


Aplicación de estadística multivariada para caracterizar palabras claves de Adwords usadas por una compañía de ventas por internet

Br. José Antonio Castillo
Tutor: Dra. Mairene Colina

Universidad Central de Venezuela
Facultad de Ciencias
Escuela de Matemática

Caracas, Julio de 2018

Introduccion

El enorme impacto que ha tenido el Internet en nuestra sociedad ha generado un cambio radical en la comunicación hoy en día. La publicidad ha encontrado en la red un espacio para atraer a los usuarios a las diferentes marcas. Muchas empresas han visto la idea de promocionarse en la web como una manera más económica, fácil y eficaz de posicionarse en sus respectivos mercados. En este sentido **Google** lanza al mercado en el año 2000 **AdWords**, su sistema de publicidad en línea y principal fuente de ingresos. Este sistema permite orientar los anuncios de las empresas con palabras claves mediante un modelo de *pago por clic (PPC)*. Conocer que palabras claves tuvieron mas impacto en el anuncio, saber cual fue el rendimiento de éstas o que características presentan, son de un gran valor informativo para las empresas, pues con esto pueden tomar mejores decisiones con respecto a que palabras vale la pena apostar e invertir. Por esta razón en este trabajo se presenta un análisis multivariado de datos usando técnicas de agrupamiento y reducción de dimensionalidad derivadas del aprendizaje no supervisado de la minería de datos para poder caracterizar a los grupos resultantes mediante estadística descriptiva a un conjunto de palabras claves de una cierta campaña publicitaria de una empresa de artículos de sombreros.

Minería de datos y sus técnicas

Minería de datos y sus técnicas

La minería de datos es el proceso no trivial de identificar, a partir de datos, patrones válidos, novedosos, potencialmente útiles y comprensibles para poder generar conocimiento y realizar procesos que permitan un mejor acierto en las tomas de decisiones. La minería de datos está conformada por un arreglo de estrategias que nos permiten llevar a cabo estas acciones tales como lo son las tecnología de bases de datos, la visualización de datos, estadística, el aprendizaje automático, la inteligencia artificial, entre otras disciplinas.

Minería de datos y sus técnicas

Minería de datos y sus técnicas

La minería de datos es el proceso no trivial de identificar, a partir de datos, patrones válidos, novedosos, potencialmente útiles y comprensibles para poder generar conocimiento y realizar procesos que permitan un mejor acierto en las tomas de decisiones. La minería de datos está conformada por un arreglo de estrategias que nos permiten llevar a cabo estas acciones tales como lo son las tecnología de bases de datos, la visualización de datos, estadística, el aprendizaje automático, la inteligencia artificial, entre otras disciplinas.

Las técnicas más usadas en la minería de datos son las *técnicas predictivas* del *Aprendizaje Supervisado* y las *técnicas descriptivas* del *Aprendizaje No Supervisado*, los cuales vamos a ver con más detalle a continuación:

Aprendizaje Supervisado

El *Aprendizaje Supervisado* tiene como objetivo hacer predicciones a futuro basadas en comportamientos o características que se observan en los datos almacenados, buscando patrones en estos para luego poder ajustar un modelo que nos permita hacer inferencia y poder así predecir con la mayor precisión posible.

Aprendizaje Supervisado

El *Aprendizaje Supervisado* tiene como objetivo hacer predicciones a futuro basadas en comportamientos o características que se observan en los datos almacenados, buscando patrones en estos para luego poder ajustar un modelo que nos permita hacer inferencia y poder así predecir con la mayor precisión posible.

Algunas de las técnicas predictivas más importantes y usadas son las siguientes:

Técnicas predictivas

- Máquinas de Soporte Vectorial.
- Arboles de clasificación o regresión.
- Redes Neuronales.
- K-vecinos más cercanos (KNN).
- Redes Bayesianas.
- Series Temporales.

Aprendizaje No Supervisado

El *Aprendizaje No Supervisado* tiene como objetivo explorar y obtener información sobre las posibles relaciones entre los datos para así poder armar una estructura que nos permita etiquetar estas asociaciones entre los datos para luego obtener conocimiento de estos.

Aprendizaje No Supervisado

El *Aprendizaje No Supervisado* tiene como objetivo explorar y obtener información sobre las posibles relaciones entre los datos para así poder armar una estructura que nos permita etiquetar estas asociaciones entre los datos para luego obtener conocimiento de estos.

Algunas de las técnicas descriptivas más importantes y usadas son las siguientes:

Técnicas descriptivas

- Análisis de agrupamiento.
- Detección de anomalías.
- Reglas de asociación.
- Análisis de Componentes Principales (ACP).

Técnicas descriptivas

- Análisis de agrupamiento.
- Detección de anomalías.
- Reglas de asociación.
- Análisis de Componentes Principales (ACP).

El presente trabajo se enfocará usando 2 técnicas del *Aprendizaje No Supervisado* como lo son el análisis de agrupamiento y el análisis de componentes principales a un conjunto de datos del sistema **AdWords** de Google, el cual vamos a ver con más detalles en la próxima sección.

Google Adwords

Google Adwrods

Google AdWords es el sistema de publicidad en línea de Google con la cual las empresas de cualquier tipo pueden promocionarse colocando sus anuncios en las páginas de busqueda de Google, así como tambien en varios sitios web que pertenezcan a la red publicitaria de Google. Este sistema permite orientar los anuncios de las empresas mediante palabras claves o *keywords* que no son más que un conjunto de palabras que usan los anunciantes para mostrar sus anuncios una vez que las mismas son colocadas en el buscador de Google.

Google Adwords

Google Adwrods

Google AdWords es el sistema de publicidad en línea de Google con la cual las empresas de cualquier tipo pueden promocionarse colocando sus anuncios en las páginas de busqueda de Google, así como tambien en varios sitios web que pertenezcan a la red publicitaria de Google. Este sistema permite orientar los anuncios de las empresas mediante palabras claves o *keywords* que no son más que un conjunto de palabras que usan los anunciantes para mostrar sus anuncios una vez que las mismas son colocadas en el buscador de Google.

El sistema **AdWords** funciona bajo un modelo de subasta donde los anunciantes compiten emitiendo pujas para posicionar sus anuncios en los primeros resultados de Google mediante las palabras claves, para así poder atraer a potenciales clientes hacia la pagina web de el anunciante. Este sistema también usa el modelo de *pago por clic (PPC)* lo cual resulta ser bastante atractivo a las empresas y anunciantes pues estos solo pagarán cuando un usuario haya hecho clic a su anuncio siendo así una manera efectiva y económica de pagar por publicidad.

Google Adwords

Algunas características relevantes de **Google AdWords** son:

- No hay requisitos de inversión mínima.
- Establecer y controlar su presupuesto.
- Medir el impacto de su anuncio.
- Las cuentas de AdWords se administran en línea.
- Modificar el texto de sus anuncios.
- Elegir dónde aparecerá el anuncio.

