¿Quién gasta más y quién se registra? Evidencia observacional en CheMarket

Adrián Arturo Suárez García
202123771
a.suarezg@uniandes.edu.co

Luis Alejandro Rubiano Guerrero
202013482
la.rubiano@uniandes.edu.co

Gabriel Alejandro Moreno Riveros
202014583
g.morenor@uniandes.edu.co

I. Introducción

A través de este informe, nuestro grupo de economistas y científicos de datos emplea diferentes técnicas de análisis estadístico y de machine learning para estudiar el comportamiento de los usuarios en **CheMarket Inc.** En particular, nos enfocamos en entender qué variables impulsan el revenue de la compañía, con el fin de generar evidencia que apoye la toma de decisiones estratégicas informadas.

El análisis se divide en dos partes complementarias. En la primera parte, trabajamos con datos observacionales para explorar patrones en el comportamiento de los usuarios: ¿quiénes gastan más?, ¿qué factores están asociados con el registro en la plataforma?, ¿qué variables ayudan a predecir el gasto individual? A partir de regresiones y modelos predictivos evaluamos si existe una diferencia sistemática en el gasto entre usuarios registrados y no registrados, y discutimos posibles sesgos que afectan la interpretación causal.

En la segunda parte, analizamos un experimento aleatorio en el que se facilitó el proceso de registro para un grupo de usuarios. Esta intervención nos permite estimar de manera más rigurosa el efecto del registro sobre el gasto, así como evaluar la efectividad del nuevo diseño en incrementar la tasa de registros. Finalmente, reflexionamos sobre la validez de los resultados, las limitaciones del experimento y presentamos recomendaciones concretas sobre si conviene escalar esta intervención.

II. Datos observacionales: ¿Qué impulsa las ventas?

1. Datos y preparación

Para el análisis observacional, contamos con un conjunto de 100.000 observaciones históricas de los usuarios de CheMarket. Las variables disponibles son `time_spent`: tiempo en el sitio durante la sesión, `past_sessions`: número de sesiones anteriores, `device_type`: tipo de dispositivo (móvil, escritorio, tablet), `os_type`: sistema operativo (OS X, Windows, otros), `is_returning_user`: si el usuario ya había visitado antes, `sign_up`: si se registró o no y `revenue`: cuanto gastó en cada sesión.

En la siguiente tabla se presentan las estadísticas descriptivas de estas variables.

Tabla I
Estadísticas descriptivas de las variables

	Tiempo	Sesiones	Dispositivo	OS	Recurrente?	Revenue
Mín	0.000125	0	Tablet: 9882	Otro: 9909	No: 4930	0.5451
25%	1.4247	2	Escritorio: 40009	OS X: 30253	Sí: 95070	2.3400
Mediana	3.4384	3	Móvil: 50109	Windows: 59838		3.1370
50%	4.9946	3.001				3.9766
75%	6.9117	4				4.5221
Máx	54.3989	14				36.2934

Estadísticas descriptivas de las variables numéricas y cantidad por clase para las categóricas.

En la siguiente figura se presentan las distribuciones originales de `Revenue` y `time_spent`. Se puede observar que la densidad de ambas variables muestra muchos valores pequeños y pocos muy grandes. Para nuestro análisis decidimos aplicar una transformación logarítmica y de raíz cuadrada respectivamente, ya que estas transformaciones comprimen la cola y acercan la distribución a algo más gaussiano, lo que beneficia métodos lineales y tests que asumen normalidad de errores.

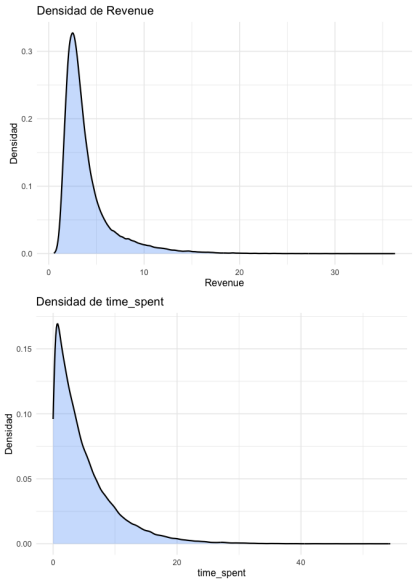

Fig. 1. Distribución de `Revenue` y `time_spent`

2. Estimación del efecto de registrarse sobre el gasto

En este caso queremos analizar el efecto de registrarse sobre el gasto de los usuarios. Sin embargo, es importante tener en cuenta que la relación entre estas variables puede estar influenciada por otras variables de nuestro modelo, en

particular, en el siguiente gráfico presentamos como cambia el gasto dentro de diferentes clases de usuarios.

