UNIVERSIDAD DE JAÉN
Facultad de Ciencias Sociales y Jurídicas

La aplicación del Análisis Discriminante en el estudio de las causas de falta de fidelidad de clientes en sector de telefonía

Estadística y Empresa

Técnicas de Asociación y Clasificación

Valyano Daniil

2020

Introducción

El nicho de mercado de telefonía iba teniendo un crecimiento brusco justo después de su aparición en el siglo XX. El invento del teléfono fijo (y móvil, aunque mucho más tarde) provocó un cambio radical de la vida cotidiana de cada ser humano de nuestro planeta, facilitando tanto la comunicación interpersonal, como las negociaciones a nivel empresarial o incluso político. Hoy en día dicho mercado representa un nicho muy prometedor con competencia moderada, con lo cual la permanencia del cliente en una empresa constituye una parte casi mayoritaria del éxito. El negocio de las telecomunicaciones está severamente afectado por el problema de la fidelidad de los clientes. Con el objetivo de conocer los factores más importantes que contribuyen a que los clientes se den de baja del servicio de telefonía, un operador ha realizado un estudio sobre sus clientes sobre los que dispone de información acerca de su permanencia o abandono del proveedor.

La estadística poco a poco no solo se implementa a todos ámbitos socioeconómicos sino a ámbitos conexos e incluso opuestos. Los procedimientos probabilísticos y estadísticos de análisis e interpretación de datos o características de un conjunto de elementos permiten investigar rigurosamente cualquier cuestión que se plantea. Para poder analizar cualquier realidad social o económica de interés es imprescindible el uso de diferentes métodos estadísticos, tanto exploratorios como inferenciales que permitan la observación del fenómeno y la recolección de datos. Con esto se pretende lograr una mayor compresión de la realidad estudiada que facilite la toma de decisiones. De esta forma, en el presente trabajo se pretende emplear los métodos estadísticos para profundizar en los conocimientos que actualmente se tienen sobre los pilares básicos de la indústria telefónica. El objetivo principal de dicho estudio consiste en analizar los datos proporcionados por la empresa prestadora de servicios de telefonía, subrayar factores que tengan el mayor y menor impacto sobre la fidelidad de la marca.Ello a su vez puede aportar numerosa información al análisis DAFO, y, al mejorar puntos débiles encontrados, los convertirá en una fuerte ventaja competitiva en el mercado-oligopolio.

Metodología

El estudio se basa en el Análisis Discriminante, una técnica estadística multivariante cuya finalidad es analizar si existen diferencias significativas entre grupos de objetos respecto a un conjunto de variables

medidas sobre los mismos. En el caso de que existan, AD permite explicar en qué sentido se dan y facilitar procedimientos de clasificación sistemática de nuevas observaciones de origen desconocido en uno de los grupos analizados.

El planteamiento estadístico del problema de discriminación o clasificación es el siguiente: se dispone de un conjunto amplio de elementos que pueden provenir de dos o más poblaciones distintas. En cada elemento se ha observado un vector aleatorio cuya distribución se conoce en las poblaciones consideradas. El objetivo es clasificar un nuevo elemento en una de las poblaciones a partir del conocimiento de los valores de las variables del vector. El AD trata de obtener variables incorreladas a partir de combinaciones lineales de las variables independientes de forma que se logre la mejor separación entre los grupos previamente denidos. La discriminación se consigue asignando ponderaciones a cada variable de forma que se maximice la varianza entre los grupos en relación a la varianza dentro de los grupos.Para realizar este análisis se han utilizado el Software R/RStudio un lenguaje de código abierto con enfoque al análisis estadístico, equipos informáticos y gráficos junto con bibliografía referente.

Un estudio realizado sobre clientes de una empresa telefónica recogió información relevante a dichas variables: duración (duración del contrato), mensajes (número de mensajes de voz), diurno (cargo total por llamadas matutinas), tarde (cargo total por llamadas vespertinas), noche (cargo total por llamadas nocturnas), internacional (cargo total por llamadas internacionales), baja (solicitó la baja si=1/no=0), atención (número de llamadas al servicio de atención al cliente).

Análisis

El fichero *telefonía* contiene información sobre una muestra aleatoria de 5000 clientes (acerca de la permanencia de aquellos o su abandono del proveedor) relativa a las siguiente variables: solicitó la baja (sí=1/no=0), duración del contrato, número de mensaje de voz, cargo total por llamadas matutinas, cargo total por llamadas vespertinas, cargo total por llamadas nocturnas, cargo total por llamadas internacionales y número de llamadas al servicio de atención al cliente.

