

Explorando sitios de comercio electrónico

Análisis Inteligente de Datos

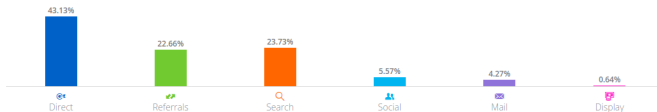
Leonardo Aranda

Agosto, 2017

Introducción

- Descubrir patrones en sitios de comercio electrónico con más visitas.
- Datos de Alexa + SimilarWeb. Ejemplo de amazon.com:

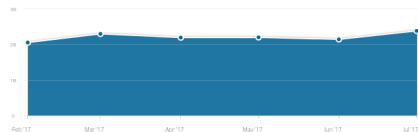
On desktop



Total Visits

On desktop & mobile web, in the last 6 months

[Embed Graph](#)



Engagement

Total Visits	2.39B ▲ 11.10%
Avg. Visit Duration	00:06:57
Pages per Visit	10.71
Bounce Rate	36.00%

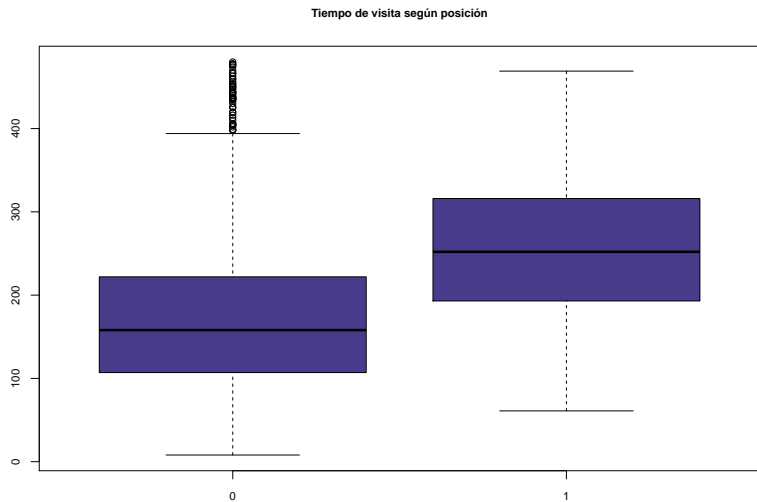
- Comportamiento de usuarios y fuentes de tráfico web.

Variable	Descripción
url	Dirección del sitio web (ID)
position	Pertenece o no al Top 50.000 mundial (Clase)
bounce	Porcentaje de visitas con una sola página vista
ppv	Páginas vistas por visita
time	Tiempo promedio de la visita
paid_search	Tráfico pago
traffic_direct	Tráfico directo
traffic_display	Tráfico de redes de publicidad
traffic_search	Tráfico de buscadores
traffic_mail	Tráfico de correo electrónico
traffic_social	Tráfico de redes sociales
traffic_referrals	Tráfico referido desde otras páginas

- Diferencia en la media de los grupos
- ¿ Son significativas ?

	V1	V2
Group.1	0	1
bounce	40.48	43.03
paid_search	13.92	18.46
ppv	4.52	5.80
time	172.99	255.88
traffic_direct	24.46	28.58
traffic_display	0.88	1.74
traffic_search	54.47	45.90
traffic_mail	4.21	5.20
traffic_social	3.12	3.84
traffic_referrals	12.86	14.74

Tiempo de visita - Distribución



■ Prueba de Shapiro-Wilk

position	p-value
0	0.0000000
1	0.0240839

■ Transformación de Box-Cox

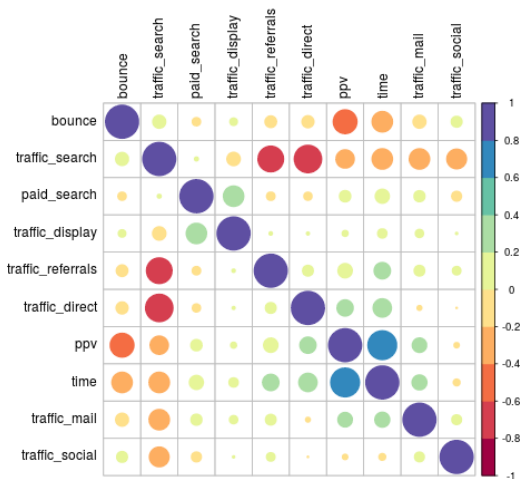
position	p-value
0	0.0002660
1	0.0019657

- H_0 : Los grupos pertenecen a la misma distribución.
- H_1 : Existen diferencias significativas en la distribución.

```
##  
## Wilcoxon rank sum test with continuity correction  
##  
## data: time by position  
## W = 148840, p-value < 2.2e-16  
## alternative hypothesis: true location shift is not equal to 0
```