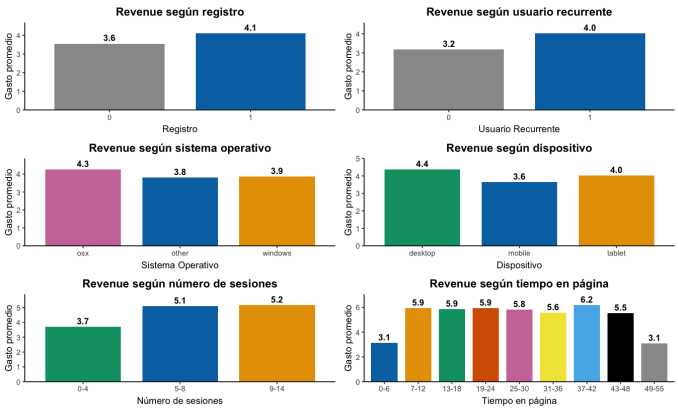

Fig. 2. Gasto promedio por clase de usuario

Queremos tener esto en cuenta ya que estimar de manera ingenua el efecto del registro sobre el gasto podría llevar a conclusiones erróneas. Por ejemplo, si observamos que los usuarios registrados tienden a gastar más, esto podría deberse a que otras variables los hicieron registrarse, y puede que estas variables también estén teniendo un efecto sobre el gasto.

3. Capacidad predictiva del modelo

4. Recomendación preliminar

III. Datos experimentales: ¿Funciona facilitar el registro?

1. Verificación del experimento

Se implementó un cambio en el diseño del sitio para facilitar el proceso de registro. Esta intervención fue asignada aleatoriamente a algunos usuarios, lo que permite evaluar su impacto sobre el gasto como un experimento de manera más rigurosa.

Queremos ver si el experimento fue bien implementado, es decir, si los grupos de control y tratamiento son comparables en términos de características observables.

Para ello, se realizó un análisis de balance entre el grupo de tratamiento y el de control. Para las variables continuas se aplicaron pruebas t de diferencias de medias, mientras que para las categóricas se emplearon pruebas de Chi-cuadrado (χ^2), además de comprobar de manera gráfica. En la siguiente figura se muestran las medias de las diferentes covariables entre los grupos de tratamiento y control. Posteriormente, se muestran los resultados estadísticos de la prueba de diferencia de medias.

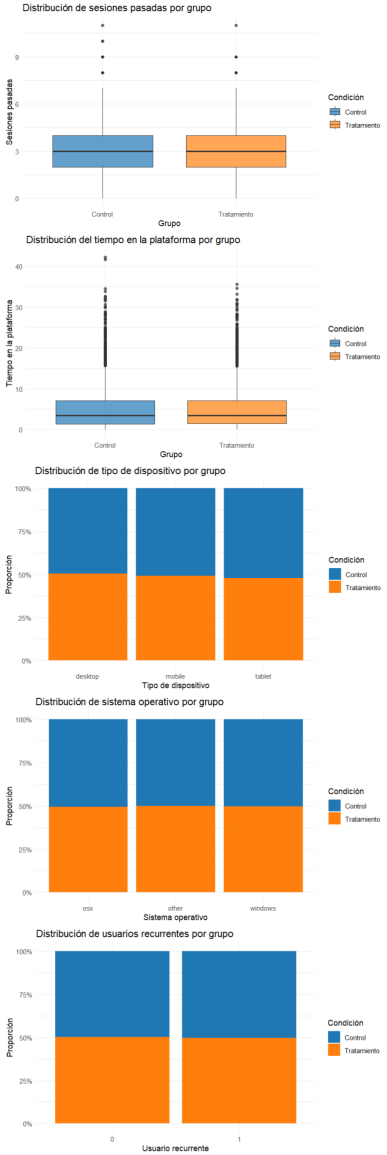

Fig. 3. Análisis de balance entre grupos de tratamiento y control

Tabla II
Balance entre grupos de control y tratamiento

Variable	Media (Control)	Media (Tratamiento)	p-valor (test)
Tiempo en plataforma	4.97	5.09	0.247
Sesiones pasadas	2.99	3.04	0.245
Tipo de dispositivo	–	–	0.210 (Chi^2)
Sistema operativo	–	–	0.924 (Chi^2)
Usuario recurrente	–	–	0.771 (Chi^2)

Prueba de balance entre grupos de control y tratamiento.

