Test Técnico para postular al cargo de Data Analyst/BI en Destácame

14 de febrero de 2019

Candidato
Fernando Andrés Caneo Mercado

Tabla de Contenidos

Información

- Este trabajo se hizo utilizando Python junto a sus librerias Pandas, Numpy, Matplotlib y StatsModels.
- Se adjunta un Notebook Jupyter con todo el código comentado ante cualquier consulta.
- Idealmente abrir el Notebook en la misma carpeta de los csv's entregados.
- Se irá explicando en el mismo orden que está hecho el Notebook.

Tabla de Contenidos

Pregunta 1

¿Transa más un usuario nuevo o uno que vuelve a la plataforma?

- A partir de las tablas entregadas, se aprecia inmediatamente que son de distintos largos y además presentan ciertos mecanismos e información distinta (pero relacionada) del negocio.
- Como en primera instancia se desea trabajar con las transacciones, **consideré que cualquier solicitud de crédito y pago es considerado una transacción**, en base a lo anterior, uní ambas tablas para generar una única tabla de transacciones.
- Luego se hizo un join de esta última tabla, con la de las cuentas de usuario, para saber con claridad los datos de creación de la cuenta del usuario que hizo cada transacción.

Antigüedad de la transacción

	user_id	product	created	id	birthday	gender	country_id	date_joined	last_login
0	3073847	total_payment	2018-10-07 10:42:03.487499	3073847	1996-03-13	m	1.0	2018-10-07 07:34:41	2018-10-07 10:34:42
1	3074936	total_payment	2018-10-07 20:46:34.488778	3074936	1982-02-16	f	1.0	2018-10-07 17:28:31	2019-01-29 13:56:39
2	3074936	99	2018-10-07 20:42:17	3074936	1982-02-16	f	1.0	2018-10-07 17:28:31	2019-01-29 13:56:39
3	3076049	total_payment	2018-10-08 16:47:49.824659	3076049	1996-10-25	f	1.0	2018-10-08 13:30:49	2018-10-08 16:30:49
4	3076370	total_payment	2018-10-08 20:35:31.764826	3076370	1991-06-27	f	1.0	2018-10-08 17:25:30	2019-01-29 17:26:00

Figura: .head() de la anterior tabla mencionada

- Con la anterior tabla se calcula la diferencia en días entre cuándo ocurrió la transacción y la creación de la cuenta del usuario que la realizó.
- Luego se hacen dos tipos de groupby de usuario:
  - Según las transacciones que ocurrieron el mismo día de la creación de la cuenta.
  - Según las que ocurrieron pasado 1 día o más desde la creación de la cuenta.

Clasificación de Usuarios

- De esta manera cada usuario quedó clasificado como:
  - **Nuevo :** Si transó más el mismo día de creación de su cuenta.
  - **Returning :** Si transó más posterior al día de creación a su cuenta.
  - **Same :** Si tiene la misma cantidad de transacciones tanto el día de creación como después.

La siguiente es una tabla resumen de la cantidad de estos usuarios.

Tipo de usuario	Cantidad
Nuevo	11895
Returning	3594
Same	466

Cuadro: Resumen de clasificación de usuarios.

Conclusión Preliminar

- De la anterior tabla se concluye que existe una clara dominancia por los usuarios nuevos en las transacciones, es más, representan casi un 75 % del total de usuarios que transaron en el tiempo del set de datos.

Interrogantes

1. ¿Y si el resultado anterior está sesgado por usuarios que sólo hacen 1 transacción el día de la creación de la cuenta?
2. ¿Cómo evoluciona la distribución de usuarios Nuevos/Returning a medida que discriminamos por cantidad de transacciones que realizan?

Sensibilización

Tipo de usuario	1 tr.	2 tr.	3 tr.	4 tr.	5 tr.	6 tr.
Nuevo	11895	1927	342	76	23	3
Returning	3594	876	334	100	39	18
Same	466	466	22	22	3	3

Cuadro: Sensibilización respecto a la cantidad mínima de transacciones

- En el anterior cuadro se puede apreciar claramente que a medida que analizamos a usuarios que hacen más transacciones, estas van tendiendo a disminuir pero la brecha entre usuarios Returning y Nuevos se hace más pequeña.
- Notar también que los usuarios Same bajan radicalmente a partir de 3 transacciones, lo que habla de que se empieza a generar una tendencia de uso por parte de los usuarios.