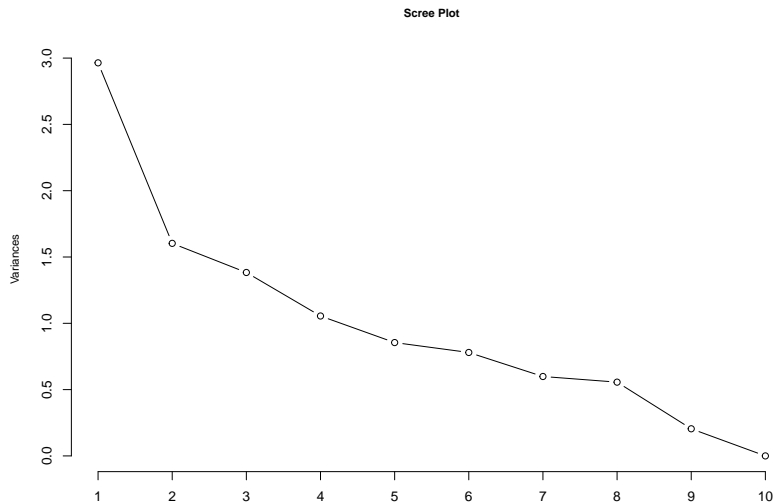
Matriz de correlación

- Correlación negativa entre Search vs Referido y Directo.
- Correlación positiva entre Directo y Tiempo de Visita.



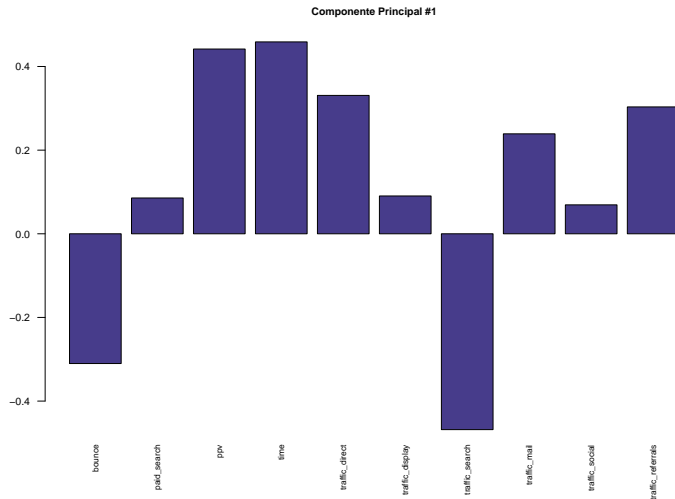
Componentes principales - Scree Plot

- Tres componentes explican el 60% de la variabilidad.