Google Adwords

Cada palabra clave o *keyword* tiene asociada variables que describen el rendimiento que desempeñaron en un período de tiempo, algunas de estas son:

- **Clics.**
- **Impresiones.**
- **Tasas de clics (CTR).**
- **Campaña.**
- **Nivel de calidad.**
- **Conversión.**
- **Costo total.**
- **Costo promedio.**
- **Oferta máxima de costo por clic.**
- **Posición promedio.**

Técnicas del análisis multivariado de datos No Supervisado

Análisis de agrupamiento

El análisis de agrupamiento consiste en un conjunto de métodos númericos cuyo objetivo en común es encontrar o descrubrir grupos o *'clusters'* de un conjunto de datos que sean homogéneos entre sí y diferentes de los otros grupos. El propósito de dicho análisis es identificar patrones en los grupos resultantes.

Técnicas del análisis multivariado de datos No Supervisado

Los métodos de agrupamiento principales son:

Técnicas del análisis multivariado de datos No Supervisado

Los métodos de agrupamiento principales son:

- Basados en particiones: Son algoritmos que dividen los datos en k clusters, donde el número entero k es especificado por el usuario.

Técnicas del análisis multivariado de datos No Supervisado

Los métodos de agrupamiento principales son:

- Basados en particiones: Son algoritmos que dividen los datos en k clusters, donde el número entero k es especificado por el usuario.
- Agrupamiento jerárquico: Este método a su vez se subdivide en dos métodos: Los algoritmos *aglomerativos* y los *divisivos*. Los algortimos aglomerativos consisten en que cada observación forma su propio cluster, luego los clusters con mayor similitud se agrupan formando un nuevo cluster, este proceso se repite hasta que solo quede un solo cluster el cual contiene a todo el conjunto de datos. Por otro lado los algoritmos divisivos empiezan con el conjunto de datos como un solo cluster y a partir de ahi este se divide en varios clusters hasta que cada observación se convierta en su propio cluster.

Medidas de distancia o similitud

Es de central importancia conocer que tan ’cerca’ o ’lejos’ están las observaciones entre sí para poder agruparlos, conocer esta medida de cercanía se le conoce como similitud (disimilitud) ó distancia. Dos observaciones son cercanas si tienen una disimilitud o distancia corta o una similitud grande.

Medidas de distancia o similitud

Es de central importancia conocer que tan ’cerca’ o ’lejos’ están las observaciones entre sí para poder agruparlos, conocer esta medida de cercanía se le conoce como similitud (disimilitud) ó distancia. Dos observaciones son cercanas si tienen una disimilitud o distancia corta o una similitud grande.

La elección de la distancia es de gran peso para las agrupaciones pues dependiendo de la similitud entre las observaciones esta influirá en la conformación de los grupos. Las medidas de disimilitud están divididas entre *las medidas de distancia* y *las medidas de distancia basadas en la correlación.*

Algunas de las medidas de distancia más usadas son:

Algunas de las medidas de distancia más usadas son:

1. *Distancia Euclídea:*

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

2. *Distancia Manhattan:*

$$d(x, y) = \sum_{i=1}^{n}|x_i - y_i|.$$

Donde $x = (x_1, ..., x_n)$ e $y = (y_1, ..., y_n)$ son dos observaciones o vectores de tamaño n.

Asociadas a las medidas de distancia basadas en correlación tenemos:

Asociadas a las medidas de distancia basadas en correlación tenemos:

1. *Distancia de correlación Pearson:*

$$d(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

2. *Distancia de correlación del Coseno Eisen:*

$$d(x,y) = 1 - \frac{\left|\sum_{i=1}^{n} x_i y_i\right|}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}.$$

Con $x = (x_1, ..., x_n)$ e $y = (y_1, ..., y_n)$ dos observaciones.

Estandarización

Cuando estandarizamos las variables, para cada observación x_i tenemos:

$$\frac{x_i - center(x)}{scale(x)}.$$

Donde *center(x)* es una medida de tendencia central de las variables, donde la más común es la media y *scale(x)* la medida de dispersión, donde la más común es la desviación estándar.

Una vez escogidas la manera de estandarizar y la medida de distancia, se procede a elegir el algoritmo a usar para el agrupamiento de los datos, para este trabajo presentaremos el algoritmo PAM y haremos mención de otros algoritmos.

Algortimos basados en particiones

K-medoids ó PAM

El objetivo de PAM es encontrar un subconjunto de observaciones (*medoids*) $\{m_1, ..., m_k\} \subset \{x_1, ..x_m\}$ del conjunto de datos tal que la suma total de la distancias de todas las observaciones a su *medoid* más cercano sea mínima, es decir, encontrar $\{m_1, ..., m_k\}$ tal que la expresión dada por:

$$\sum_{i=1}^{k} \sum_{x_j \in G_i} d(x_j, m_i).$$

sea mínima, donde $G_i = \{x_j : d(x_j, m_i) = \min_{i=1,..,k} d(x_j, m_i)\}$.

Otros algoritmos de agrupaciones son:

Otros algoritmos de agrupaciones son:

- K-medias
- AGNES (Aglomeración por anidación)
- DIANA (Análisis divisivo)
- MONA (Análisis monotético)

Validación

Método de la silueta

El análisis de la Silueta mide que tan bién se agrupó una observación comparando su similitud con el resto de observaciones de su cluster frente a las de los otros clusters. Para conocer esta medida calculamos el índice de silueta $S_{(x_j)}$ para cada observación $x_j \in \{x_1, .., x_m\}$, donde $\{x_1, .., x_m\}$ son todas las observaciones del conjunto de datos. El valor $S_{(x_j)}$ esta entre los valores -1 y 1, siendo el valor 1 un indicativo que la observación se ha asignado al grupo correcto y -1 como una mala asignación.

El método de la silueta consiste en calcular el promedio de todos los índices de silueta $S_{(x_j)}$ de las observaciones $\{x_1, .., x_m\}$, es decir, calcular:

$$\bar{S} = \frac{1}{m} \sum_{i=1}^{m} S_{(x_j)}.$$

Validación

Obtener un promedio $\bar{S}$ lo más grande posible nos indica una mejor calidad del agrupamiento formado, podemos interpretar el valor $\bar{S}$ de la siguiente forma:

Valor $\bar{S}$	**Interpretación**
0.71-1	Fuerte estructura
0.51-0.70	Estructura razonable
0.26-0.50	Estructura débil o superficial
<0.25	Estructura no encontrada

Table: Interpretación del valor $\bar{S}$

Análisis de componentes principales

El análisis de componentes principales es un método de reducción de dimensionalidad que consiste en extraer la información de las variables (usualmente correlacionadas entre sí) de un conjunto multivariado de datos y expresar esta información en un conjunto nuevo de variables no correlacionadas conocidas como *componentes principales*. Estas nuevas variables corresponden a una combinación lineal de las variables originales. La cantidad de componentes principales es menor o igual a la cantidad de las variables originales.

Análisis de componentes principales

El análisis de componentes principales es un método de reducción de dimensionalidad que consiste en extraer la información de las variables (usualmente correlacionadas entre sí) de un conjunto multivariado de datos y expresar esta información en un conjunto nuevo de variables no correlacionadas conocidas como *componentes principales*. Estas nuevas variables corresponden a una combinación lineal de las variables originales. La cantidad de componentes principales es menor o igual a la cantidad de las variables originales.

El objetivo del ACP es representar en las primeras dos ó tres componentes principales la mayor cantidad de variabilidad de todo el conjunto de datos y poder así visualizarlas gráficamente obteniendo de esta manera una mínima perdida de información. El resto de las componentes principales representan el resto de la variabilidad de los datos.

