UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Junio 2016

Comparación de modelos predictivos para propensión de compra y su caracterización mediante análisis de datos textuales

Comparison of predictive models for purchase propensity and their characterization by textual data analysis

Paula Jiménez Quintero[a]
paulajimenezquin@gmail.com

Daniel Cruz Castro[b]
lecruk@gmail.com

Resumen

Recientemente las aplicaciones de CRM (customer relationship management) y los planes de lealtad se han masificado, permitiendo a las empresas obtener mayor información de sus clientes y monitorear sus transacciones. Está información además de mostrar cómo está el cliente actualmente, puede permitir conocerlo a profundidad y con ello poder por ejemplo marcar a los clientes más valiosos, segmentar la población para crear productos llamativos, predecir algunos comportamientos futuros como la deserción, establecer la próxima mejor oferta, definir las oportunidades de venta cruzada y el perfil del cliente más propenso en una determinada campaña comercial. En este articulo se presenta el tema de la comparación de modelos predictivos de la posibilidad de compra para un individuo en las campañas comerciales y su caracterización mediante análisis de datos textuales. Se aplicaron métodos predictivos de Redes Neuronales, Maquinas de Soporte Vectorial, Regresión Logística binomial, modelo Bayesiano y Árboles de Clasificación. Se realiza la selección del mejor modelo mediante validación cruzada, su acuraccy, precisión y el área bajo la curva ROC. Adicionalmente, se realizó un análisis de datos textuales para relacionar las opiniones de los clientes y su posibilidad de compra.

***Palabras clave*:** Modelos predictivos, segmentación, datos textuales, estadística.

Abstract

Recently CRM (customer relationship management) applications and loyalty schemes have become massive, allowing companies to get more information about their customers and monitor their transactions. It is information in addition to showing how is the current customer, you can allow to know in depth and thus can for example mark the most valuable customers, segment the population to create eye-catching products, predict some future behaviors such as desertion, set the next best deal , defining cross-selling opportunities and customer profile more prone in a given marketing year. In this article the issue of comparison of predictive models of the possibility of purchase for an individual in commercial campaigns and their characterization is presented by analysis of textual data. predictive methods Neural Networks, Support Vector Machines, Binomial Logistic Regression, Bayesian Model and Classification Trees were applied. selecting the best model by cross-validation, your acuraccy, accuracy and area under the ROC curve is performed. In addition, a textual data analysis was performed to relate the opinions of customers and their ability to purchase.

***Keywords*:** Predictive modeling, segmentation, textual data, statistics.

[a]Estudiante Pregrado en Estadística, U. SantoTomas, sede Bogotá, Colombia.
[b]Docente de Planta. U. Santo Tomas, sede Bogotá, Colombia.

1. Introducción

En las áreas comerciales, de ventas y de mercadeo, es común la generación de campañas para la fidelización y profundización de clientes, las cuales, están enmarcadas en el propósito de aumentar el valor del cliente y la lealtad del consumidor. Algunos de los detalles que se deben tener en cuenta para asegurar el éxito de dichas campañas son: las características del producto a ofrecer, los beneficios de la campaña, el guión de contacto por Call Center y el perfil del consumidor. Todas estas características en su conjunto hacen una campaña efectiva. Algunos parámetros pueden ser optimizados utilizando diversos métodos estadísticos. Por ejemplo, es posible que existan clientes con preferencia de ciertos productos o promociones, o incluso clientes que no quieran volver a tener algún tipo de relación con la entidad por razones personales o simplemente por falta de interés hacia la marca.

Actualmente se logra tener mayor información de los clientes y de los monitoreos de sus transacciones, evidenciando el comportamiento del cliente a través del tiempo y con ello poder: establecer una próxima mejor oferta, identificar clientes valiosos, segmentarlos según sus necesidades y así crear productos llamativos que fidelicen clientes, al igual que predecir algunos comportamientos futuros y el perfil del cliente más propenso en una determinada campaña comercial.

En este trabajo se busca estimar si al contactar un cliente vía telefónica, este será propenso en una determinada campaña comercial, para ello se realiza la comparación de distintos modelos predictivos utilizando como variables explicativas lo sociodemográfico, el habito de pago, historial crediticio, productos adquiridos y la capacidad de endeudamiento del cliente; donde la variable respuesta es la efectividad de la compra. Se aplicarón métodos predictivos de redes neuronales, maquinas de soporte vectorial (kernel model), regresión logística binomial, modelo bayesiano (naive bayes) y árboles de Clasificación (CART, Dtree, random tree y random forest). La selección del mejor modelo se da mediante validación cruzada, con el acuraccy, precisión y el área bajo la curva ROC mas altos; y como complemento de la aplicacón del modelo predictivo se caracterizan y agrupan las opiniones que los clientes dieron al contactarlos por call-center.

2. Análisis predictivos

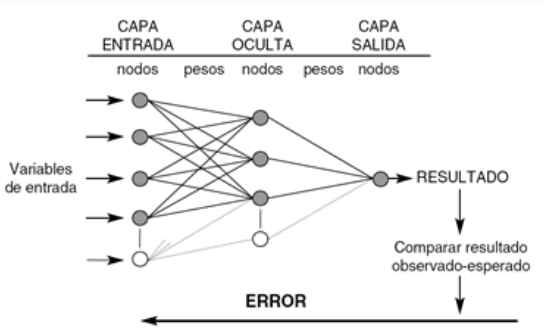
Los análisis predictivos nos ayudan a tener conjeturas de predicción, mostrando el ahora, y el hacia dónde dirigirse. Logrando analizar tendencias, patrones y relaciones en datos estructurados y no estructurados, aplicando dicha información de valor para prever sucesos futuros, y actuar para alcanzar los resultados deseados.

2.1. Redes neuronales artificiales (RNA)

El neurofísico Warren McCulloch y el matemático Walter Pitts en base a sus estudios del sistema nervioso biológico, proponen un modelo de redes neuronal implementada mediante circuitos eléctricos (McMulloch, 1943). Con Rosenblatt, F. (1957) y su desarrolló del Perceptrón, crea un modelo de red que posee la capacidad de generalización, el cual se utiliza hasta el día de hoy en diversas aplicaciones, generalmente en el reconocimiento de patrones. Las RNA se aplican a distintos campos del conocimiento como: la clasificación de patrones, el procesamiento de señales, el reconocimiento de escritura y habla, la visión artificial y robótica, entre otras.

Una RNA puede definirse como un sistema de procesamiento de información compuesto por un gran número de elementos de procesamiento (neuronas), conectadas entre sí a través de canales de comunicación (Reguero, 1995). Estas conexiones establecen una estructura jerárquica, basadas en el aprendizaje secuencial y sobre todo en la no linealidad, permitiendo aprender en contextos difíciles, en general de un tratamiento previo de los datos. La computación neuronal permite desarrollar sistemas que resuelvan problemas complejos como: el aprendizaje adaptativo, la autoorganización, la tolerancia a fallos, la ope-

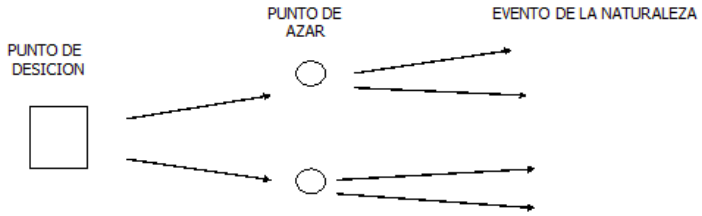
ración y su fácil inserción en la tecnología existente. Su principal inconveniente es que para el usuario puede ser una caja negra.



2.2. Árboles de decisión (CART)

Los arboles de clasificación y regresión (CART) fueron desarrollados por Breiman, Olshen, Freidman y Stone en el libro Classification and regression trees de Breiman (1984). La metodología CART busca dividir datos históricos en base a una prueba de atributo que optimice cierto criterio, en la que se representa gráficamente un proceso de decisión secuencial. En la construcción de una árbol de decisión se utiliza un cuadrado que representa el punto de decisión (situación planteada), se extienden lineas rectas que enlazan a círculos los cuales representa los puntos de azar (proceso aleatorio) es decir, las diferentes opciones a tomar y ligados a estos círculos lineas rectas que nos llevan a los eventos de la naturaleza, de esta forma el gráfico final es similar a un árbol, por esta razón su nombre.

Los puntos de decisión son planteados por el individuo y los puntos de azar no se encuentran bajo control de individuo no dependen de el, por tal razon van acompañados de su probabilidad de ocurrencia.



Se plantea de izquierda a derecha y se resuelve de derecha a izquierda hasta llegar al punto de decisión, es decir, que es posible simplificar el arbol hasta llegar a la respuesta deseada. Entre sus ventajas esta su robustez a outliers, la invarianza en la estructura de sus arboles de clasificación o de regresión a transformaciones monótonas de las variables independientes, y sobre todo, su fácil interpretación al tener resultados de forma visual.