Empezamos nuestro análisis con la exploración de los datos proporcionados. El fichero de datos está compuesto por 5000 observaciones de 8 variables, además comprobamos que no contiene datos faltantes. Una de las variables (baja) es categórica, tiene dos niveles 1 y 0, correspondientes al hecho del abandono o permanencia del cliente en la empresa. A continuación observamos 7 variables numéricas que van a tratar de explicar la capacidad de separación de grupos en el factor “baja”. Observamos que los niveles del factor “baja” no tienen tamaños muestrales iguales (4293 clientes permanentes/707 se han dado de baja), lo que

supone una condición no deseable para la realización del AD. Con el fin de evaluar la capacidad de discriminación en las variables numéricas, procedemos a la búsqueda de la función discriminante. En este caso se trata de una única función discriminante, dado que el factor "baja" tiene solo 2 niveles.

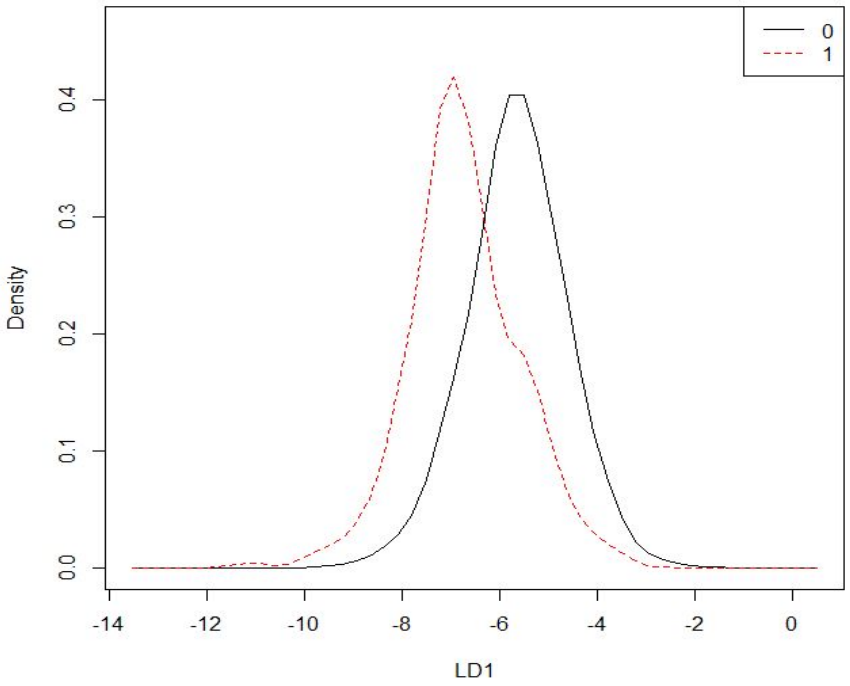


Ilustración 1

A la izquierda se puede observar el gráfico en el que se visualizan los dos histogramas suavizados correspondientes a los grupos 1 y 0 en dimensión de una sola función discriminante (Ilustración 1). Según dicho gráfico se observa que las 7 variables tienen cierta capacidad de separación de grupos. No obstante, existe un cierto grado de solapamiento entre ambos.

A continuación nos aparecen los centroides en esta dimensión discriminante, grupo 0 : -5.65 y grupo 1: -6.7203. Ello se verifica claramente en la representación gráfica de los histogramas, como aparece en color rojo más a la izquierda proporcionando un valor más negativo del centroide que el histograma negro, perteneciente al grupo 0.

Según la función discriminante estandarizada aparece un valor de variable "duración" proporcionalmente inferior a los coeficientes de variables restantes, lo que pueda indicar que posteriormente podremos prescindir de ella sin perder capacidad discriminante.

Función discriminante estandarizada:

duracion	mensajes	diurno	tarde	noche	internac	atención
-0.068	0.32	-0.664	-0.314	-0.154	-0.228	-0.677

Con la intención de sacar algunas conclusiones, previamente tenemos que asegurarnos que la función discriminante es significativa, es decir, las 7 variables aportan suficiente información para separar los grupos. Para ello comprobamos la significación global de la función discriminante. Como salida aparece un estadístico *T2=646.053*, con el *p-valor=0*, así que seremos capaces de separar ambos grupos.