Resultados

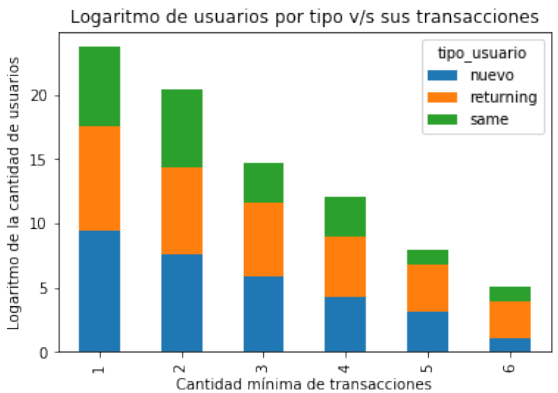

Figura: Logaritmo de usuarios por tipo v/s sus transacciones

Se puede apreciar gráficamente la baja en cantidad de usuarios nuevos y totales.

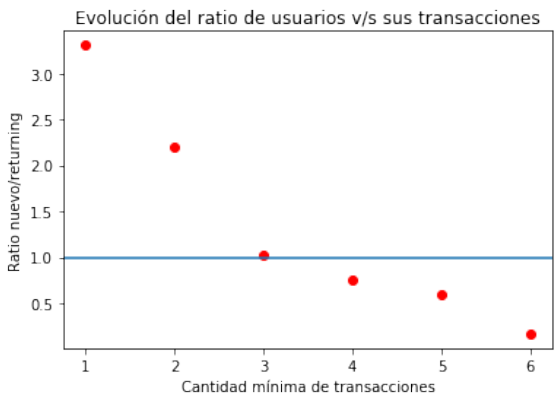

Figura: Evolución del ratio de usuarios v/s sus transacciones

La relación Nuevo/Returning también baja abrúptamente, a partir de 4 transacciones los Returning son dominantes.

Resultados Generales de Transacciones

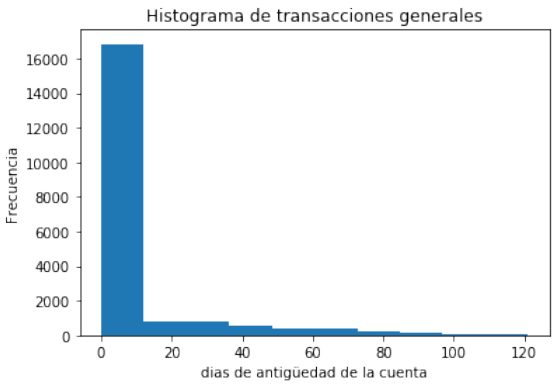

Figura: Histograma de transacciones generales

Considerando sólo las transacciones, claramente estas son dominadas por las que se realizan dentro de las primeras 24 horas de creación.

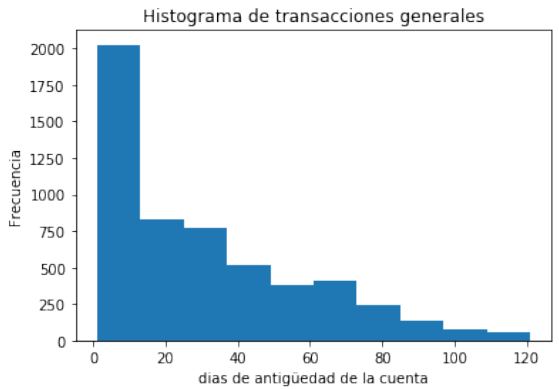

Figura: Histograma de transacciones generales

Por otro lado, si se descartan las transacciones de usuarios nuevos, estas siguen siendo más frecuentes en bajas antigüedades de cuenta.

Conclusiones Generales

- Considerando sólo el espectro de transacciones, estas son claramente más frecuentes en periodos más cercanos a la creación de la cuenta que las realizó, lo que indica que una transacción nueva es altamente probable que venga de un usuario nuevo.
- Considerando a todos los usuarios, existen muchos más que interactúan con la web el primer día de registro respecto a los que llevan más tiempo.
- A medida que un usuario hace más transacciones, es más probable que estas sean después del primer día de registro.

