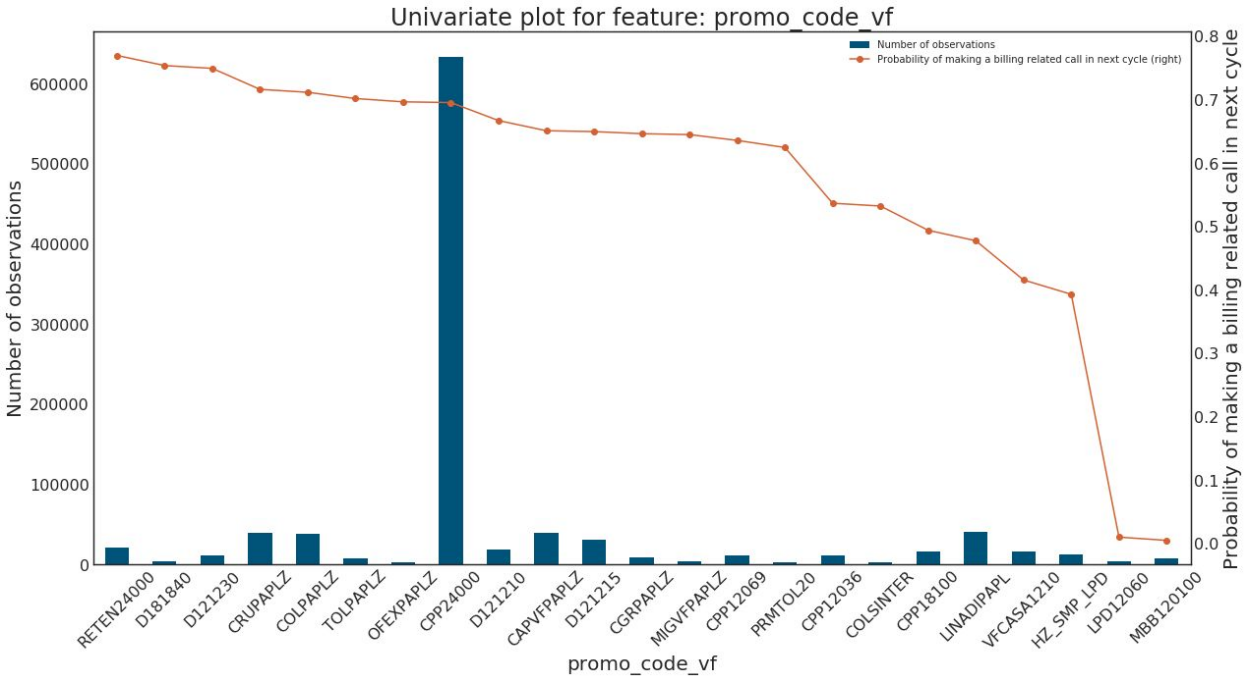

Figura 50. Gráfico univariante. Código de promoción de tipo Vodafone

- Código de promoción tipo tarifa:

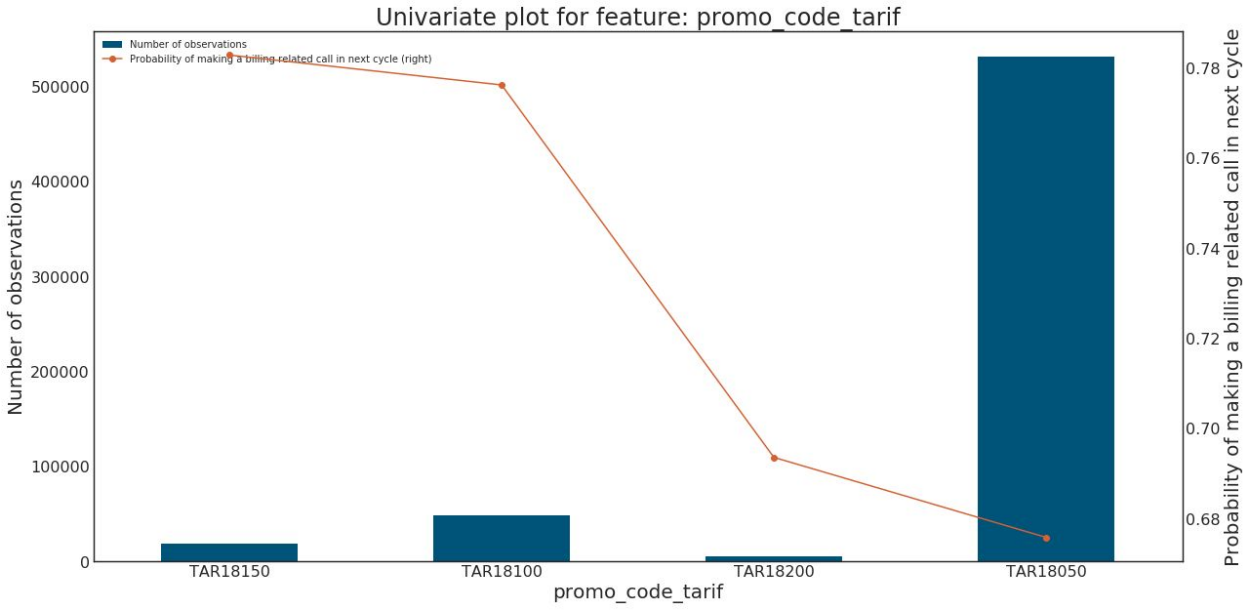

Figura 51. Gráfico univariante. Código de promoción de tipo tarifa

- Código postal:

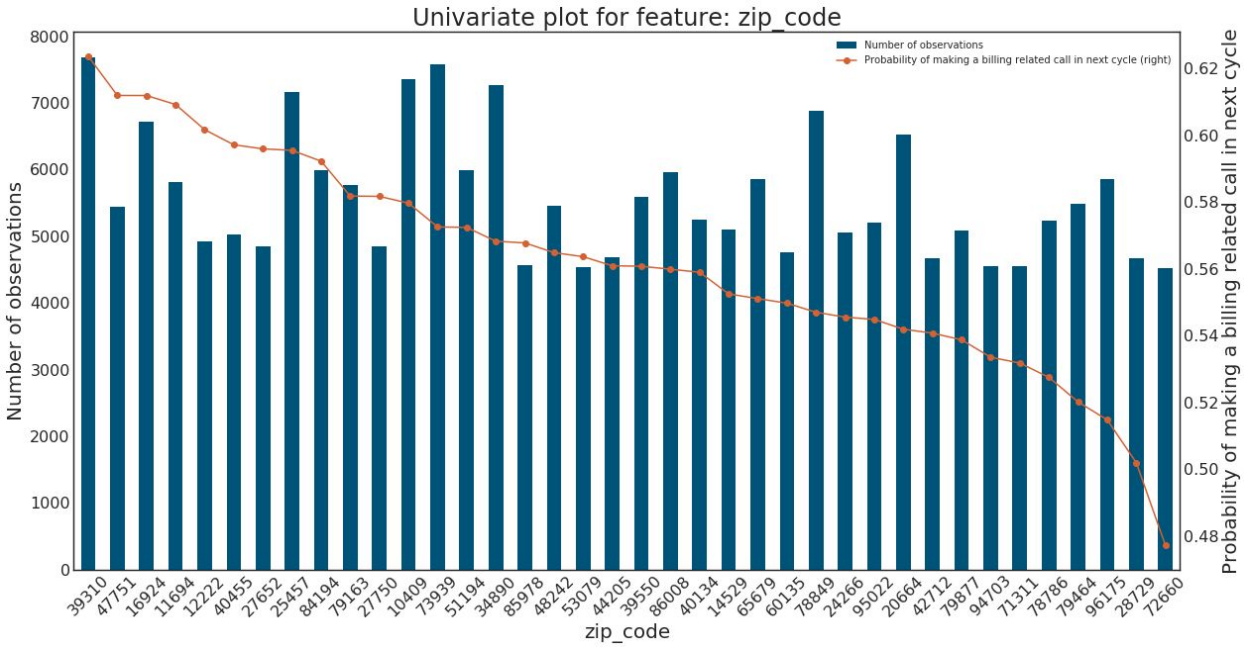

Figura 52. Gráfico univariante. Código Postal

- Región (definida por los dos primeros caracteres del código postal, correspondientes a la provincia):