Dado que la función discriminante es significativa, podemos fiarnos de sus cocientes que nos van a conllevar a varias conclusiones respecto a la influencia de las variables en la clasificación de una observación a uno de los grupos. Así, por ejemplo, podemos considerar cocientes elevados negativos (marcados en rojo) como los más influyentes en el abandono del proveedor. Mientras tanto los que representan un valor más grande positivo (marcados en verde) influyen positivamente en la permanencia del cliente en la empresa.

Retomamos nuestra cuestión planteada anteriormente sobre la importancia de variable "duración" en la separación de grupos, evaluando los estadísticos lambda-parciales que examinan la contribución individual de cada variable. Se rechaza la hipótesis H_0 con un *p-valor* = *0.113* , por lo tanto llegamos a la conclusión de que podemos prescindir de la variable "duración" (Tabla 1). Para ello vamos a realizar una comprobación adicional. Otras variables son necesarias mantenerlas para conseguir explicar la separación de grupos, como representan p-valores próximos a *0*.

VARIABLE	**LAMBDA-PARCIALES**	**P-VALOR**
duración	0.999	0.113
mensajes	0.989	0.000
diurno	0.953	0.000
tarde	0.989	0.000
noche	0.997	0.000
internacional	0.994	0.000
atención	0.951	0.000

Tabla 1

Sin embargo, el AD por pasos: método forward nos lleva a la conclusión contraria de dejar la variable "duración" (Tabla 2)

VARIABLE	Wilks.lambda	F.statistics.overall	p.value.overall	F.statistics.diff	p.value.diff
atención	0.9548165	236.51350	3.492304e-52	236.513504	3.492304e-52
internacional	0.9119174	241.33145	8.888623e-101	235.072696	0.000000e+0
noche	0.9024601	179.99294	8.051482e-111	52.355476	5.333511e-13
tarde	0.8931732	149.35512	7.756339e-121	51.936347	6.589174e-13
diurno	0.8882494	125.65897	1.093847e-125	27.682992	1.488859e-07
mensajes	0.8859785	107.09611	2.253504e-127	12.797550	3.503634e-04
duración	0.8855339	92.18249	7.021935e-127	2.506831	1.134170e-01

Tabla 2

Pasamos al análisis de la posible multicolinealidad o alta dependencia lineal en un conjunto variables métricas, obteniendo una matriz de correlaciones. Sin duda alguna se observa que la cantidad más elevada en valor absoluto corresponde a *0.017* que no es suficiente grande para generar problemas.

	duracion	**mensajes**	**diurno**	**tarde**	**noche**	**internac**	**atencion**
duracion	1.000	-0.015	-0.001	-0.010	0.001	0.001	-0.001
mensajes	-0.015	1.000	0.005	0.019	0.006	0.003	-0.007
diurno	-0.001	0.005	1.000	-0.011	0.012	-0.019	0.003
tarde	-0.010	0.019	-0.011	1.000	-0.017	0.000	-0.014
noche	0.001	0.006	0.012	-0.017	1.000	-0.007	-0.009
internac	0.001	0.003	-0.019	0.000	-0.007	1.000	-0.012
atencion	-0.001	-0.007	0.003	-0.014	-0.009	-0.012	1.000

Comprobamos que el menor de los autovalores no es demasiado pequeño (Tabla 3), por lo tanto el número/indice de condición (IC) calculado como $\lambda_1/\lambda_7 = 1.079$ (un valor superior a 30 indica la presencia de multicolinealidad severa).

Autovalores de R:

1.04	1.026	1.012	0.995	0.983	0.981	0.964

Tabla 3

Los factores de inflación de la varianza son cercanos a 1, cuando valores por encima de 5 son indicio de problemas por causa de la multicolinealidad (Tabla 3.1).

FIVs:

1	1.001	1.001	1.001	1.001	1.001	1

Tabla 3.1

Finalmente, una medida comprendida entre 0 y 1 que utiliza los inversos de los autovalores de R fue propuesta por *Heo = 0.001* y el *det(R) = 0.998,* con lo cual sacamos la conclusión de que la multicolinealidad en este estudio no va a ser ningún problema.

Procedemos al análisis de la capacidad predictiva de nuestro modelo. En esta parte de nuestro estudio vamos a fijar 2 modelos diferentes: con variable "duracion" y sin ella. Dado que sacamos conclusiones contradictorias a la hora de analizar la colaboración de dicha variable,analizaremos la tasa de clasificaciones correctas (*CCR*) en ambos casos para comprobar si realmente perdemos precisión a la hora de separar los grupos (Tabla 4).