Para lograr este objetivo se realiza el siguiente proceso:

Para cada componente principal $y_j = \alpha_j x^T$, nos sujetamos a las condiciones $\alpha_j \alpha_j^T = 1$ y $\alpha_j \alpha_i^T = 0 (i \neq j)$. La aplicación de la técnica de los multiplicadores de Lagrange demuestra que el vector de pesos α_j de la j-ésima componente principal, es el autovector asociado al j-ésimo autovalor (λ_j) más grande de la matriz de correlación\covarianza y además que la varianza de la j-ésima componente principal (Var[y_j]) está dada por λ_j.

Resultados

Se realizó un análisis de componentes principales al conjunto de datos debido a la alta dimensionalidad de variables que estos presentan, para esto se usaron solamente las variables de tipo numérica.

Resultados

Se realizó un análisis de componentes principales al conjunto de datos debido a la alta dimensionalidad de variables que estos presentan, para esto se usaron solamente las variables de tipo numérica.

Las variables usadas fueron:

- Clics.
- Tasas de clic (CTR).
- Oferta máxima de costo por clic.
- Impresiones.
- Costo total.
- Costo promedio.
- Posición promedio.
- Calidad del anuncio.
- Conversiones.

La varianza asociada a cada componente principal, su porcentaje de varianza explicada (P.V.E) y el porcentaje de la varianza explicada acumulada (P.V.E.A) se muestran en la siguiente tabla. En ésta observamos que las dos primeras componentes principales explican un 57.1% de la variabilidad del conjunto de datos, para tres componentes se tiene un 68.7% y así sucesivamente. En general la elección de la cantidad de componentes principales a ser utilizadas se hace considerando el P.V.E.A más significativo, para el presente trabajo se consideró el uso de dos y cuatro componentes principales las cuales explican el 57% y 79.5% de variabilidad de los datos respectivamente.

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8
Varianza	3.62	1.51	1.04	0.97	0.82	0.50	0.24	0.21
P.V.E	40.2%	16.8%	11.6%	10.8%	9.1%	5.6%	2.7%	2.3%
P.V.E.A	40.2%	57.1%	68.7%	79.5%	88.7%	94.3%	97.0%	99.4%

	Comp 9
Varianza	0.04
P.V.E	0.5%
P.V.E.A	100%

Table: Varianza de las componentes.

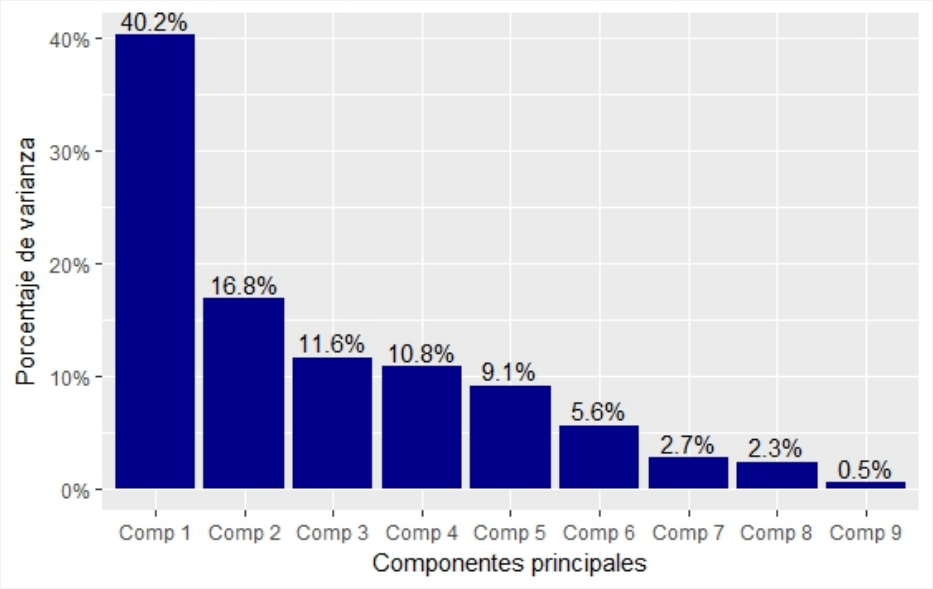

Figure: Porcentaje de varianza explicada por cada componente principal.

La proyección de los datos estandarizados (con medida de tendencia central la media y medida de dispersión la desviación estándar) en las dos primeras componentes principales se muestran en la siguiente gráfica.

La proyección de los datos estandarizados (con medida de tendencia central la media y medida de dispersión la desviación estándar) en las dos primeras componentes principales se muestran en la siguiente gráfica.

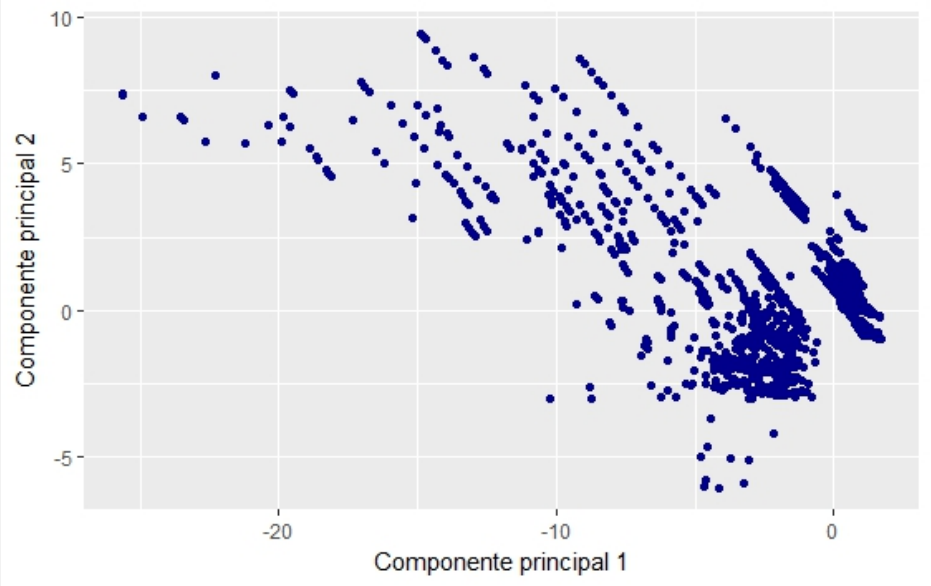

Figure: Proyección de los datos en las dos primeras componentes principales.

Agrupaciones

El algoritmo que tuvo mejor desempeño fue PAM, las variables usadas para las agrupaciones fueron las que más información aportaban en las primeras dos y cuatro componentes principales respectivamente, éstas son:

Agrupaciones

El algoritmo que tuvo mejor desempeño fue PAM, las variables usadas para las agrupaciones fueron las que más información aportaban en las primeras dos y cuatro componentes principales respectivamente, éstas son:

- Clics,
- Tasas de clics,

para las dos primeras componentes principales y para las primeras cuatro componentes:

Agrupaciones

El algoritmo que tuvo mejor desempeño fue PAM, las variables usadas para las agrupaciones fueron las que más información aportaban en las primeras dos y cuatro componentes principales respectivamente, éstas son:

- Clics,
- Tasas de clics,

para las dos primeras componentes principales y para las primeras cuatro componentes:

- Clics,
- Tasa de clics,
- Posición promedio,
- Conversiones.