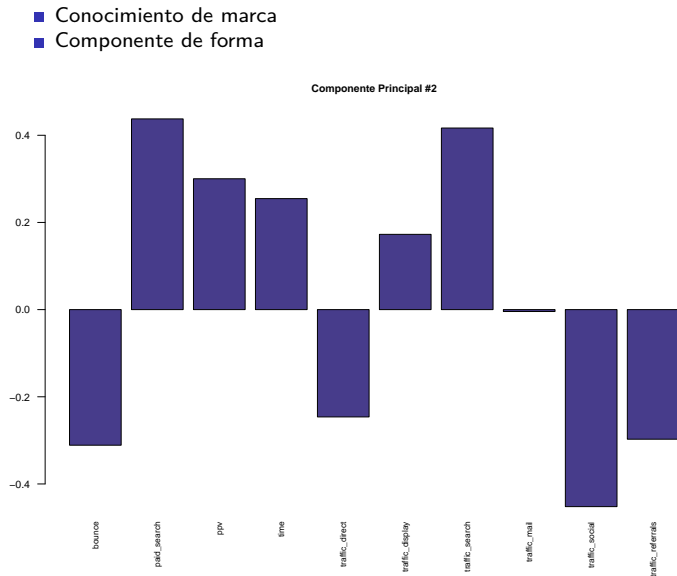


Componentes principales - Componente 1

- Nivel de interacción
- Componente de forma

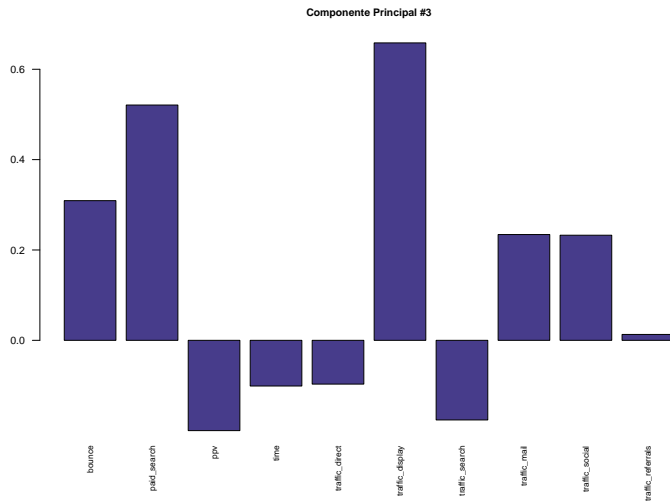


Componentes principales - Componente 2



Componentes principales - Componente 3

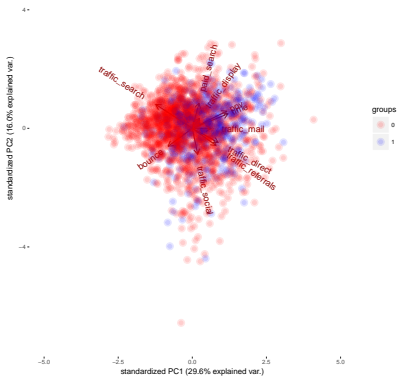
- Tráfico pago
- Componente de forma



Componentes principales - Biplot

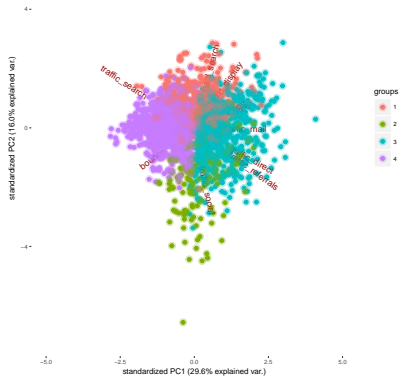
- Grupo 1 proyecta valores positivos sobre la primer componente.
- Grupo 1 tiene mayor Tiempo de Visita, tráfico Directo, Referido y desde Email.
- Grupo 0 tiene mayor Bounce.

Position sobre componentes principales



Clustering - K-means

K-means sobre componentes principais



- Cluster 3 concentra a los primeros sitios del ranking.
- Cluster 4 concentra a los últimos sitios del ranking.

cluster	0	1
1	259	106
2	122	21
3	440	187
4	704	89

Clustering - Grupos identificados

- Cluster 3: Baja tasa de rebote, elevado tiempo de visita, elevado tráfico directo.
- Cluster 4: Bajo nivel de páginas por visita, elevado tráfico pago.
- ¿ Hub vs Authority ?

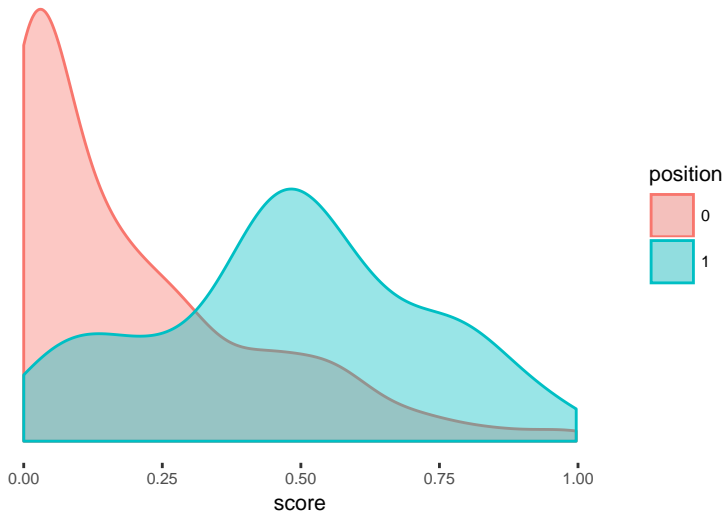
	V1	V2	V3	V4
Group.1	1	2	3	4
bounce	40.46	46.59	34.99	45.03
paid_search	38.13	7.37	11.38	8.28
ppv	5.02	3.83	6.58	3.44
time	212.41	142.62	263.66	130.75
traffic_direct	22.66	21.16	34.77	19.82
traffic_display	3.64	0.55	0.63	0.30
traffic_search	54.04	33.90	37.91	67.13
traffic_mail	5.60	12.36	5.12	1.89
traffic_social	2.17	15.56	3.01	1.77
traffic_referrals	11.89	16.46	18.56	9.10

Análisis discriminante - Resumen

- No se satisfacen supuestos de normalidad y homocedasticidad
- Análisis discriminante cuadrático
- 70% training / 30% testing.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##           0 333  57
##           1  48  44
##
##           Accuracy : 0.7822
##           95% CI : (0.7426, 0.8182)
##           No Information Rate : 0.7905
##           P-Value [Acc > NIR] : 0.6956
##
##           Kappa : 0.3201
##           Mcnemar's Test P-Value : 0.4350
##
##           Sensitivity : 0.43564
##           Specificity : 0.87402
##           Pos Pred Value : 0.47826
##           Neg Pred Value : 0.85385
##           Prevalence : 0.20954
##           Detection Rate : 0.09129
```

Puntuaciones discriminantes



- Python: Preparación de datos.
- RStudio: Análisis y documentación.
- GitHub: Almacenamiento y versionado del código.

- Los resultados permiten hacer un diagnóstico de los sitios de comercio electrónico.
- El Tiempo de Visita es significativamente más elevado en sitios con mayor volumen de tráfico.
- Sus principales fuentes de tráfico son Directo, Email y Referido.
- Los sitios con poco nivel de tráfico presentan mayor Tasa de Rebote.
- Para trabajos futuros, se podrían aplicar otros métodos que intenten mejorar el desempeño de clasificación del análisis discriminante.