Los resultados muestran que no existen diferencias estadísticamente significativas entre los grupos en las covariables observadas. Esto respalda la validez del procedimiento de asignación aleatoria, garantizando que el tratamiento y el control son comparables al inicio del experimento.

2. Efecto sobre el registro

Con el objetivo de evaluar si una interfaz más rápida en el proceso de registro incrementa la probabilidad de que los usuarios completen exitosamente su inscripción en la plataforma, se estimó un modelo de regresión logística para identificar el efecto del tratamiento sobre la probabilidad de registro, los resultados se presentan en la siguiente tabla.

Tabla III
EFECTO DEL TRATAMIENTO SOBRE EL REGISTRO

	Coeficiente	**Error estándar**
Intercepto	-3.06e-14***	(4.35e-16)
Tratamiento (easier_signup)	1.000***	(6.17e-16)
Observaciones	10,000	

Errores estándar entre paréntesis.
$***p < 0.01, **p < 0.05, *p < 0.1$.

Regresión logística para estimar el efecto del tratamiento sobre la probabilidad de registro.

Los resultados muestran que crear una interfaz de rápido acceso hace que prácticamente todos los individuos del grupo de tratamiento completaron el registro, en contraste con el grupo de control. Sin embargo, la magnitud absoluta del efecto (100%) sugiere la presencia de separación perfecta, lo que obliga a ser prudentes en la generalización a otros contextos.

3. Efecto sobre el gasto

Con el objetivo de evaluar si una interfaz más rápida en el proceso de registro incrementa la gasten en la plataforma, se estimó un modelo de regresión para identificar el efecto del tratamiento sobre la ganancias. Los resultados se presentan en la siguiente tabla.

Tabla IV
EFECTO DEL TRATAMIENTO SOBRE LOS INGRESOS

	Coeficiente	**Error estándar**
Intercepto	3.978***	(0.051)
Tratamiento (easier_signup)	0.491***	(0.073)
Observaciones	10,000	
R-cuadrado	0.0045	

Errores estándar entre paréntesis.
$***p < 0.01, **p < 0.05, *p < 0.1$.

Los resultados muestran que las personas que usan una interfaz mas rapida generaron 0.49 unidades monetarias más de ingreso que las personas con la interfaz original, sin embargo este efecto no es significativo sólo explica una pequeña parte de la variación de los ingresos por $R^2(0.0045)$.

4. Limitaciones y robustez

El experimento aleatorio proporciona evidencia clara de que la interfaz más rápida aumenta los registros y produce un incremento promedio en ingresos. Sin embargo hay problemas con la separación perfecta entre tratados y controles, lo que puede implicar que la magnitud del experimento no sea totalmente fiable. La presencia potencial de variables omitidas podría explicar la separación perfecta además de sesgar la estimación. Otra limitación del experimento es el bajo poder explicativo que tiene el experimento ya que la mayor parte de la variación en Revenue queda sin explicar, lo que sugiere heterogeneidad relevante entre usuarios y limita la capacidad del experimento para predecir resultados económicos agregados.

5. Recomendación final

El experimento muestra que la interfaz de registro más rápida genera un efecto positivo tanto en los ingresos como en la tasa de registros. Sin embargo, recomendamos mantener la intervención en la escala actual antes de una expansión, dado que la base de información disponible es limitada y puede ocultar factores relevantes que influyen en la relación observada, lo que introduce un riesgo de sesgo por omisión. Para reducir esa incertidumbre y mejorar la capacidad explicativa del modelo, es necesario ampliar la recolección de datos incorporando variables adicionales sobre el comportamiento de los usuarios, características de perfil, factores contextuales. de modo que el análisis econométrico pueda ofrecer estimaciones más sólidas y confiables que sustentan una decisión estratégica con mayor certeza.

IV. CONCLUSIONES

Resumen
Recomendaciones accionables.