Agrupación 1: Clics, Tasa de clics

La agrupación se realizó con los datos estandarizados con medida de tendencia central la media y medida de dispersión la desviación estándar, la medida de distancia usada fue la distancia Euclídea. Para esta agrupación se consideró $K = 4$.

Agrupación 1: Clics, Tasa de clics

La agrupación se realizó con los datos estandarizados con medida de tendencia central la media y medida de dispersión la desviación estándar, la medida de distancia usada fue la distancia Euclídea. Para esta agrupación se consideró $K = 4$.

La distribución de las palabras claves en los 4 grupos formados por el algoritmo PAM quedó de la siguiente manera.

Grupo 1	Grupo 2	Grupo 3	Grupo 4
13.617	1.054	157	1.478

Table: Distribución de las palabras claves.

La proyección de los datos en las dos primeras componentes principales segmentadas por estos cuatro grupos la observamos en la siguiente figura.

La proyección de los datos en las dos primeras componentes principales segmentadas por estos cuatro grupos la observamos en la siguiente figura.

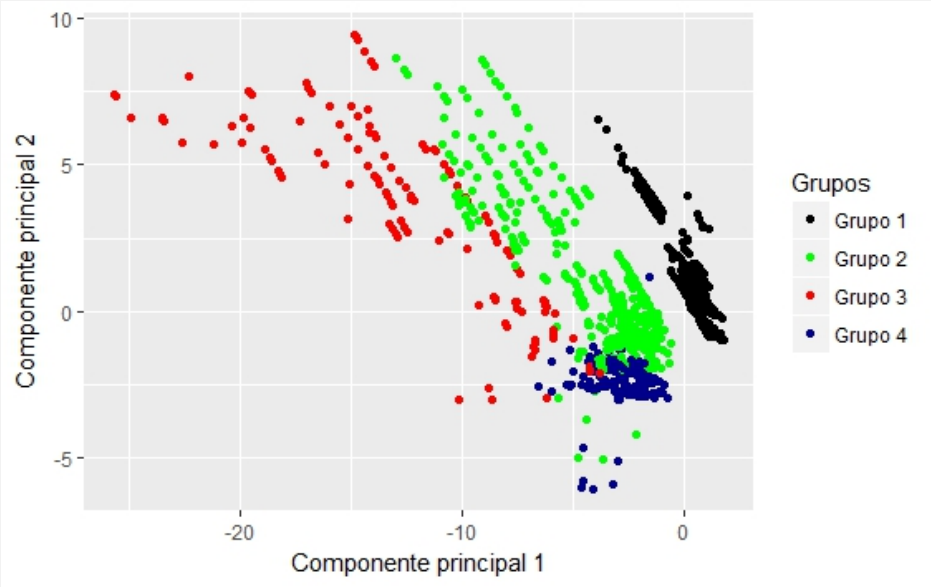

Figure: Proyección de los datos originales segmentados por los cuatro grupos.

La proyección de los datos en las dos primeras componentes principales segmentadas por estos cuatro grupos la observamos en la siguiente figura.

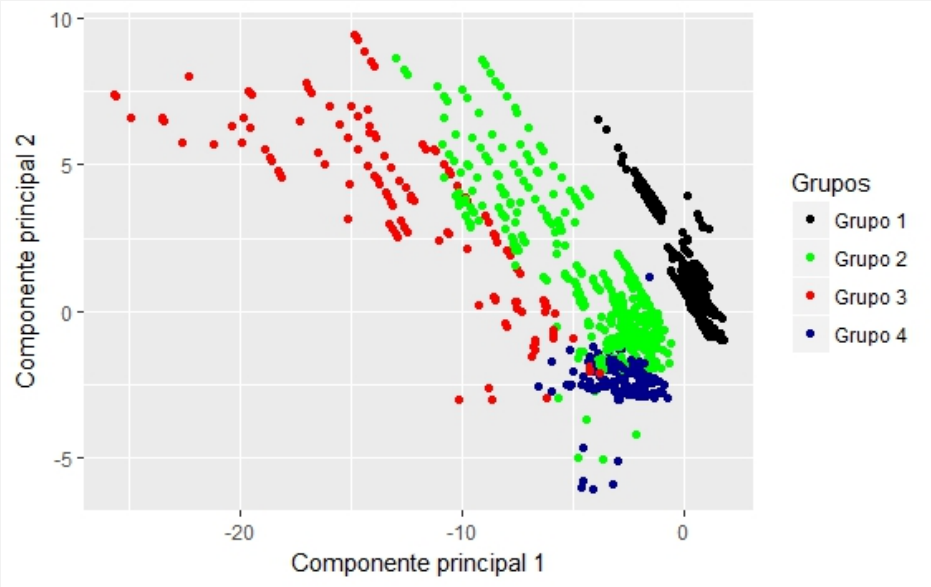

Figure: Proyección de los datos originales segmentados por los cuatro grupos.

Veamos las características de las palabras claves en los cuatro grupos formados.

Grupo 1:

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total
Min	0.13	1	0	0	0
Mediana	1	1	0	0	0
Media	1.07	1.75	0	0	0
Max	5	60	0	0	0
	Costo promedio	Posicion promedio	Calidad anuncio	conversiones	
Min	0	1	0	0	
Mediana	0	1	8	0	
Media	0	1.43	8.69	0	
Max	0	7	10	0	

Table: Resumen del grupo 1.

Grupo 2:

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total
Min	0.34	2	1	0.92	0.01
Mediana	1.14	4	1	33.33	0.92
Media	1.56	10.02	1.1	29.19	1.05
Max	5	117	2	50	4.46
	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones	
Min	0.01	1	5	0	
Mediana	0.88	1	8	0	
Media	0.92	1.31	8.76	0.02	
Max	2.54	3.5	10	1	

Table: Resumen del grupo 2.

Grupo 3:

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total
Min	0.59	3	3	2.27	0.42
Mediana	5	34	3	10.81	4.4
Media	3.41	43.41	3.99	21.55	5.43
Max	5	132	9	100	13.94
	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones	
Min	0.14	1	7	0	
Mediana	1.39	1	8	0	
Media	1.32	1.13	8.03	0.01	
Max	2.38	2.3	10	1	

Table: Resumen del grupo 3.

Grupo 4:

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total
Min	0.13	1	1	66.67	0.01
Mediana	1	1	1	100	0.66
Media	1.12	1.05	1.05	99.8	0.73
Max	2.8	3	2	100	3.87
	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones	
Min	0.01	1	0	0	
Mediana	0.64	1	10	0	
Media	0.67	1.17	9.33	0.04	
Max	2.16	4	10	1	

Table: Resumen del grupo 4.

Validación

Se usó el método de la silueta (Silhouette method) para conocer la calidad del agrupamiento formado, el valor $\bar{S}$ se ilustra en la siguiente gráfica.