Para especificar la prueba se debe tener en cuenta el tipo de atributo (si es nominal, ordinal o continuo) y del numero de divisiones (way split). La diferencia en la construcción de los arboles de clasificación y los árboles de regresión son el criterio de división de los nodos (way split) , es decir, la medida de impureza y la bondad de una división es diferente para los arboles de clasificación y de regresión. Existen dos maneras de dividir una es "binary split" la cual divide los valores en dos subconjuntos buscando la partición óptima, y la o multi-way split) que utiliza tantas particiones como valores distintos.

2.3. Maquinas de soporte vectorial (SMV)

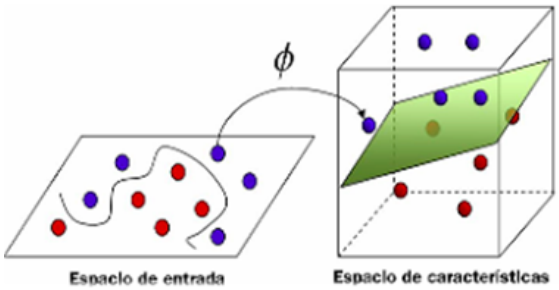
La teoría de las Máquinas de Soporte Vectorial (Support Vector Machines) es una técnica de clasificación, desarrollado principalmente por Vapnik, V. N (1992) y sus colaboradores, su formulación incorpora el principio de Minimización del Riesgo Estructural (SRM), el cual se ha demostrado que es superior al más tradicional principio de minimización del riesgo empírico (ERM). Las SMV proporcionan una habilidad mayor para generalizar, siendo una herramienta robusta para regresión y clasificación en dominios complejos y ruidosos; pueden ser usadas para extraer información relevante y para construir algoritmos de clasificación o de regresión rápidos para datos masivos.

Las SMV, aplicadas al problema de clasificacón, mapean los datos a un espacio de características altodimensional, donde se puede hallar más fácilmente un hiperplano de separación. Este mapeo puede ser llevado a cabo aplicando el *kernel*, el cual transforma implícitamente el espacio de entrada en un espacio de características de alta dimensión. El hiperplano de separación se calcula al maximizar la distancia de los patrones más cercanos, es decir la maximización del margen. Las SVM pueden ser definidas como un sistema para el entrenamiento eficiente de máquinas de aprendizaje lineal en un espacio de características inducido por un *kernel*, mientras respeta los principios de la teoría de la generalización y explota la teoría de la optimización.

$$(w.x) + b = 0; w \epsilon R^2, b \epsilon R$$

$$f(x) = sign(w.x + b) = sign(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x_j) + b)$$

Con función de densidad $f(x)$, donde $\alpha = (\alpha_1, ..., \alpha_l)$ es una vector de multiplicadores de Lagrange positivos asociados con las constantes, y utilizando un *kernel* polinomial de grado d tal que $K(x_i, x_j) = (1 + x_i * x_j)^d$.



2.4. Regresión logistica

Los modelos de regresión logística son modelos de regresión que permiten estudiar si una variable binomial depende, o no, de otra u otras variables (no necesariamente binomiales), dando a conocer la relación entre: Una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos valores (regresión logística multinomial). Y una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas, siendo la ecuación inicial del modelo de tipo exponencial, con transformación logarítmica (logit).

Las covariables cualitativas deben ser dicotómicas, tomando valores entre 0 y 1. Pero si la covariable cualitativa tuviera más de dos categorías, para su inclusión en el modelo se podría realizar una transformación de la misma en varias covariables cualitativas dicotómicas de diseño (variables dummy), tomando una de las categorías como de referencia, asi todas las categorías entrarían en el modelo de forma indi-

vidual. En general, si la covariable cualitativa posee n categorías, habrá que realizar $n-1$ covariables ficticias.

Los modelos de regresión logística por sus características permite:

- Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente, lo que lleva implícito también clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente (es decir, conocer la razon de odds para cada covariable).
- Clasificar individuos dentro de las categorías (1- 0) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables.

Logrando modelar la probabilidad de aparición de un suceso (generalmente dicotómico), la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías (politómico).

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

$$P(y=1|x) = \frac{1+e^{(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}{e^{(\beta_0+\beta_1 x)}} = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}$$

Con $P(y=1|x)$ es la probabilidad de que y tome el valor 1 (presencia de la característica estudiada), en presencia de las covariables x; donde x es un conjunto de n covariables $\{x1, x1, ..., xn\}$ que forman parte del modelo; β_0 es la constante del modelo o término independiente y β_i los coeficientes de las covariables.

3. Criterios de la estimación

3.1. Curva ROC

El Análisis ROC (Receiver operating characteristics) es una metodología desarrollada para analizar un sistema de decisión. Tradicionalmente se ha usado en ámbitos de detección de señales y, en las últimas décadas, se ha utilizado mucho en Medicina.

		Realidad y_0	
		1	**0**
Modelo y_e	**1**	VP	FP
	0	FN	VN

Tabla 1: Clasificación

- Sensibilidad = VP/(VP+FN)
- Especificidad = VN/(VN+FP)
- Area bajo la curva ROC construida para todos los posibles puntos de corte de π para clasificar los individuos.

La curva ROC trabaja con las nociones de Sensibilidad y Especificidad, la información contenida en la curva se puede interpretar de la siguiente manera:

- Si la prueba fuera perfecta, hay una región en la que cualquier punto de corte tiene sensibilidad y especifidad iguales a 1: la curva sólo tiene el punto (0,1). AUC=1

- Si la prueba fuera inútil: ambas nociones coinciden y la sensibilidad (verdaderos positivos) es igual a la proporción de falsos positivos, la curva sería la diagonal de (0,0) a (1,1). AUC=0.5

- Las pruebas habituales tienen curvas intermedias. AUC=0.8

Cálculo del área bajo la curva ROC: Se guardan los valores que predice el modelo (esperados) donde n_1 y n_0 son respectivamente el número esperado de "1" y "0", y se calcula la U de Mann-Whitney respecto a los esperados :

$$AUC = 1 - \frac{U}{n_1 n_0}$$

3.2. Validación cruzada

La idea es también muy simple, aunque computacionalmente algo laborioso. Se estima el error de predicción dividiendo al azar el conjunto de datos en varias partes. En cada paso una de las partes se convierte en una muestra de prueba que sirve para validar el modelo y las restantes partes constituyen lo que es llamado una muestra de entrenamiento que sirve para construir el modelo.

Si por ejemplo, se usasen 10 partes, se llamaría una "10 fold cross-validation" , por lo general se usa 1 parte y en ese caso es llamado el método "leave-one-out" (dejar uno afuera).

Sea $y_j^{[-i]}$ el valor predicho (la predicción que hacemos de la observación y_j) para la j-ésima observación usando una línea de regresión que ha sido estimada sin haber usado las observaciones de dicha parte.

El cálculo del error por validación cruzada usando p partes es:

$$CV_{(t)} = \frac{\sum_{i=1}^{p} \sum_{j=1}^{l} (y_j - \widehat{y}_j^{[-i]})^2}{p}$$

Entonces el mejor modelo (el mejor factor de regularización k por validación cruzada) es aquel k que tiene el error de validación cruzada promedio más pequeño:

$$k = \arg\, minCV$$

En principio, calcular $CV_{(k)}$ para un valor de k requeriría llevar a cabo l regresiones, excluyendo cada vez una observación distinta.

4. Metodologia

4.1. Selección de variables

Un modelo debe ser aquél más reducido que explique los datos (principio de parsimonia); ya que entre mayor sea el número de variables que tenga un modelo, mayor será su error estándar. Deben incluirse todas aquellas variables que se consideren importantes para el modelo, con independencia para demostrar o no su significación estadística. Por otro lado, no debería dejarse de incluir las variables que en un análisis previo demostraran una relación con la variable dependiente; para este trabajo se utilizaron las siguientes variables para explicar la efectividad de compra:

Figura 1: Variables del modelo

4.2. Detección de colinealidad entre variables independientes

En un modelado de regresión se busca que las variables independientes no posean ningún tipo de dependencia lineal entre ellas. Cuando una variable independiente posee alta correlación con otra ú otras ó puede ser explicada como una combinación lineal de alguna de ellas, se dice que el conjunto de datos presenta el fenómeno denominado multicolinealidad García (2006).

4.2.1. Análisis univariado

Valores perdidos

Se excluyeron las siguientes variables por el criterio de completitud(Figura 2): Mes Ultima Cancelación Negada, el estrato y todas las NumProdBank1, 2, 3 y 4 que son el número de productos que tienen los clientes en otros bancos.