Pregunta 2

¿Hay alguna característica que nos permita identificar a un usuario que transa versus uno que no lo hace?

Esta pregunta fue abordada con dos enfoques

1. Teniendo en cuenta a los usuarios que llenaron todos los datos, ¿Cuáles de sus características hacen que sea más probable que transe?
2. Considerando la insuficiencia de datos (no rellenado), ¿Cuáles características hacen que sea **menos** probable que se transe?

En la práctica se "reciclaron" muchas partes de código anteriormente usado para la pregunta anterior (Ver Notebook).

Enfoque sin NaN's

- En este primer enfoque se eliminaron todos los NaN's para que queden sólo características puras.
- Se estimó una regresión logística, en donde la variable explicada se consideró como la categoría de si el usuario ha transado o no (1 para transar al menos una vez, 0 en otro caso).
- Las variables regresoras consideradas fueron:
  - País
  - Salario
  - Si pertenece al mercado laboral o no
  - Género
  - Edad

Pregunta 2

Resultados sin NaN's

Model:	Logit	Pseudo R-squared:	0.291
Dependent Variable:	transacciones	AIC:	75481.1257
Date:	2019-02-14 15:36	BIC:	75541.5371
No. Observations:	174305	Log-Likelihood:	-37735.
Df Model:	5	LL-Null:	-53219.
Df Residuals:	174299	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	8.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
country_id	0.6382	0.0198	32.1668	0.0000	0.5993	0.6771
entry	0.0000	0.0000	0.6939	0.4877	-0.0000	0.0000
trabaja	3.3055	0.0288	114.5974	0.0000	3.2490	3.3621
gender_f	-4.7048	0.0577	-81.5343	0.0000	-4.8179	-4.5917
gender_m	-4.6689	0.0579	-80.6365	0.0000	-4.7824	-4.5555
edad	-0.0194	0.0012	-16.0966	0.0000	-0.0217	-0.0170

Figura: Tabla de resultados de la regresión logística.

- Se aprecia que el salario declarado no es significativo.
- Si el usuario es del país 2, es más probable que transe respecto al país 1.
- Si el usuario pertenece a la fuerza laboral es más probable que transe.
- Si bien según el modelo, llenar la casilla de género disminuye las posibilidades de transar, no existe diferencia apreciable entre que se vea más afectado un sexo u otro.
- A mayor edad, menor probabilidad de realizar una transacción.

Enfoque con NaN's

- Se trabajo con todas las tablas, pero se hicieron variables categóricas de si el usuario relleno o no tal característica.
- Al igual que el anterior enfoque se estimó una regresión logística, con la misma variable explicada (si ha transado o no).
- Las variables consideradas fueron:
  - Si **no** llenó el género o si.
  - Si no llenó el salario o si.
  - Si no llenó su trabajo o si.

Pregunta 2

Resultados con NaN's

Model:	Logit	Pseudo R-squared:	0.208
Dependent Variable:	transacciones	AIC:	84659.1863
Date:	2019-02-14 15:33	BIC:	84689.4127
No. Observations:	175512	Log-Likelihood:	-42327.
Df Model:	2	LL-Null:	-53466.
Df Residuals:	175509	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	9.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
gender_nan	-1.4267	0.1471	-9.6977	0.0000	-1.7151	-1.1384
entry_nan	-1.9392	0.1064	-18.2226	0.0000	-2.1478	-1.7306
employment_status_nan	-3.2648	0.1134	-28.7796	0.0000	-3.4872	-3.0425

Figura: Tabla de resultados de la regresión logística.

- Todos los coeficientes son negativos, por lo que no rellenar alguna de las categorias analizadas disminuye la probabilidad de transar.
- No rellenar la categoría de trabajo es la situación que más reduce las posibilidades de transar.

Conclusiones Generales

- Existe una clara distribución de todas las transacciones hechas en la web, la cual muestra que estas se concentran en el primer día de creación de cuentas.
- Es más probable que un usuario que haga pocas transacciones, las haya hecho en el primer día de creación de su cuenta y a medida que transe más están sean en días posteriores.
- Existen características de un usuario que hacen aumentar la probabilidad de que un usuario transe, además existen características de la información que entrega este que disminuyen la probabilidad de que transe.

Test Técnico para postular al cargo de Data Analyst/BI en Destácame

14 de febrero de 2019