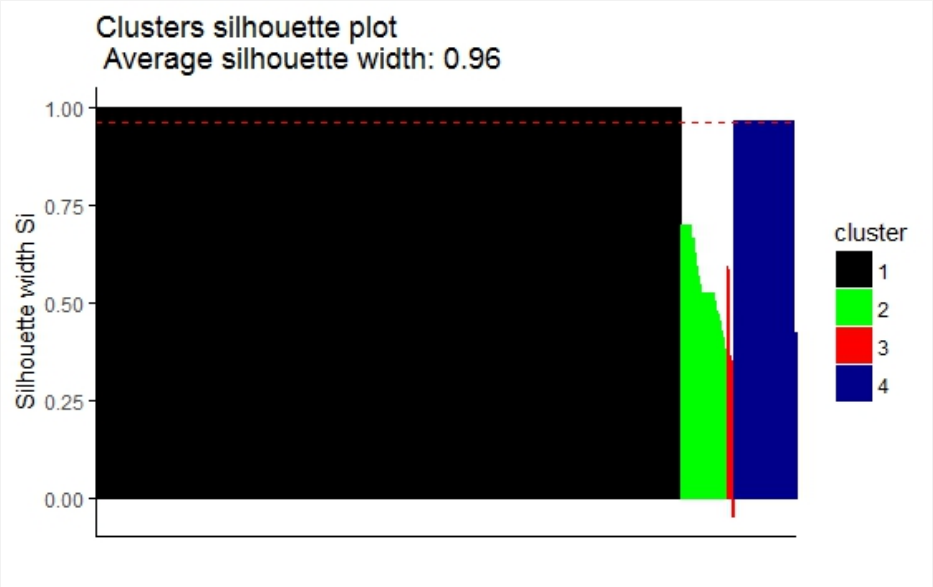

Figure: Promedio de los índices de silueta.

Agrupación 2: Clics, Tasa de clics, Posición promedio, Conversiones

Al igual que la Agrupación 1, se usó como medida de tendencia central la media, medida de dispersión la desviación estándar y medida de disimilitud la distancia Euclídea. Para ésta agrupación se consideró $K = 6$.

Agrupación 2: Clics, Tasa de clics, Posición promedio, Conversiones

Al igual que la Agrupación 1, se usó como medida de tendencia central la media, medida de dispersión la desviación estándar y medida de disimilitud la distancia Euclídea. Para ésta agrupación se consideró $K = 6$.

La distribución de las palabras claves en los 6 grupos formados quedó de la siguiente manera.

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
9.600	1.112	2.660	1.420	1.433	81

Table: Distribución de las palabras claves de la agrupación 2.

La proyección de los datos originales segmentadas por estos seis grupos la observamos en la siguiente figura.

La proyección de los datos originales segmentadas por estos seis grupos la observamos en la siguiente figura.

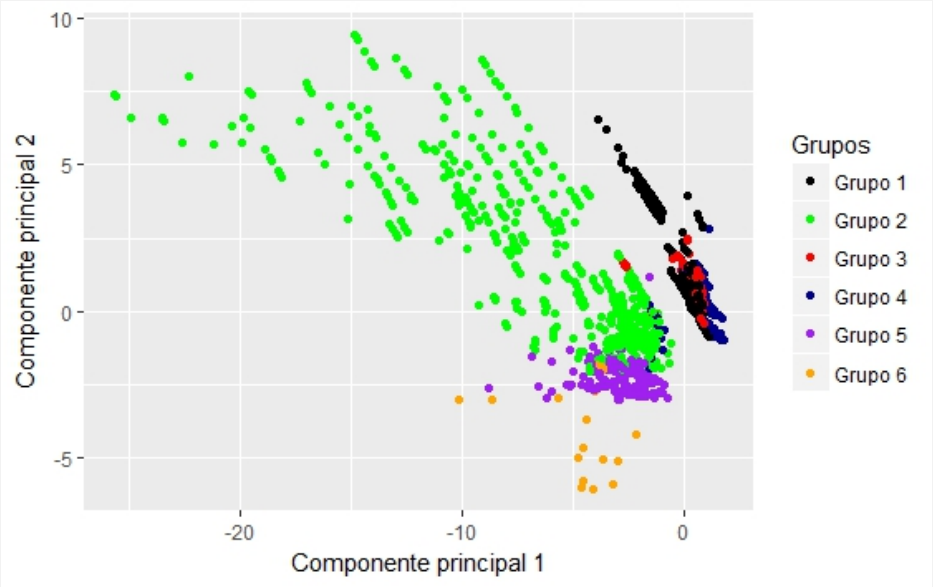

Figure: Proyección de los datos originales segmentados por los seis grupos.

La proyección de los datos originales segmentadas por estos seis grupos la observamos en la siguiente figura.

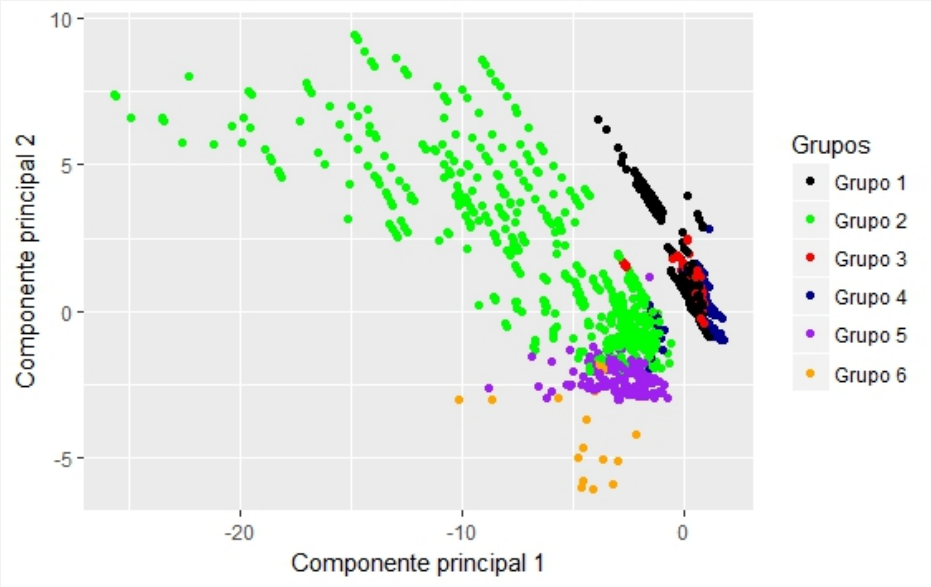

Figure: Proyección de los datos originales segmentados por los seis grupos.

Veamos en detalle las características de cada grupo formado.

Grupo 1:

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total
Min	0.13	1	0	0	0
Mediana	1	1	0	0	0
Media	1.14	1.68	0	0	0
Max	5	60	0	0	0
	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones	
Min	0	1	0	0	
Mediana	0	1	8	0	
Media	0	1.01	8.74	0	
Max	0	1.4	10	0	

Table: Resumen del grupo 1.

Grupo 2:

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total
Min	0.34	2	1	0.92	0.01
Mediana	1.31	4	1	25	1.08
Media	1.89	14.98	1.49	27.46	1.67
Max	5	132	9	60	13.94
	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones	
Min	0.01	1	5	0	
Mediana	0.98	1	8	0	
Media	1	1.2	8.64	0	
Max	2.54	2.8	10	0	

Table: Resumen del grupo 2.

Grupo 3:

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total
Min	0.25	1	0	0	0
Mediana	1	1	0	0	0
Media	1.02	2.06	0.002	0.01	0.001
Max	2.44	37	1	11.11	0.63
	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones	
Min	0	1.5	5	0	
Mediana	0	2	8	0	
Media	0.001	1.93	8.51	0	
Max	0.63	2.4	10	0	

Table: Resumen del grupo 3.