	N										Percentiles			
	Válidos	Missing	Completitud	Media	Mediana	Desv. típ.	Varianza	Rango	Mínimo	Máximo	25	50	75	Uso
MES	700	0	1,00	1,02	1,00	,135	,018	1	1	2	1,00	1,00	1,00	Excluir
NOMBRE_CAMP	700	0	1,00	1,00	1,00	,000	,000	0	1	1	1,00	1,00	1,00	Excluir
REGIONAL	700	0	1,00	4,59	6,00	2,634	6,940	7	1	8	2,00	6,00	7,00	A Categorizar
ZONAL	700	0	1,00	17,90	17,00	9,434	88,994	39	1	40	11,00	17,00	27,00	A Categorizar
CALL	700	0	1,00	2,00	2,00	,038	,001	1	1	2	2,00	2,00	2,00	Excluir
NumProdCred	695	5	,99	15,55	9,00	22,651	513,067	262	1	263	4,00	9,00	19,00	A Categorizar
NumProdCapt	693	7	,99	3,95	3,00	2,763	7,633	23	1	24	2,00	3,00	5,00	A Categorizar
NumTarCred	518	182	,74	5,96	4,00	5,677	32,223	33	1	34	2,00	4,00	8,00	A Categorizar
TotalProd	700	0	1,00	24,20	17,00	26,112	681,837	282	0	282	9,00	17,00	31,00	A Categorizar
NumEntRela	700	0	1,00	6,74	6,00	4,196	17,607	20	0	20	3,00	6,00	9,00	A Categorizar
NumAperAnio	693	7	,99	,18	,00	,436	,190	2	0	2	,00	,00	,00	A Categorizar
NumCredAperAnio	695	5	,99	1,37	1,00	4,548	20,686	94	0	94	,00	1,00	2,00	A Categorizar
NumAper6mes	693	7	,99	,10	,00	,333	,111	2	0	2	,00	,00	,00	A Categorizar
NumCredAper6mes	695	5	,99	,80	,00	2,256	5,091	41	0	41	,00	,00	1,00	A Categorizar
NumAper3mes	693	7	,99	,05	,00	,240	,058	2	0	2	,00	,00	,00	A Categorizar
NumCredAper3mes	695	5	,99	,53	,00	1,269	1,610	19	0	19	,00	,00	1,00	A Categorizar
MesUltimaCancel	621	79	,89	19,51	11,00	25,617	656,253	210	0	210	3,50	11,00	26,00	A Categorizar
MesUltimaCancelNeg	87	613	,12	64,84	45,00	57,892	3351,532	256	0	256	18,00	45,00	94,00	Excluir
IngEstimados	684	16	,98	7478,07	7256,00	2461,353	6058260,776	13306	2660	15966	6094,25	7256,00	8966,00	A Categorizar
EgrEstimados	684	16	,98	22,17	14,00	28,288	800,208	291	0	291	8,00	14,00	25,00	A Categorizar
XEndsobreIng	684	16	,98	,03	,00	,380	,144	7	0	7	,00	,00	,00	A Categorizar
IndCartVcda	685	15	,98	,27	,00	2,616	6,845	50	0	50	,00	,00	,00	A Categorizar
SaldoMora	684	16	,98	,67	,00	12,558	157,692	324	0	324	,00	,00	,00	A Categorizar
SaldoTotal	684	16	,98	131,09	83,50	156,051	24351,824	1466	0	1466	41,00	83,50	159,75	A Categorizar
Cupo_sobre_Monto	684	16	,98	216,36	144,00	246,815	60917,527	2867	2	2869	79,00	144,00	263,75	A Categorizar
UtilizaTC12meses	431	269	,62	34,65	29,00	28,545	814,836	104	0	104	9,00	29,00	56,00	A Categorizar
XAmort	676	24	,97	39,94	38,85	21,833	476,661	92	0	92	24,90	38,85	54,14	A Categorizar
MoraMaxHist	694	6	,99	,66	,00	1,459	2,130	8	0	8	,00	,00	1,00	A Categorizar
NumProd	684	16	,98	2,84	2,00	2,552	6,515	37	1	38	1,00	2,00	3,00	A Categorizar
NumCons6mes	296	404	,42	1,77	1,00	1,316	1,732	13	0	13	1,00	1,00	2,00	A Categorizar
MesesUltConsulta	295	405	,42	3,05	3,00	1,709	2,922	6	0	6	2,00	3,00	5,00	A Categorizar
Edad	700	0	1,00	50,90	51,00	11,468	131,522	80	0	80	44,00	51,00	58,00	A Categorizar
Estrato	231	469	,33	3,16	3,00	1,477	2,182	6	0	6	3,00	3,00	4,00	Excluir
MesesPrimApertura	700	0	1,00	222,92	211,00	113,306	12838,304	656	0	656	135,00	211,00	291,00	A Categorizar
MesesProdReciente	696	4	,99	12,80	6,00	15,906	252,993	116	0	116	2,00	6,00	20,00	A Categorizar
EdadPromPortafolio	694	6	,99	67,63	61,00	32,001	1024,076	191	11	202	45,00	61,00	84,00	A Categorizar
NumEntVinc	435	265	,62	2,31	2,00	1,485	2,204	10	1	11	1,00	2,00	3,00	A Categorizar
NumEntVincCapta	388	312	,55	1,44	1,00	,666	,443	3	1	4	1,00	1,00	2,00	A Categorizar
NumProdBank1	243	457	,35	2,15	1,00	1,855	3,441	13	1	14	1,00	1,00	3,00	Excluir
NumProdBank2	91	609	,13	2,67	2,00	1,606	2,579	7	1	8	1,00	2,00	3,00	Excluir
NumProdBank3	126	574	,18	2,72	1,00	2,442	5,962	9	1	10	1,00	1,00	4,00	Excluir
NumProdBank4	41	659	,06	2,34	1,00	5,097	25,980	32	1	33	1,00	1,00	2,00	Excluir
NumProdReal	335	365	,48	3,06	2,00	2,578	6,646	17	1	18	1,00	2,00	4,00	A Categorizar

Figura 2: Análisis univariado

4.2.2. Análisis bivariado

Matriz de correlaciones

(Figura 3) Se observa la presencia de variables correlacionadas que son: ($NumAperAnio$) el Numeró de productos aperturados en el último año, ($NumAper6mes$) el numeró de productos aperturados en los últimos 6 meses, ($NumAper3mes$) el numeró de productos aperturados en los últimos 3 meses, ($EgrEstimados$) egresos estimados, ($\%EndsobreIng$) el porcentaje de endeudamiento sobre los ingresos, ($IndCartVcda$) el porcentaje de cartera vencida a la fecha, ($SaldoTotal$) en productos a la fecha (Total de productos), ($NumProd$) el número de productos al día a la fecha, ($NumCons6mes$) el numero de consultas en los últimos 6 meses, ($MesesUltConsulta$) meses desde la última consulta.

Figura 3: Matriz de correlaciones entre las posibles variables del modelo

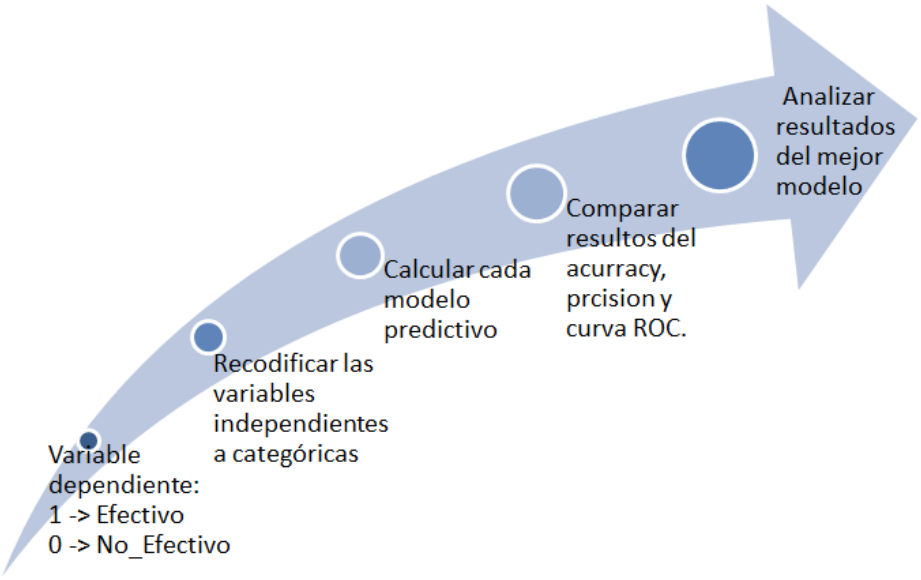


Figura 4: Procedimiento

4.3. Análisis de Componentes Principales (ACP)

Componente	Autovalores iniciales			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	4,921	28,949	28,949	4,339	25,525	25,525
2	3,542	20,835	49,784	3,056	17,974	43,499
3	1,878	11,045	60,828	2,625	15,441	58,940
4	1,737	10,216	71,044	2,058	12,103	71,044
5	1,055	6,208	77,252			
6	,927	5,454	82,706			
7	,859	5,056	87,762			
8	,575	3,384	91,145			
9	,454	2,670	93,815			
10	,307	1,807	95,623			
11	,255	1,501	97,124			
12	,181	1,063	98,187			
13	,159	,933	99,120			
14	,081	,478	99,598			
15	,045	,265	99,863			
16	,013	,076	99,939			
17	,010	,061	100,000			

Método de extracción: Análisis de Componentes principales.

a. Sólo aquellos casos para los que Efectivo = 1, serán utilizados en la fase de análisis.