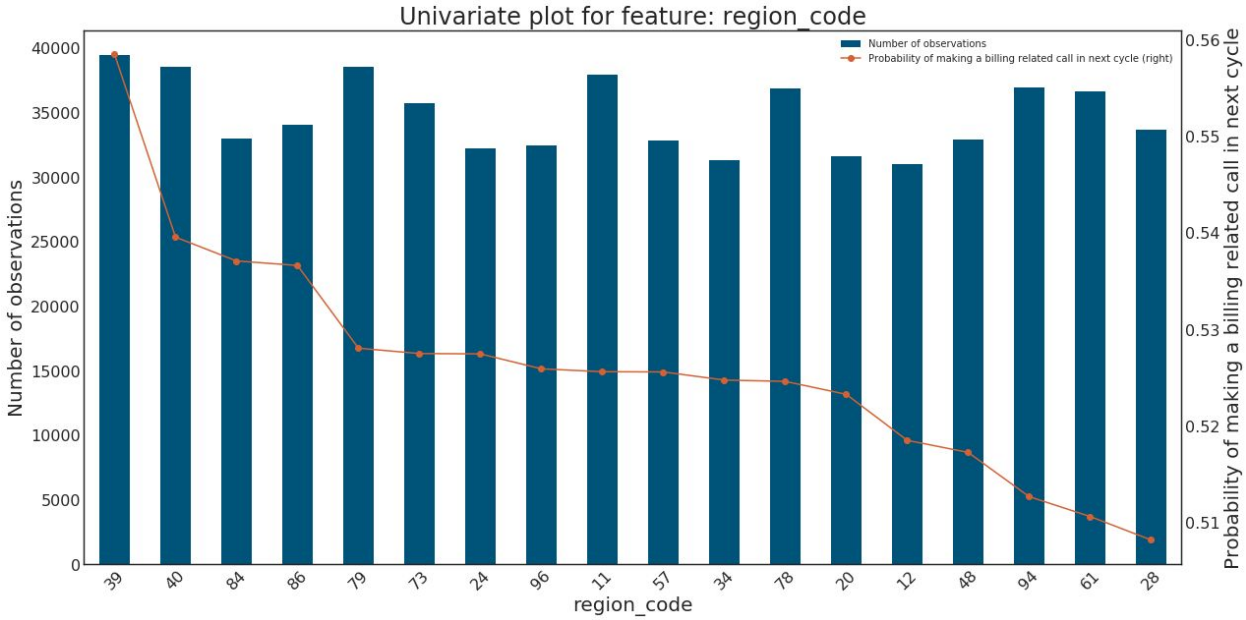

Figura 53. Gráfico univariante. Región

- Género:

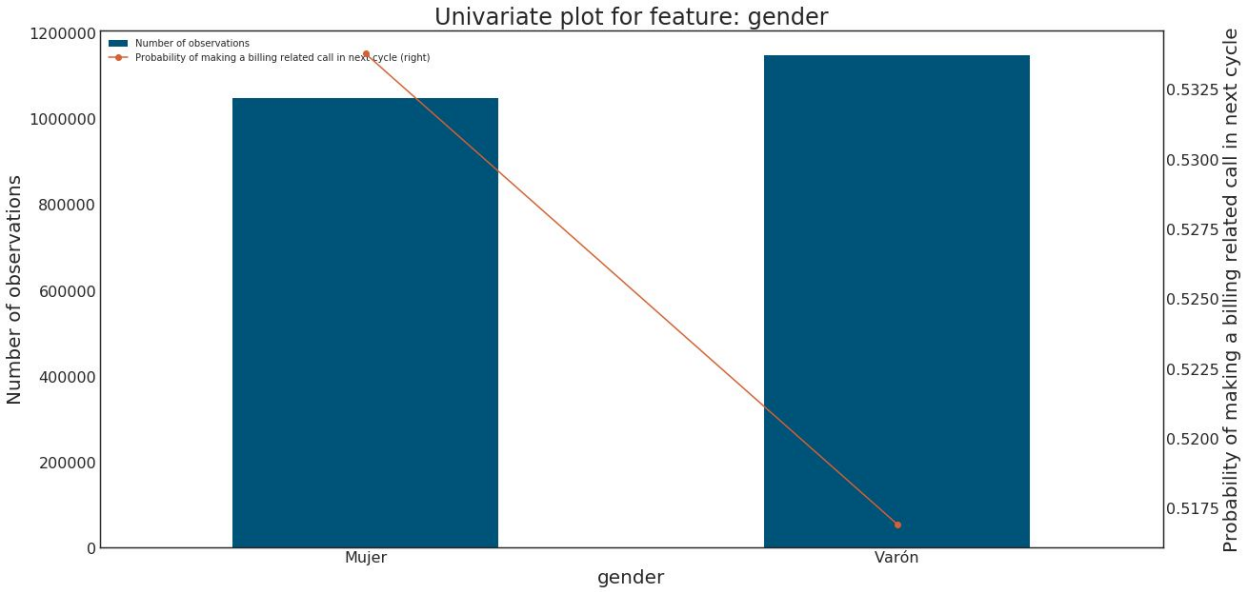

Figura 54. Gráfico univariante. Género

- Tipo de documento de identificación:

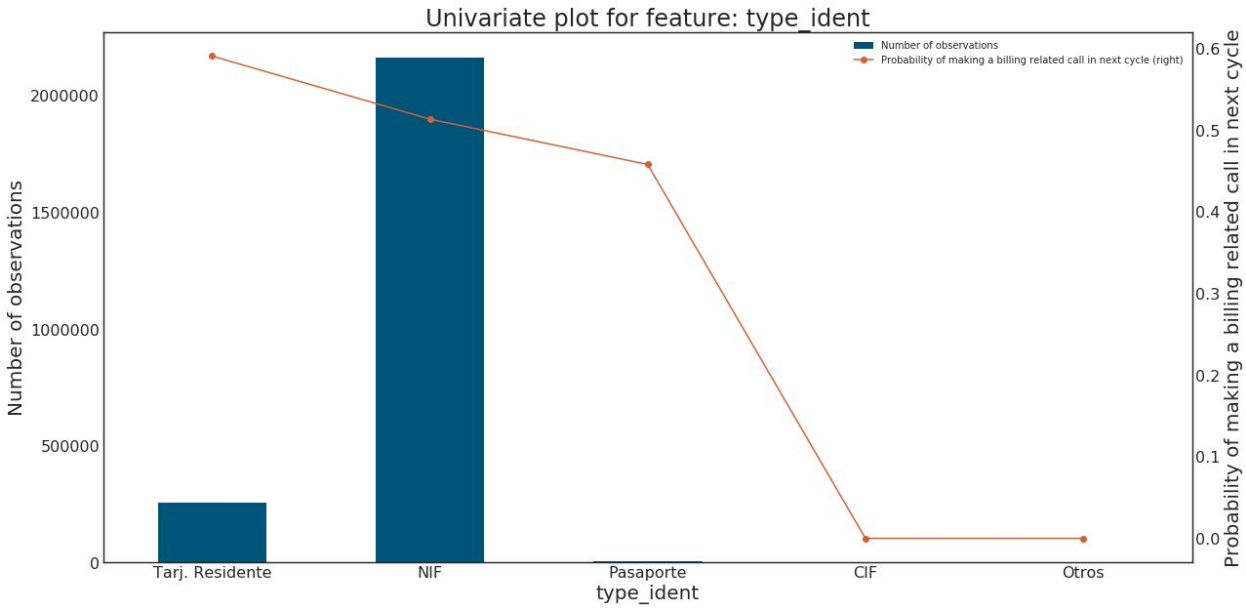

Figura 55. Gráfico univariante. Tipo de documento de identificación

- Nacionalidad:

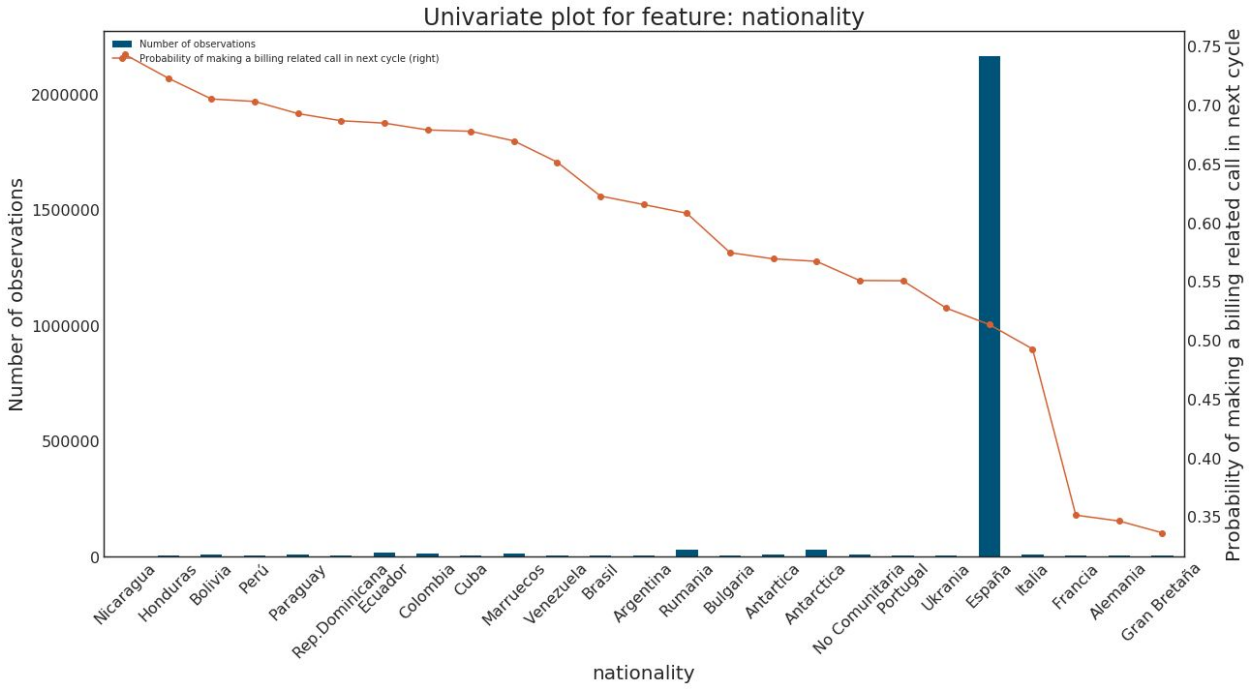

Figura 56. Gráfico univariante. Nacionalidad

- Edad:

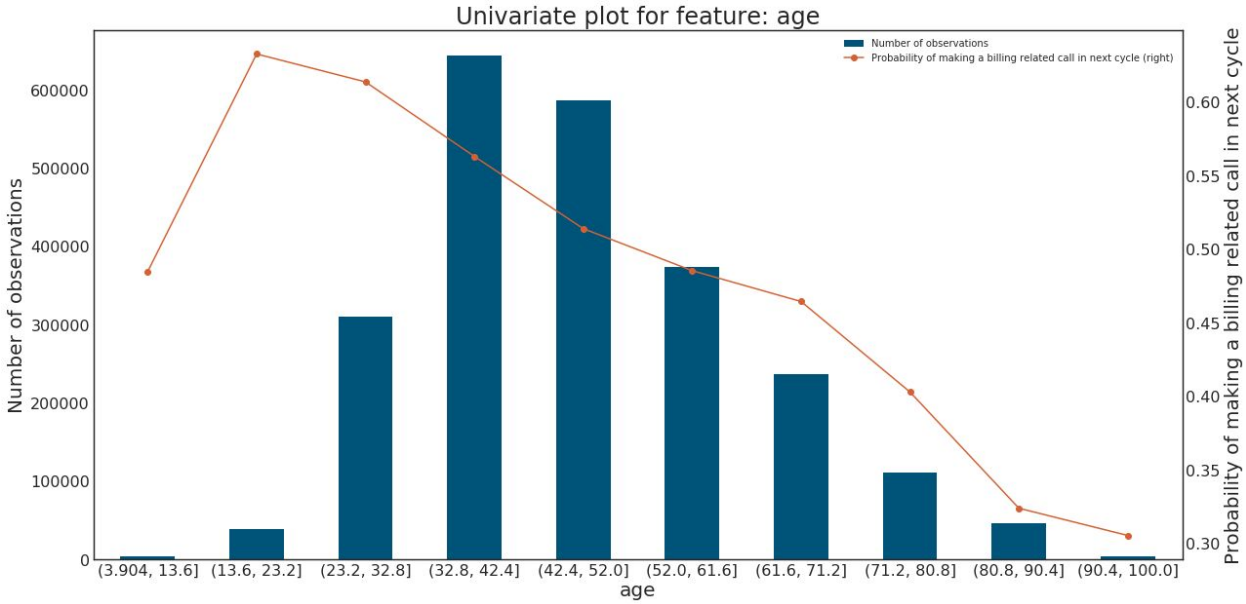

Figura 57. Gráfico univariante. Edad

Anexo 2. Otras métricas de rendimiento del modelo

El umbral es de 0.5 para todas las métricas de este apartado.

- Matriz de confusión en conjunto de entrenamiento:

		True label	
		0	1
Predicted label	0	6816980	100012
	1	6638	4065

Figura 58. Matriz de confusión. Conjunto de entrenamiento.

- Matriz de confusión en conjunto de test 1:

		True label	
		0	1
Predicted label	0	1966299	39154
	1	2520	1652

Figura 59. Matriz de confusión. Conjunto de test 1.

- Matriz de confusión en conjunto de test 2:

		True label	
		0	1
Predicted label	0	1476517	21980
	1	1236	724

Figura 60. Matriz de confusión. Conjunto de test 2.

- Accuracy:

Conjunto	Accuracy
Entrenamiento	0.985
Test 1	0.980

Test 2	0.985

Figura 61. Accuracy

- Precision:

Conjunto	Precision
Entrenamiento	0.380
Test 1	0.396
Test 2	0.369

Figura 62. Precision

- Recall:

Conjunto	Recall
Entrenamiento	0.040
Test 1	0.040
Test 2	0.032

Figura 63. Recall

- F1 Score:

Conjunto	F1 Score
Entrenamiento	0.071
Test 1	0.073
Test 2	0.059

Figura 64. F1 Score