Grupo 4:

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total
Min	0.24	1	0	0	0
Mediana	0.63	1	0	0	0
Media	0.7	1.93	0.04	1.31	0.02
Max	2.44	25	1	50	0.63
	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones	
Min	0	2.5	0	0	
Mediana	0	3	8	0	
Media	0.02	3.37	8.68	0	
Max	0.63	7	10	0	

Table: Resumen del grupo 4.

Grupo 5:

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total
Min	0.13	1	1	66.67	0.01
Mediana	1	1	1	100	0.65
Media	1.12	1.07	1.06	99.77	0.75
Max	2.8	4	3	100	5.6
	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones	
Min	0.01	1	0	0	
Mediana	0.63	1	10	0	
Media	0.66	1.17	9.33	0	
Max	2.16	4	10	0	

Table: Resumen del grupo 5.

Grupo 6:

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total
Min	0.44	1	1	3.33	0.18
Mediana	1.2	1	1	100	0.93
Media	1.18	4.31	1.12	77.07	0.9
Max	2	36	6	100	3.32
	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones	
Min	0.18	1	7	1	
Mediana	0.93	1	10	1	
Media	0.83	1.2	9.11	1	
Max	1.4	2.7	10	1	

Table: Resumen del grupo 6.

Validación

Para la validación de esta agrupación, se usó el método de la silueta al igual que la agrupación 1, el valor $\bar{S}$ se muestra en la siguiente gráfica.

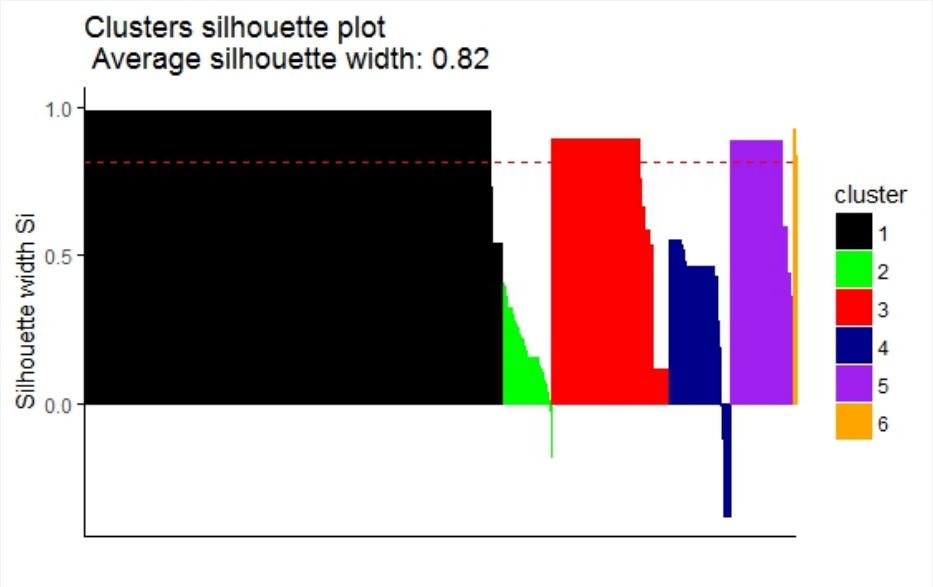

Figure: Promedio de los índices de silueta.

Modelos de regresión

Modelos de regresión

Con la intención de dar una caracterización un poco más fuerte a los grupos de cada agrupación se buscaron patrones de dependencia o estructuras entre las variables con la finalidad de establecer modelos de regresión que pudieran servir para predecir valores futuros para determinadas variables de cada grupo. En esta sección mostraremos los modelos más significativos que se obtuvieron.

Agrupación 1

Grupo 1

Para el grupo 1 de ésta agrupación no se obtuvieron clics, por lo que realizó el modelo de regresión comparando el total de las impresiones por cada posición promedio dada, el gráfico de esta comparación es el siguiente.

Agrupación 1

Grupo 1

Para el grupo 1 de ésta agrupación no se obtuvieron clics, por lo que realizó el modelo de regresión comparando el total de las impresiones por cada posición promedio dada, el gráfico de esta comparación es el siguiente.

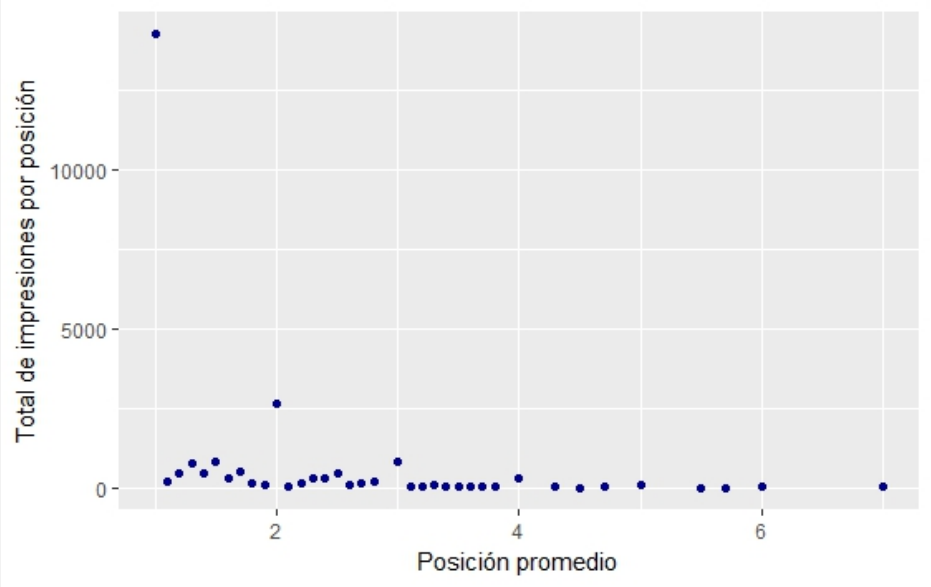

Figure: Total de impresiones por posición.

Agrupación 1

Para la gráfica mostrada en la figura anterior se ajusto un modelo de tipo exponencial obteniendo un R^2 de 0.83, con una desviación estándar de los errores de 968.1.

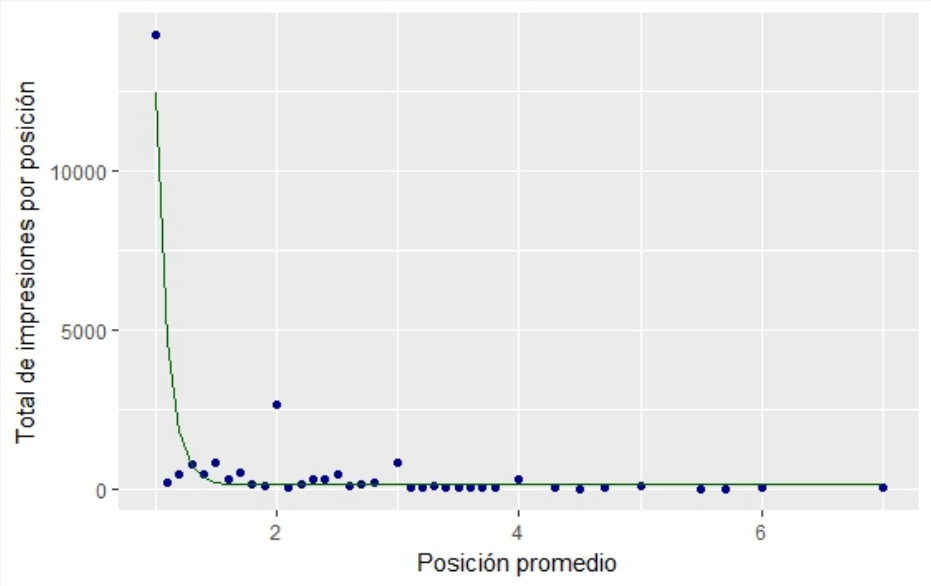

Figure: Modelo ajustado.