Figura 5: ACP

Con los cuatro primeros factores se esta reteniendo el 71.04 % de la varianza.

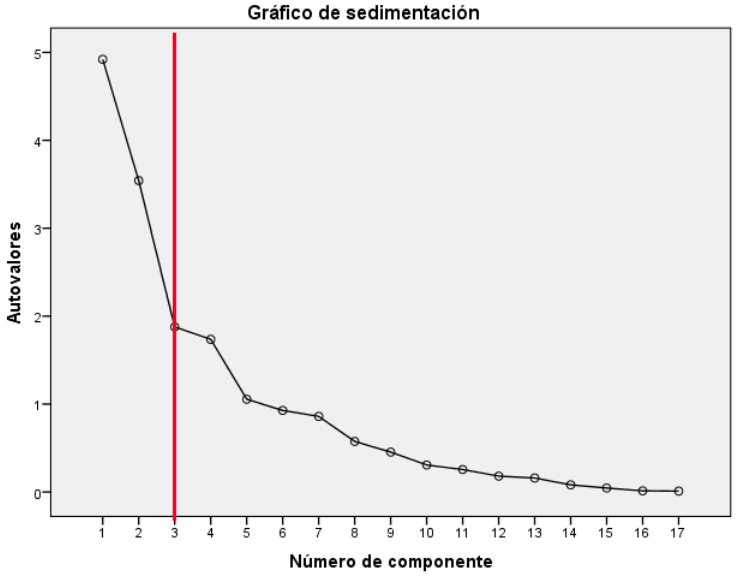


Figura 6: Autovalores

Matriz de componentes rotados[a,b]

	Componente			
	1	2	3	4
NumCredAper6mes	,955	,009	-,040	,029
NumCredAperAnio	,946	,006	-,050	-,059
NumCredAper3mes	,893	,083	-,022	,162
TotalProd	,753	-,070	,265	-,173
SaldoTotal	,675	-,008	,607	,093
IndCartVcda	,220	,137	-,176	-,151
revisados	,037	,955	,116	,058
Tramites_radicados	,022	,954	,113	,005
devueltos	,078	,732	,027	,026
tramites_decididos	-,120	,724	,263	,351
Monto_decididos	-,133	,297	,779	,180
Monto_negados	-,106	,029	,685	-,469
Cupo_sobre_Monto	,640	,002	,654	,019
IngEstimados	,452	,050	,649	,198
NumAper3mes	,024	,087	,295	,062
obg_desembolsadas	,000	,070	,027	,906
monto_desembolsos	,002	,195	,206	,850

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.[a,b]

a. La rotación ha convergido en 6 iteraciones.

b. Sólo aquellos casos para los que Efectivo = 1, serán utilizados en la fase de análisis.

Figura 7: Análisis de Componentes Principales

5. Resultados

5.1. Resultados de la aplicación y comparación de los modelos

Se aplicaron métodos de Redes Neuronales, Maquinas de Soporte Vectorial, Regresión Logística, Redes Bayesianas y Árboles de Clasificación:

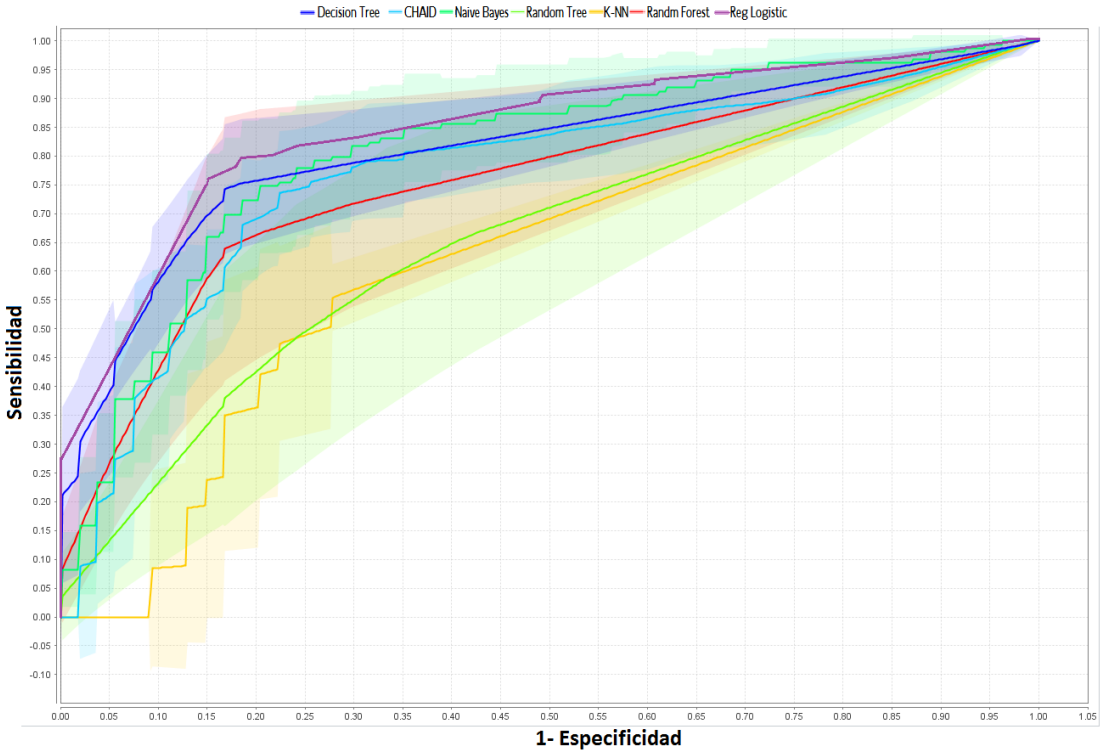


Figura 8: Comparación de la curva Roc de los modelos probados

Modelo		Accuracy	Presicion (Efectiva)	Area bajo la Curva ROC
Lazy Modeling	k-nn (vecinos mas cercanos)	75.29%	45.51%	0.500 +/- 0.000
Modelo Bayesiano	Naive Bayes	79.29%	53.21%	0.808 +/- 0.070
Arboles CART	CHAID	80.43%	59.32%	
	Dtree	82.86%	91.49%	0.814 +/- 0.058
	Random Forest	77.43%	100.00%	0.780 +/- 0.105
	Random Tree	78.71%	70.83%	0.611 +/- 0.141
Logit	Regresón Logistica	83.80%	71.10%	0.847
Redes Neuronales	NeuralNet	83.00%	72.73%	0.812 +/- 0.076
SMV	Kernerl Model	82.43%	61.11%	0.835 +/- 0.055

Figura 9: Comparación de los modelos probados

Al comparar el acuracy, precisión y curva bajo la area ROC de los metodos predictivos utilizados, se identifico que el modelo Logistico es quien mejor cumple los tres criterios por ello se elige como el mejor modelo para predecir la propensión de compra de un cliente.

5.2. Resultados del mejor modelo: Modelo Logistico binomial

Tabla de clasificación[a,b]

			Pronosticado		
			Efectivo		Porcentaje correcto
	Observado		no efectiva	efectiva	
Paso 0	Efectivo	no efectiva	540	0	100,0
		efectiva	159	0	,0
	Porcentaje global				77,3

a. En el modelo se incluye una constante.

b. El valor de corte es ,500
Para el analisis de regresion logistica el bloque 0 indica que hay un 77.3% de prob. de acierto en el resultado de la variable dependiente; Asumiendo que la efectividad en las ventas para todos los clientes fue NO EFECTIVA.

Figura 10: Clasificación 0

Tabla de clasificación[a]

			Pronosticado		
			Efectivo		Porcentaje correcto
	Observado		no efectiva	efectiva	
Paso 4	Efectivo	no efectiva	472	68	87,4
		efectiva	46	113	71,1
	Porcentaje global				83,7

a. El valor de corte es ,500
La tabla de clasificacion indica que hay un 83.7% de prob. de acierto en el resultado de la variable dependiente.

Figura 11: Tabla de clasificación final

La tabla de clasificación final del modelo logistico binomial indica que hay un 83.7 % de probabilidad de casos que acierta al predecir correctamente la variable dependiente (compra efectiva).

Pruebas omnibus sobre los coeficientes del modelo[a]

		Chi cuadrado	gl	Sig.
Paso 4	Paso	5,479	3	,140
	Bloque	238,306	8	,000
	Modelo	238,306	8	,000

a. Mirar resultados del modelo, estoy mejorando significativamente el modelo

b. X^2 es la puntuacion de eficiencia estadistica de ROA, indica que hay una mejora significativa en la prediccion de la prob de ocurrencia de las categorias chi:238.306, gl:8, p<0.001

Figura 12: Prueba de Omnibus sobre los coeficientes del modelo

Sobre la bondad del modelo la prueba de omnibus indica que el modelo efectivamente ayuda a explicar el evento en este caso la efectividad de compra del cliente.

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
4	511,299[a]	,604	,844

a. La estimación ha finalizado en el número de iteración 20 porque se han alcanzado las iteraciones máximas. No se puede encontrar una solución definitiva.