Condición	M.Confusión	M. Confusión dejando uno fuera
Clas. Lineal "duración" inluida	*CCR = 0.86*	*CCR = 0.86*
Clas. Cuadr "duración" inluida	*CCR = 0.895*	*CCR = 0.894*
Clas. lineal "duración" excluida	*CCR = 0.859*	*CCR = 0.859*
Clas. Cuadr"duración" excluida	*CCR = 0.894*	*CCR = 0.894*

Tabla 4

Optamos por prescindir de la variable "duración", debido a que el porcentaje de clasificaciones correctas disminuye muy poco (89.4%), en comparación con con la Clasificación cuadrática con ella (89.5%).

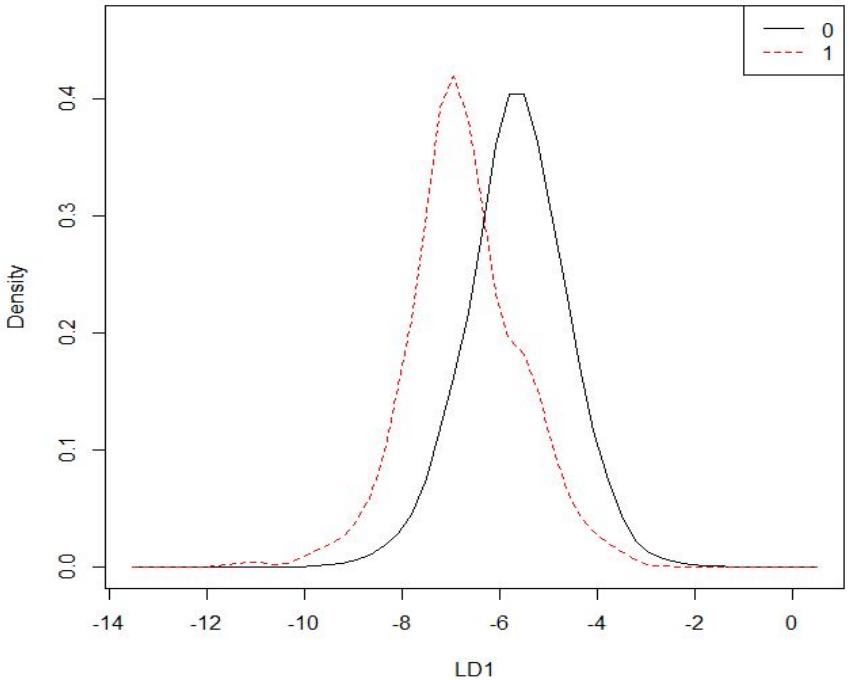
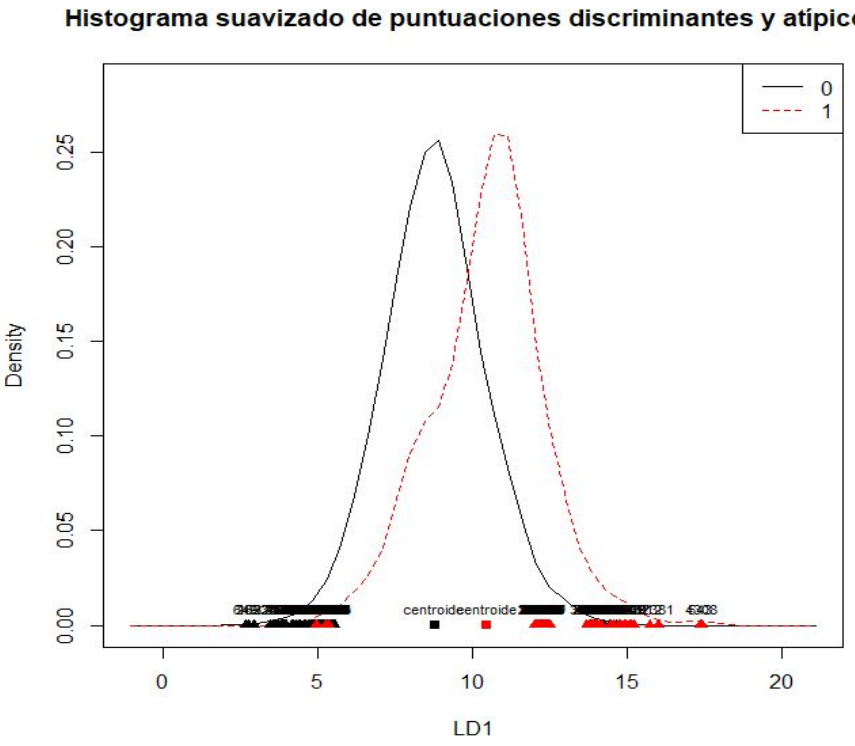


Ilustración 2

El gráfico de los histogramas suavizadas prescindiendo de la variable "duración" aparece en la Ilustración 2. Se nota poca diferencia visual con respecto a la Ilustración 1, además de que se puede observar una evidente separación de grupos y un grado moderado de solapamiento entre ellos.