Agrupación 1

Grupo 3

El grupo 3 representa a las palabras que más recibieron clics, sin embargo, el rendimiento de estas palabras no fue el mejor pues obtuvieron en promedio una baja tasa de clics y ninguna conversión. Para este modelo de regresión comparamos las impresiones de las palabras con el promedio de tasas de clics, es decir, por cada impresión cual fue su promedio de tasas de clics.

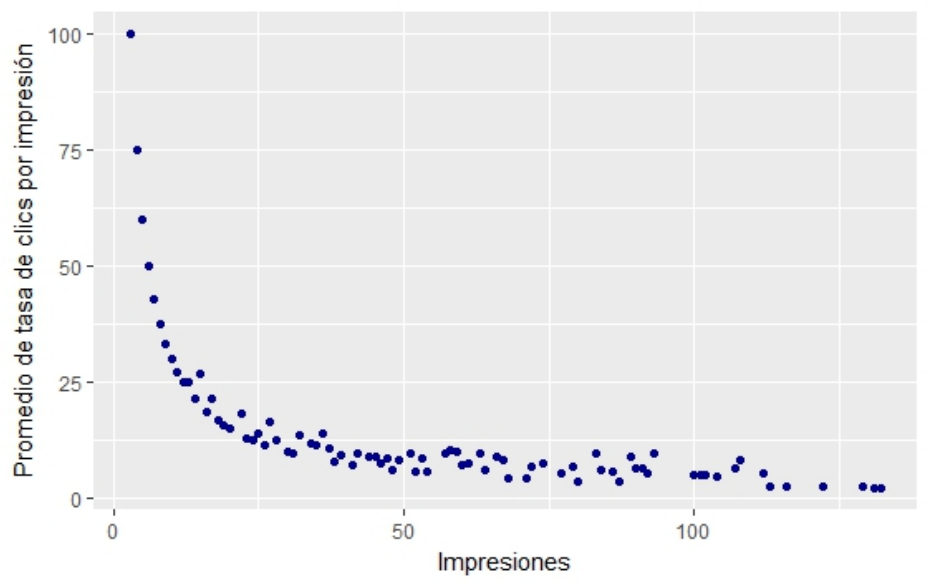

Figure: Promedio tasas de clics por impresión.

Agrupación 1

Al comportamiento mostrado en la figura anterior se le ajustó un modelo del tipo $\frac{1}{x}$ cuyo R^2 obtuvo un valor de 0.98 y una desviación estándar de los residuales de 1.87.

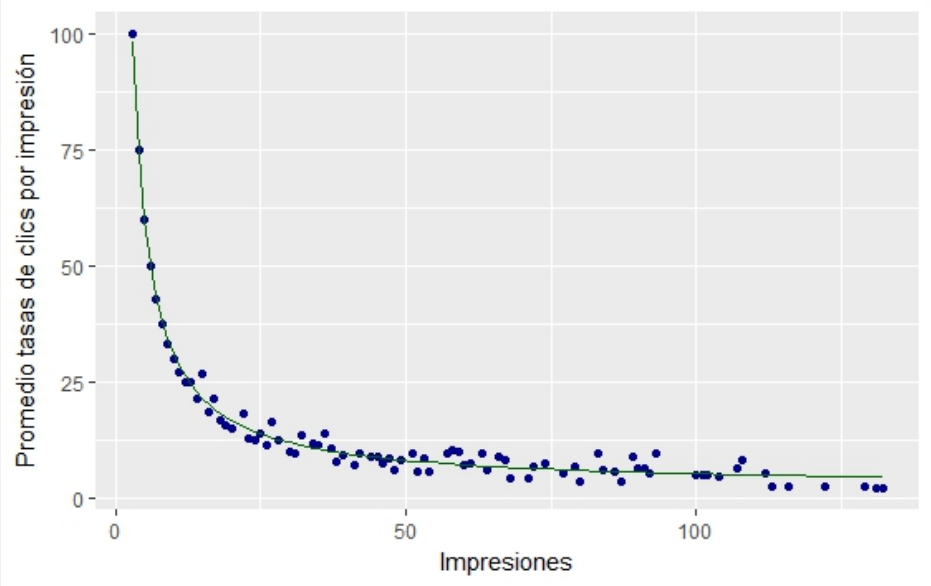

Figure: Modelo ajustado.

Agrupación 2

Grupo 2

La principal característica de este grupo es la tasa de clics la cual no supera el 60%, también este grupo obtiene en promedio el mayor costo de todos los grupos formados en la Agrupación 2. Para este modelo de regresión comparamos la cantidad total de costo por cada posición promedio dada, el gráfico es el siguiente.

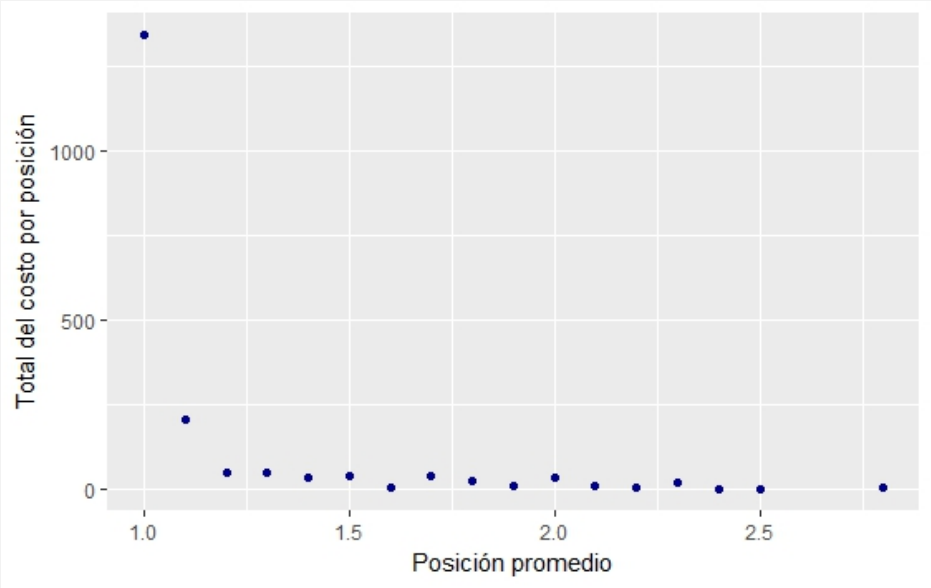

Figure: Total de costo por posición.

Agrupación 2

Se ajustó un modelo de tipo exponencial obteniendo un R^2 de 0.94, con una desviación estándar de los errores de 77.93.