El R cuadrado de Nagelkerke indica que el modelo propuesto indica el 84,4% de la varianza de la variable dep. (,844)

Prueba de Hosmer y Lemeshow

Paso	Chi cuadrado	gl	Sig.[a]
4	7,360	6	,289

a. Dato adicional: la prueba de Hosmer y Lemeshow indica que la varianza es significativa con un 7.36% de la variable dependiente.

Figura 13: Resumen del modelo

El R^2 de Cox y Snell y el R^2 de Naegelkerke indican la parte de la varianza de la variable dependiente explicada por el modelo. Cuanto mas grande el R^2 mas explicativo es el modelo, es decir, que las variables independientes si estan explicando la efectividad de compra del cliente.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
								Inferior	Superior
Paso 4[b]	NumCredAper6mes_A			16,218	1	,000			
	NumCredAper6mes_A(1)	-1,803	,448	16,218	1	,000	,165	,069	,396
	NumCredAper3mes_A			21,874	2	,000			
	NumCredAper3mes_A(1)	21,780	17975,320	,000	1	,999	2876061441	,000	.
	NumCredAper3mes_A(2)	23,599	17975,320	,000	1	,999	1,774E+10	,000	.
	XAmort_A			5,508	3	,138			
	XAmort_A(1)	-,789	,629	1,574	1	,210	,454	,132	1,559
	XAmort_A(2)	-,804	,618	1,696	1	,193	,447	,133	1,501
	XAmort_A(3)	-1,344	,649	4,285	1	,038	,261	,073	,931
	NumEntVinc_A			25,367	2	,000			
	NumEntVinc_A(1)	-1,515	,306	24,462	1	,000	,220	,121	,401
	NumEntVinc_A(2)	-1,799	,614	8,580	1	,003	,165	,050	,551
	Constante	-8,517	5991,773	,000	1	,999	,000		

a. Se ha detenido un procedimiento por pasos ya que al eliminar la variable menos significativa se obtuvo un modelo previamente ajustado.

b. Variable(s) introducida(s) en el paso 4: XAmort_A.

c. <1 Si aumenta el valor de la indep., disminuye la var depen. (x=1, y=0)
>1 Si aumenta la indp tambien aumentara el de la depn. (x=1,Y=1)

Figura 14: Valores del modelo

El modelo es significativo, explica entre el 0.604 y el 0.844 de la variable dependiente (compra efectiva) y clasifica correctamente el 83.7 % de los casos, por tanto se acepta el modelo. En general es un buen modelo

$$P(y=1|x) = \frac{1}{e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}+1} =$$

$$\frac{1}{e^{-(8{,}51-1{,}8*NumCredAper6meses_A(1)+21{,}8*NumCredAper3meses_A(1)+23{,}6*NumCredAper3meses_A(2)}}$$

$$\frac{}{e^{-0{,}8*XAmort_A(1)-0{,}8*XAmort_A(2)-1{,}3*XAmort_A(3)-1{,}5*NumEntVinc_A(1)-1{,}8*NumEntVinc_A(2))}+1} \quad (1)$$

Las variables que mejor explican que la propensión de compra sea efectiva son: el numero de productos de crédito aperturados en los últimos 6 y 3 meses, el porcentaje de cartera vencida a la fecha y el NumEntVinc.

6. Caracterización mediante Análisis de Datos Textuales

El Análisis de Datos Textuales es una aplicación de los métodos de análisis multidimensionales exploratorios, esto se dio con las primeras aplicaciones realizadas por Benzécri (1973) quien desarrolló el análisis de correspondencias para discutir las tesis de Chomsky sobre la lengua. Ante la necesidad de postcodificar las preguntas abiertas con mayor rapidez y no de forma manual, se desarrollan métodos más automáticos: L. Lebart con el programa DtmVic y Bécue, M (1989) con un paquete informático para el tratamiento de datos textuales, el SPAD.T .

El Análisis de Datos Textuales consiste en aplicar estos métodos, en especial el análisis de correspondencias y la clasificación a tablas específicas, creadas a partir de los datos textuales. Estos métodos se completan con métodos propios del dominio textual como los glosarios de palabras, las concordancias y la selección del vocabulario más específico de cada texto, con lo que tenemos una herramienta comparativa de los mismos.

La información recogida muchas veces tiene errores al digitar o de ortografía por ello se debe realizar un primer paso de corrección ortográfica que se programa en R (mirar anexo), tomando como referencia el articulo de Bååth (2014) y de Norvig (2007), Enfocados en modelar las causas de los errores de ortografía directamente, y codificarlos en un algoritmo o un modelo de error. Utilizando la distancia Damerau-Levenshtein que ayuda a detectar errores de ortografía, esta distancia busca hallar el número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra. Sea: una inserción, eliminación, sustitución o transposición de dos caracteres; esta última cuenta como una sola operación de edición a cualquiera de las tres primeras, pero cuenta la transposición como dos operaciones de edición.

La distancia de Damerau-Levenshtein entre dos palabras a y b esta dada por $d_{a,b}(|a|,|b|)$, cuando:

$$d_{a,b}(i,j) = \begin{cases} \text{máx}(i,j) & \text{Si mín}(i,j) = 0, \\ \text{mín} \begin{cases} d_{a,b}(i-1,j)+1, \text{ borrado} \\ d_{a,b}(i,j-1)+1, \text{ inserción} \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)}, \text{ sustitución} \\ d_{a,b}(i-2,j-2)+1, \text{ transposición} \end{cases} & \text{Si } i,j > 1; \text{ con } a_i = b_{j-1} \text{ y } a_{i-1} = b_j \end{cases}$$

Ahora dada una palabra se elige la corrección ortográfica más probable para esa palabra (la corrección puede ser la misma palabra). Para ello se utilizan probabilidades que logran encontrar la corrección c, de entre todas las posibles correcciones que maximizan la probabilidad de c dada la palabra original w.

El modelo de error $P(c/w)$ esta dado por el Teorema de Bayes, equivalente a: $P(w/c) * P(c)/P(w)$, como $P(w)$ es el mismo para cada posible c, queda: $\mathbf{P(w/c) * P(c)}$. (Norvig (2007)). Donde $P(c)$ es la probabilidad de que una corrección propuesta c vale por sí mismo. Se conoce como el modelo de lenguaje, y $P(w/c)$ es la probabilidad de que dijo w pero el autor quería decir c.

Después del corrector ortográfico se prepara el corpus que es el texto completo que se somete a análisis. Los segmentos de las palabras se conocen como "formas gráficas", se define una variable léxica cuyas modalidades son formas gráficas del corpus y se construyen tablas de contingencia, estas tablas contienen los perfiles léxicos del individuo y su frecuencia. Cada presencia de una palabra en el corpus es una ocurrencia y el # de ocurrencias se denomina el tamaño del corpus.

7. Pre-tratamiento del corpus textual

El pre-tratamiento del corpus textual pretende reducir el numero de palabras sin perder información, por ello requiere una codificación ya que el lenguaje presenta normas gramaticales y sintácticas propias, para ello se utilizan una serie de herramientas computacionales que permiten tratar estadísticamente el

corpus facilitando los conteos y la construcción de tablas para analizar.

7.1. Reducción del Vocabulario

La reducción del vocabulario necesita seguir con el proceso de: corrección ortográfica de las palabras, dejar una sola palabra para todos sus sinónimos y crear palabras compuestas (varias palabras asociadas a un significado). En DtmVic el procedimiento Cortex es muy útil, ya que recuenta las palabras (formas gráficas) del corpus.

7.2. Lematización

La lematización del corpus es automática y se realizo a través de un programa llamado TreeTagge el cual hace un análisis morfológico de cada palabra dejandola con su categoría gramatical y la palabra lematizada. Al lematizar se reduce la variabilidad entre las respuestas y se pueden hacer análisis parcial del corpus, como de los sustantivos, adjetivos, verbos, etc.

7.3. El umbral de frecuencia

Para seleccionar formas y segmentos es preciso utilizar umbrales de frecuencia, es decir, definir la frecuencia mínima que debe tener la palabra, esto permite efectuar filtros sobre la información de base. Generalmente se utilizan umbrales entre 10 y 20, o superiores; pero esto depende de los objetivos del estudio a realizar. En este caso se tiene una muestra de 16.574 y se decide utilizar un umbral de 6 (figura 1).

Definición 7.1. *Las formas y segmentos se crean al fragmentar o reagrupar una muestra utilizando un proceso secuencial, que delimita grupos homogéneos según los criterios de una variable respuesta.*

En la figura 15. se muestran las palabras más asociadas (recuentos de las palabras seleccionadas para el análisis), con una frecuencia mínima (umbral) de 6.

8. Análisis multivariado sobre las palabras retenidas

El análisis de datos suele tratar grandes tablas de datos creadas a partir de variables nominales, ordinales o cuantitativas. Para aplicar métodos como el análisis de correspondencias y los métodos de clasificación, se construyen tablas de contingencia (tablas léxicas).