En la Ilustración 3 podemos observar la clasificación lineal prescindiendo de la variable "duración" junto con los centroides y atípicos. La elección de clasificación cuadrática se debe a la salida del Test de homogeneidad de las matrices de covarianza. Según este, las matrices de covarianza tienen diferencias significativas. Esta conclusión se verifica mediante un estadístico *Chi-q = 797.38* y *p-valor* asociado de 0. Cabe mencionar que en el caso de heterogeneidad de las matrices de covarianza, en el que estamos debido al test realizado anteriormente, no tiene sentido el análisis de los puntos atípicos.



lustración 3

La función de clasificación para ambos grupos aparece a continuación:

baja	mensajes	diurno	tarde	noche	internac	atención	cte
0	0.022	0.406	0.984	1.803	5.271	1.267	-30.897
1	-0.001	0.481	1.057	1.871	5.578	1.797	-38.708

Tabla 5

Se observa a simple vista una obvia diferencia entre los grupos de variable-factor "baja" en todas las variables métricas, mientras tanto la variable "duración" mostraba coeficientes -0.066 y -0.067 para "baja" = 0 y "baja" = 1 respectivamente. De ello, concluimos que dicha variable la podemos considerar como poca importante en la separación de grupos.

Examinemos la función discriminante al haber eliminado una variable para asegurarnos que sigue siendo significativa. Empezamos con los centroides de ambos grupos: $1 = -6.5379$, $0 = -5.4783$ que evidentemente no han cambiado de manera radical. La función discriminante se considera significativa con $T2 = 643.22$ y *p-valor* $= 0$.

La última cuestión que nos pueda preocupar es la normalidad multivariante de nuestro conjunto de datos, por ello realizamos una serie de tests de normalidad multivariante: de Mardia de curtosis y asimetría, Henze-Zirkler para cada uno de los niveles de la variable categórica "baja". Prescindimos del test de Royson porque el tamaño muestral de una de las muestras es superior a 2000. Resulta que en ambos casos se rechaza la hipótesis de la normalidad multivariante con *p-valores* ≈ 0. No obstante, es una situación muy común en AD y en nuestro caso, donde la clasificación resulta correcta en 89.5% casos, no causa ningún problema.

Resultados

Realizados varios procedimientos y técnicas del AD llegamos a identificar una función discriminante significativa y capaz de describir de manera óptima la separación de grupos. Además se consiguió encontrar una función de clasificación con buena capacidad predictiva para la distribución posterior de observaciones nuevas según variables métricas ya existentes. Se elaboraron contenidos tanto visuales como escritos para justificar métodos y procedimientos llevados a cabo durante dicho estudio.

Conclusiones y Recomendaciones

Siendo una técnica de análisis multivariante AD pueden ser muy útil a la hora de reducir una cantidad de información irrelevante. En nuestro caso hemos justificado que la variable "duración" no ha sido significativa a la hora de explicar la fidelidad de la marca del consumidor. Con lo cual para esta cuestión se recomienda: prescindir de la duración del contrato durante dicho análisis, recopilar datos nuevos sobre clientes que parecen tener influencia en la permanencia del cliente con los servicios de la compañía, volver a solicitar una investigación parecida con el transcurso de tiempo con el fin de obtener conclusiones relevantes al día.

Se ha revelado que el cargo total por llamadas matutinas junto con el número de llamadas al servicio de atención al cliente constituyen los factores subyacentes sobre la ruptura del contrato entre su empresa y el cliente. Además, entre las variables que puedan tener cierta influencia negativa en la permanencia del uso de sus servicios por el consumidor son: cargo total por llamadas vespertinas, internacionales y nocturnas ordenadas por el grado de su influencia respectivamente.

Bibliografía

Trevor Hastie; Robert Tibshirani; Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*(second ed.). Springer. p. 128.

Garson, G. D. (2008). Discriminant function analysis.

Alpaydin, Ethem (2010). *Introduction to Machine learning*. MIT Press. p. 9

Fisher, R. A. (1938). "The Statistical Utilization of Multiple Measurements"

Páginas web:

https://rstudio.cloud/

https://www.rdocumentation.org/packages/MVN/versions/5.8/topics/mvn