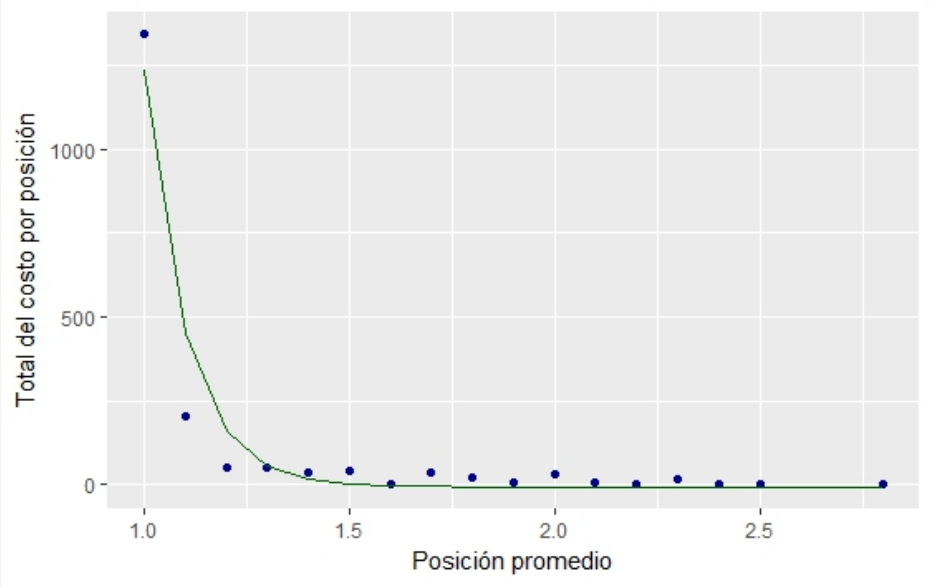

Figure: Modelo ajustado.

Agrupación 2

Grupo 5

Este grupo está formado por las palabras que obtuvieron una tasa de clic mayor al 60%. Para este grupo se ajustó un modelo comparando la oferta máxima por clic de cada palabra y su media del promedio de costo por clic. La gráfica es la siguiente.

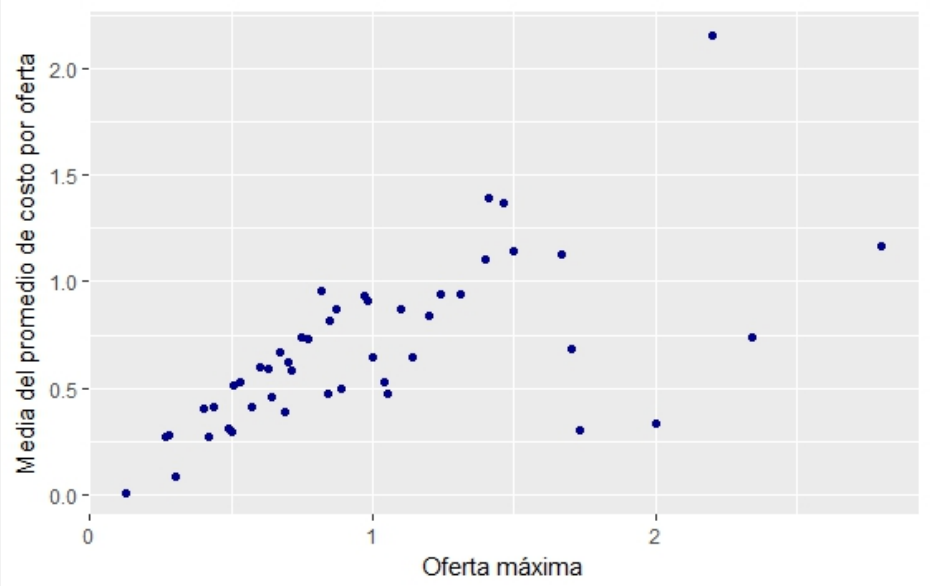

Figure: Media del promedio de costo por oferta.

Agrupación 2

Se ajustó un modelo del tipo logarítmico, resultando un R^2 de 0.47 y una desviación estándar de los residuos de 0.28.

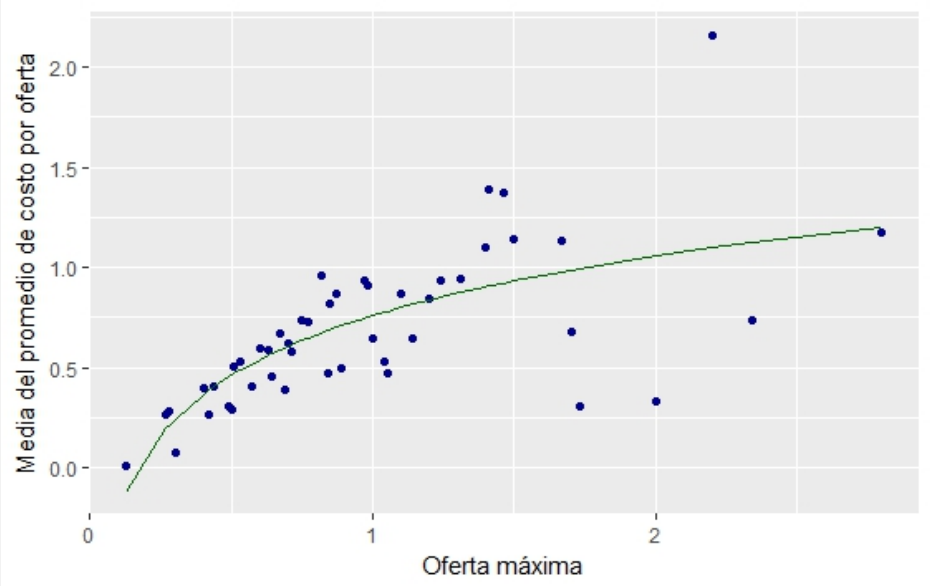

Figure: Modelo ajustado.

Conclusiones

Para las dos agrupaciones hechas pudimos notar una buena segmentación de las palabras y una calidad óptima de agrupación lo cual se corroboró con el método de la silueta, obteniendo características muy marcadas e importantes en cada grupo formado, permitiendo observar grupos de palabras que tuvieron un mejor rendimiento que otras, obteniendo grupos en los cuales habian clics y no conversiones, grupos sin clics, grupos con solo las palabras que convirtieron, entre otros. La aplicación del análisis de componentes principales fue de gran utilidad, gracias a el se consiguió reducir la dimensionalidad del espacio de variables de nueve a dos y de nueve a cuatro, con estas reducciones fue posible caracterizar y clasificar al conjunto de palabras. Otro resultado a resaltar son los modelos de regresiones, donde se encontraron buenos ajustes para el comportamiento de las variables en los diferentes grupos.

Biblografía

[1] B. Everitt, T. Hothorn (2011).
An Introduction to Applied Multivariate Analysis with R.

[2] G. Dunteman (1989).
Principal Components Analysis.

[3] A. Kassambara (2017).
Practical Guide To Cluster Analysis in R.

[4] I.T Jolliffe (2002).
Principal Component Analysis, second edition.

[5] L. Kaufman, P. Rousseeuw (1990).
Finding Groups in Data: An introduction to cluster Analysis

[6] Cluster Analysis
https://www.stat.berkeley.edu/~s133/Cluster2a.html

[7] H. Nuñez.
Talle Minería de datos UCV.

[8] Google AdWords.
https://support.google.com/adwords/answer/6319?hl=es-419