La tabla léxica contiene la frecuencia con la cual una forma gráfica es empleada por cada uno de los individuos. El análisis de correspondencias, aplicado esta tabla léxica, procede por comparación de las distribuciones de las formas en los individuos, es decir compara los perfiles léxicos de los individuos. También esta la tabla léxica agregada, que se da cuando existe una o varias particiones del corpus (como puede ser el estrato, la edad o el sexo del individuo, etc.), se puede construir la tabla de contingencia que contiene la frecuencia de cada forma en cada parte del corpus.

A estas tablas se les puede aplicar el análisis factorial de correspondencias simples (ACS) y los métodos de clasificación automática; revelando diferencias y características de los individuos asociado a un tipo de vocabulario y a la opinión respecto al crédito o producto que se le ofrece. La clasificación automática de los individuos completa y enriquece el análisis, identificando si un grupo de individuos dice lo mismo o no (Montenegro & Pardo 1998).

```
 Summary of results
-------------------
          total number of responses =     700
              total number of words =    3252
          number of distinct words  =     381
          percent.of distinct words =      11.7

selection of words
------------------
               frequency threshold  =       5
                        kept words  =    2785
                  distinct kept word =      88
```

num.	used words	freq.
1	acercar	183
2	agenda	21
3	agendar	11
4	ampliar	16
5	asistir	28
6	banco	39
7	brindar	19
8	buzon	8
9	buzón	12
10	citar	172
11	ciudad	8
12	cliente	146
13	con	139
14	confirmar	17
15	congestión	14
16	contactar	7
17	contestador	38
18	contestar	19
19	crédito	41
20	cualquier	9
21	dejar	28
22	del	20
23	dirigir	20
24	disponible	8
25	día	96
26	efectivo	174
27	encontrar	19
28	equivocar	13
29	estar	9
30	evidenciar	17
31	facilitar	7
32	fecha	14
33	gerente	7
34	haber	19
35	hablar	6
36	hora	29
37	indicar	137
38	información	39
39	informar	13
40	interesar	51
41	inválido	14
42	ir	20
43	junio	48
44	llamar	45

num.	used words	freq.
45	lleno	18
46	manifestar	24
47	maximo	14
48	mayo	10
49	mañana	11
50	mencionar	12
51	mensaje	36
52	momento	34
53	mucho	7
54	muy	6
55	ni	7
56	no	176
57	numerar	17
58	oficina	143
59	otro	6
60	pagar	6
61	para	13
62	poder	35
63	por	44
64	posible	6
65	progarmado	6
66	próximo	19
67	querer	7
68	recibir	8
69	renovar	13
70	requerir	11
71	saber	27
72	semana	18
73	ser	20
74	sin	12
75	sobre	12
76	solicitar	6
77	telefono	14
78	tener	20
79	tiempo	12
80	titular	9
81	transcurso	7
82	tratar	7
83	varios	6
84	venta	13
85	vivir	7
86	volver	6
87	voz	16
88	ya	84

Figura 15: Vocabulario del corpus en orden alfabético

8.1. Análisis de tablas léxicas

Al tener el corpus en textos se busca detectar por medio de un modelos estadísticos tales como el ACM y la segmentacion las formas características y poder crear clusters o grupos del corpus; (Montenegro

& Pardo 1998) consideran el texto como una posible muestra del corpus y se sitúa en conjunto de todas las muestras posibles de una forma que pueden ser obtenidas. La variacion total de las frecuencias entre formas se analiza con respecto a la totalidad de sus ocurrencias en el corpus. Con el propósito de establecer la probabilidad de las formas características.

El modelo de probabilidad se establece como: sea X la variable aleatoria definida como el numero de veces que la forma i (que tiene frecuencia total en el corpus $(f_i.)$) aparece en una muestra de tamaño $f._j$, por lo tanto la probabilidad de que la v.a X tome el valor x esta dada por:

$$prob(X = x) = \frac{\binom{f_i.}{x}\binom{f..-f_i.}{f._j-x}}{\binom{f..}{f._j}}$$

Donde X tiene una distribución hipergeometrica con parámetros $f_i., f._j, f..$; donde $f_i.$:es la frecuencia de la forma i en todo el corpus, $f._j$:es el tamaño de la parte y $f..$:es la longitud del corpus .

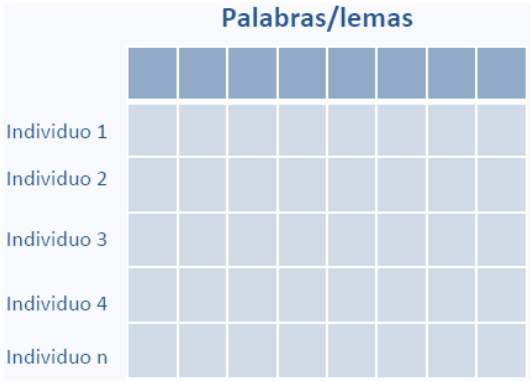


Figura 16: Balance general de los resultados

La probabilidad de encontrar por lo menos (f_{ij}) ocurrencias de la forma (palabra i) en el texto j (individuo) bajo las hipótesis de una extracción al azar sin reposición de $f._j$ (palabra de cada indidividuo) entre las $f..$ ocurrencias del corpus. Si $prob(f_{ij})$ es inferior que un cierto umbral, (normalmente 0.025) se declara la forma característica de especificidad positiva. Para facilitar se lectura se asocia a $prob(f_{ij})$ el valor de prueba (V. test) correspondiente a la distribución normal reducida. Un valor test se considera en general significativo si es mayor que 1.96.

Distribución de palabras y tabla léxica: La clasificación de los sujetos en base a sus respuestas se ve reflejado en las figuras 2, 3 y 4; la figura 2. muestra los ocho grupos dónde se indica la composición del número de individuos que la componen y el número de respuestas características de cada grupo.

number of text	identifier	number of individ.	number of responses
1	class 1 / 8	231	231
2	class 2 / 8	199	199
3	class 3 / 8	14	14
4	class 4 / 8	34	34
5	class 5 / 8	163	163
6	class 6 / 8	13	13
7	class 7 / 8	12	12
8	class 8 / 8	34	34
t o t a l		700	700

Figura 17: Balance general de los resultados

repartition of terms in texts/

number of text	identifier	number of words	/1000 of total	mean per response	number of (distinct) words	/1000 words of text	number of words kept
1 =	class 1 / 8	1489	457.9	6.4	72	48.4	1163
2 =	class 2 / 8	1020	313.7	5.1	54	52.9	892
3 =	class 3 / 8	84	25.8	6.0	6	71.4	84
4 =	class 4 / 8	127	39.1	3.7	12	94.5	119
5 =	class 5 / 8	455	139.9	2.8	11	24.2	451
6 =	class 6 / 8	26	8.0	2.0	3	115.4	26
7 =	class 7 / 8	17	5.2	1.4	4	235.3	16
8 =	class 8 / 8	34	10.5	1.0	1	29.4	34
	g l o b a l	3252	1000.0	4.6			2785

Figura 18: Tabla Resumen del número de palabras asociadas a cada grupo

table Words - Texts

		clas 1	clas 2	clas 3	clas 4	clas 5	clas 6	clas 7	clas 8
acercar	i	46.	137.	0.	0.	0.	0.	0.	0.
agenda	i	0.	0.	0.	0.	21.	0.	0.	0.
agendar	i	5.	5.	0.	0.	1.	0.	0.	0.
ampliar	i	16.	0.	0.	0.	0.	0.	0.	0.
asistir	i	1.	27.	0.	0.	0.	0.	0.	0.
banco	i	24.	15.	0.	0.	0.	0.	0.	0.
brindar	i	18.	1.	0.	0.	0.	0.	0.	0.
buzon	i	0.	0.	0.	0.	0.	8.	0.	0.
buzón	i	0.	0.	0.	7.	0.	5.	0.	0.
citar	i	15.	5.	0.	0.	152.	0.	0.	0.
ciudad	i	7.	1.	0.	0.	0.	0.	0.	0.
cliente	i	85.	61.	0.	0.	0.	0.	0.	0.
con	i	19.	3.	0.	0.	117.	0.	0.	0.
confirmar	i	0.	17.	0.	0.	0.	0.	0.	0.
congestión	i	0.	0.	14.	0.	0.	0.	0.	0.
contactar	i	2.	5.	0.	0.	0.	0.	0.	0.
contestador	i	2.	0.	0.	2.	0.	0.	0.	34.
contestar	i	19.	0.	0.	0.	0.	0.	0.	0.
crédito	i	41.	0.	0.	0.	0.	0.	0.	0.
cualquier	i	1.	8.	0.	0.	0.	0.	0.	0.
dejar	i	0.	0.	0.	28.	0.	0.	0.	0.
del	i	13.	7.	0.	0.	0.	0.	0.	0.
dirigir	i	1.	19.	0.	0.	0.	0.	0.	0.
disponible	i	5.	3.	0.	0.	0.	0.	0.	0.
día	i	10.	86.	0.	0.	0.	0.	0.	0.
efectivo	i	34.	1.	0.	0.	139.	0.	0.	0.
encontrar	i	14.	5.	0.	0.	0.	0.	0.	0.
equivocar	i	1.	0.	0.	0.	0.	0.	12.	0.
estar	i	9.	0.	0.	0.	0.	0.	0.	0.
evidenciar	i	10.	0.	0.	7.	0.	0.	0.	0.
facilitar	i	0.	7.	0.	0.	0.	0.	0.	0.
fecha	i	0.	14.	0.	0.	0.	0.	0.	0.
gerente	i	7.	0.	0.	0.	0.	0.	0.	0.
haber	i	16.	1.	0.	2.	0.	0.	0.	0.
hablar	i	6.	0.	0.	0.	0.	0.	0.	0.
hora	i	2.	27.	0.	0.	0.	0.	0.	0.
indicar	i	60.	77.	0.	0.	0.	0.	0.	0.
información	i	39.	0.	0.	0.	0.	0.	0.	0.
informar	i	10.	3.	0.	0.	0.	0.	0.	0.
interesar	i	51.	0.	0.	0.	0.	0.	0.	0.
inválido	i	0.	0.	14.	0.	0.	0.	0.	0.
ir	i	9.	10.	0.	0.	1.	0.	0.	0.
junio	i	2.	44.	0.	0.	2.	0.	0.	0.
llamar	i	37.	2.	0.	1.	5.	0.	0.	0.

		clas 1	clas 2	clas 3	clas 4	clas 5	clas 6	clas 7	clas 8
lleno	i	0.	0.	0.	5.	0.	13.	0.	0.
manifestar	i	21.	3.	0.	0.	0.	0.	0.	0.
maximo	i	0.	0.	14.	0.	0.	0.	0.	0.
mayo	i	0.	10.	0.	0.	0.	0.	0.	0.
mañana	i	0.	11.	0.	0.	0.	0.	0.	0.
mencionar	i	1.	11.	0.	0.	0.	0.	0.	0.
mensaje	i	2.	0.	0.	34.	0.	0.	0.	0.
momento	i	23.	11.	0.	0.	0.	0.	0.	0.
mucho	i	6.	1.	0.	0.	0.	0.	0.	0.
muy	i	5.	1.	0.	0.	0.	0.	0.	0.
ni	i	1.	6.	0.	0.	0.	0.	0.	0.
no	i	140.	31.	0.	4.	0.	0.	1.	0.
numerar	i	1.	0.	14.	0.	0.	0.	2.	0.
oficina	i	35.	108.	0.	0.	0.	0.	0.	0.
otro	i	5.	1.	0.	0.	0.	0.	0.	0.
pagar	i	6.	0.	0.	0.	0.	0.	0.	0.
para	i	6.	7.	0.	0.	0.	0.	0.	0.
poder	i	18.	15.	0.	1.	1.	0.	0.	0.
por	i	28.	1.	14.	0.	1.	0.	0.	0.
posible	i	4.	2.	0.	0.	0.	0.	0.	0.
progarmado	i	0.	6.	0.	0.	0.	0.	0.	0.
próximo	i	2.	17.	0.	0.	0.	0.	0.	0.
querer	i	7.	0.	0.	0.	0.	0.	0.	0.
recibir	i	8.	0.	0.	0.	0.	0.	0.	0.
renovar	i	13.	0.	0.	0.	0.	0.	0.	0.
requerir	i	11.	0.	0.	0.	0.	0.	0.	0.
saber	i	10.	17.	0.	0.	0.	0.	0.	0.
semana	i	3.	15.	0.	0.	0.	0.	0.	0.
ser	i	16.	3.	0.	0.	0.	0.	1.	0.
sin	i	1.	0.	0.	0.	11.	0.	0.	0.
sobre	i	12.	0.	0.	0.	0.	0.	0.	0.
solicitar	i	6.	0.	0.	0.	0.	0.	0.	0.
telefono	i	0.	0.	14.	0.	0.	0.	0.	0.
tener	i	18.	2.	0.	0.	0.	0.	0.	0.
tiempo	i	11.	1.	0.	0.	0.	0.	0.	0.
titular	i	3.	6.	0.	0.	0.	0.	0.	0.
transcurso	i	1.	6.	0.	0.	0.	0.	0.	0.
tratar	i	7.	0.	0.	0.	0.	0.	0.	0.
varios	i	6.	0.	0.	0.	0.	0.	0.	0.
venta	i	13.	0.	0.	0.	0.	0.	0.	0.
vivir	i	5.	2.	0.	0.	0.	0.	0.	0.
volver	i	5.	1.	0.	0.	0.	0.	0.	0.
voz	i	0.	0.	0.	16.	0.	0.	0.	0.
ya	i	66.	6.	0.	12.	0.	0.	0.	0.

Figura 19: Tabla de contingencia entre palabras y clases (cluster's)

Las figuras 17, 18 y 19, reflejan que el cluster (class 1) contiene a la mayoría de los individuos y de palabras distintas: 2.677, por esta razón se decide segmentar nuevamente a los sujetos de este grupo.

8.2. Respuestas características

Estas no son respuestas artificiales construidas a partir de las formas características, sino respuestas reales, escogidas según un criterio como el vocabulario representantes del texto.

Criterio del Ji-cuadrado: Un texto es un conjunto de vectores fila, si cada respuesta se considera como un vector fila cuyas componentes son las frecuencias de cada una de las palabras en esta respuesta. Según Bécue, M. "la respuesta más característica de un texto es la más próxima al perfil medio del texto que se obtiene haciendo la media de los perfiles de las respuestas del mismo"(Ob.cit.). Permitiendo calcular distancias entre respuestas y textos. La distancia seleccionada entre textos y respuestas es la distancia Ji-cuadrado. La respuesta mas característica será aquella mas cercana al perfil medio del texto, lo que se hace es ordenar las respuestas en orden decreciente de distancia al perfil medio; este criterio tiende a favorecer las respuestas largas.

Criterio del valor medio (v-test): Cuando se calcularon las formas características se ha asociado a cada par forma, texto un valor test, que puede ser positivo o negativo. Según la pertenencia de una respuesta a un texto, se le puede atribuir la media de los valores test correspondiente a las formas que componen la respuesta. La respuesta mas característica será aquella cuya media sea mas alta. Este criterio tiende a favorecer a las respuestas cortas.

Las respuestas características son respuestas originales pronunciadas por los individuos entrevistados. En general se extraen varias respuestas características para cada texto (10 a 20 según el caso). Una sola respuesta no resume en general todo el texto. Tampoco un único individuo es un buen representante de todo un grupo de individuos.

```
Selection of characteristic individuals or responses (criterion: frequency of words)
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
  criterion              characteristic response/individual
   of selection
-------------------------------------------------------------------------------
text number    1       class   1 /  8
----------------

  9.84 -  1  no interesar

  9.84 -  2  no interesar

  9.84 -  3  no interesar

  9.84 -  4  no interesar

  9.84 -  5  no interesar
-------------------------------------------------------------------------------
text number    2       class   2 /  8
----------------

 10.92 -  1  acercar día junio

 10.92 -  2  acercar día junio

 10.92 -  3  acercar día junio

 10.92 -  4  acercar día junio

 10.92 -  5  acercar día junio
-------------------------------------------------------------------------------
text number    3       class   3 /  8
----------------

  9.08 -  1  telefono inválido por numerar maximo congestión

  9.08 -  2  telefono inválido por numerar maximo congestión

  9.08 -  3  telefono inválido por numerar maximo congestión

  9.08 -  4  telefono inválido por numerar maximo congestión

  9.08 -  5  telefono inválido por numerar maximo congestión
-------------------------------------------------------------------------------
text number    4       class   4 /  8
----------------

 56.63 -  1  dejar mensaje

 56.63 -  2  dejar mensaje

 56.63 -  3  dejar mensaje

 56.63 -  4  dejar mensaje

 56.63 -  5  dejar mensaje
-------------------------------------------------------------------------------
=============================== ===============================
-------------------------------------------------------------------------------
text number    5       class   5 /  8
----------------

 19.89 -  1  efectivo con citar

 19.89 -  2  efectivo con citar

 19.89 -  3  efectivo con citar

 19.89 -  4  efectivo con citar

 19.89 -  5  efectivo con citar
-------------------------------------------------------------------------------
text number    6       class   6 /  8
----------------

 55.10 -  1  buzón lleno

 55.10 -  2  buzón lleno

 55.10 -  3  buzón lleno

 55.10 -  4  buzón lleno

 55.10 -  5  buzón lleno
-------------------------------------------------------------------------------
text number    7       class   7 /  8
----------------

 11.12 -  1  equivocar

 11.12 -  2  equivocar

 11.12 -  3  equivocar

 11.12 -  4  equivocar

 11.12 -  5  equivocar
-------------------------------------------------------------------------------
text number    8       class   8 /  8
----------------

 99.99 -  1  contestador

 99.99 -  2  contestador

 99.99 -  3  contestador

 99.99 -  4  contestador

 99.99 -  5  contestador
-------------------------------------------------------------------------------
====== ===============================
```

Figura 20: Caracterización de los grupos o clases

8.3. Resultados del análisis multivariado sobre las palabras retenidas

Los algoritmos de segmentación (clustering) pertenecen al grupo de métodos de minería de datos definido como no supervisados. El objetivo del clustering es entender la estructura macroscópica y relaciones entre objetos, considerando las maneras en las que estos son similares y/o diferentes. Estableciendo una medida de proximidad entre individuos y a partir de ahí, buscar los grupos de individuos más parecidos entre sí.

Las figuras 5 y 7 muestran las respuestas mas características para las catorce categorías creadas, logrando describir los sentimientos que los clientes tienen al ofrecerles un producto de la entidad financiera, tales como:

- "No interesar" (compra no efectiva, clientes que no están interesados)
- "Acercarse, día, Junio" (no agenda cita pero menciona que puede dirigirse en una fecha u hora especifica)
- "Teléfono invalido, linea telefónica congestionada" (no hubo comunicación)
- "Dejar mensaje" (no hubo comunicación)
- "Efectivo agenda cita"
- "Buzón lleno" (no hubo comunicación)
- "Número equivocado"
- "Contestador (buzón de voz)"
- "Requiere información para ampliar o renovar el crédito"

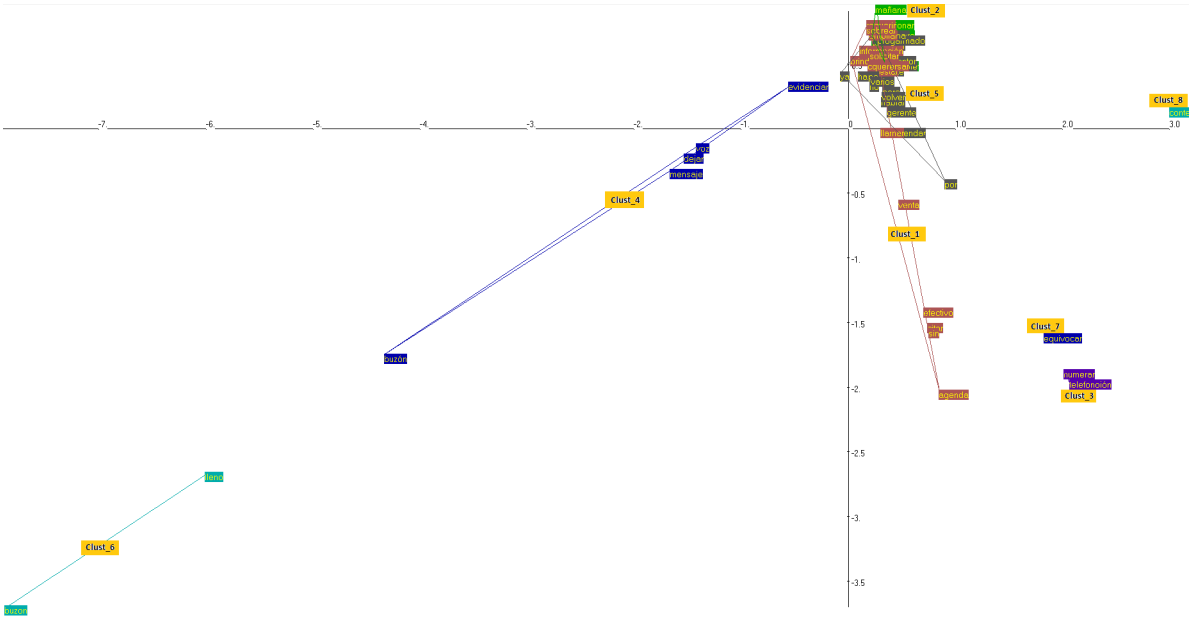


Figura 21: Cluster

9. Conclusiones

1. De los modelos que se analizaron, el mejor modelo fue el de regresión logística.
2. Las variables que mejor explican la efectividad de compra de un clientes son las de endeudamiento y hábitos de pago.
3. Con el análisis de cluster se identićaron clientes que no desean volver a tener algún tipo de relación con la entidad por razones personales o simplemente por falta de interés hacia la entidad.
4. Un reto es mejorar el guión de contacto Call Center, de tal manera que al contactar el cliente este se interese por conocer una nueva oferta y sea efectivo.

10. Motivación y trabajos futuros

- Crear una libreria en R de un corrector ortográfico y que también lematice para el idioma español.
- Construir un modelo bajo una metodología diferente y comparar resultados.

Agradecimientos

Agradezco a mi tutor de tesis por acompañarme, despejar dudas y estar pendiente del proceso del trabajo, agradezco a mi familia por el constante apoyo durante toda la carrera en especial a mis abuelos y a mi mama, muchas gracias por ser un ejemplo y estar presentes en mi vida.

Recibido: Si
Aceptado: 28 de Junio del 2016

Referencias

Bååth, R. (2014), 'Peter norvig's spell checker in two lines of base r'.
*http://www.sumsar.net/blog/2014/12/peter-norvigs-spell-checker-in-two-lines-of-r/

Bécue, M., Lebart, L. & Rajadell, N. (1992), 'El análisis estadístico de datos textuales, la lectura según los escolares de enseñanza primaria.', *Anuario de Psicología. Facultad de Psicología U.B* (55), 7–22.

Betancourt, G. (2005.), 'Las máquinas de soporte vectorial (svms)', *Scientia Et Technica* pp. 67–72.

de Estadística Aplicada. Universidad de Salamanca., G. (2006), 'Regresión y correlación. introducción a la estadística.'.
*http://biplot.usal.es/problemas/libro/index.html

Fagerland, M., Hosmer, D. & Bofin, A. (2008), 'Multinomial goodness-of-fit tests for logistic regression models.', *Statist Med.* **27**, 38–53.

Flórez, A., Gutiérrez, A. & Zea, J. (2015), 'Estimación por muestreo del índice de gini para las localidades de bogotá d.c. usando funciones en r', *Comunicaciones en Estadística* **8**(1), 59–79.

García, J. (2006), 'Efectos de la colinealidad en el modelado de la regresión y su solución', *Cultura Científica y Tecnológica* **16**, 23–34.

Hosmer, DW, J. & Lemeshow, S. (1980), 'Goodness-of-fit tests for the multiple logistic regression model.', *Communications in Statistics -Theory and Methods* **9**, 1043–1069.

Lebart, L., Salem, A. & Bécue, M. (2000), *Análisis Estadístico de Textos*, Vol. 2, ilustrada edn, Colección Educación. Serie Instrumentos.

Norvig, P. (2007), 'How to write a spelling corrector'.
*http://norvig.com/spell-correct.html

Pardo, C. E., Ortiz, J. E. & Cruz, D. L. (2012), Analisis de datos textuales con dtmvic, *in* 'XXII Simposio Internacional de Estadistica, Bucaramanga.'.

Vélez, M., Egurrola, J. & Barragán, F. (2012), 'Uso de la puntuación de propensión (propensity score) en estudios no experimentales'.

A. Códigos

```
library(stringdist)
CREA_total <-read.table("/Volumes/textual/CREA_total.TXT", header=T, quote="\")
CREA_total<-CREA_total[1:80000,] #Palabras mas frecuentes del vocabulario español
orden=as.character(CREA_total$Orden)
sorted_words <- orden
rm(list=c("orden","CREA_total"))
correct <-function( word ) {  c(sorted_words [adist (word , sorted_words)
 <= min(adist (word , sorted_words), 2)], word)[ 1 ]  }

correct1<-function(a){
b=c(b=strsplit(tolower(a)," "))
for(i in 1:length(b$b)){
b$b[i]=correct(b$b[i])
}
a=paste0(b$b,collapse=" ")
return(a)}

base<-read.csv("/Volumes/BASE_CALLCENTER.csv", sep=";")
obs=base[,13]  #Columna de la respuesta abierta

strip.text <-function(texto) {
  txt=as.matrix(unlist(texto))  #divide y genera un vector
  txt <-tolower(txt)   #convertir caracteres en mayusculas a minusculas
  txt <-gsub("[^a-z0-9]"," ",txt) #cambia caracteres alfanumericos por espacios
  txt <-gsub("[0-9]+"," ",txt)  # cambia los numeros por espacio
  txt <-gsub("[ ]+"," ",txt)   # cambiar mas de un espacio por 1 solo espacio
  return(txt)
}
obs<-strip.text(obs)

a=vector()
for(i in 1:30000){
  a[i]<-ifelse(obs[i]=="","",correct1(as.character(obs[i])))
}

write.table(a,paste("/Volumes/","muestra",length(a),".txt",sep="")) #Guardo muestra
```

23 de Junio de 2016

Señores
UNIVERSIDAD SANTO TOMÁS DE AQUINO
Carrera 9 No 51-11
Bogotá
Colombia

Asunto: Procedencia de los datos utilizados en el trabajo de grado.

Por medio de la presente se hace constar que los datos utilizados para el trabajo de grado del estudiante Paula Andrea Jiménez Quintero de la facultad de estadística identificado con número de cedula 1.013.645.805 expedida en Bogotá fueron simulados.

Gracias por la atención prestada.

Paula Jimenez & Daniel Cruz
Universidad Santo Tomás de Aquino
Facultad de Estadística