Profundizando el análisis de *CheMarket Inc.*

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

Gianluca Cicco
202020881
g.cicco@uniandes.edu.co

Juan Sebastián Sierra
202020881
j.sierrat@uniandes.edu.co

I. Introducción

A través de este informe, nuestro grupo de economistas y científicos de datos emplea nuevas técnicas de análisis estadístico y de machine learning para descubrir relaciones no lineales y efectos heterogéneos en 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.

II. Datos observacionales: ¿Qué impulsa las ventas?

1. Datos y preparación

Para el análisis observacional, al igual que en nuestro primer informe 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.

Aquí dividimos la base en entrenamiento (70%) y validación (30%) y transformamos la variable objetivo a logartimico (log_revenue) para estabilizar la dispersión. Luego, enumeramos todas las variables que el modelo puede usar y, en algunos casos incluimos transformaciones simples o interacciones como en nuestro primer informe.

2. Regresiones Lineales

Vamos a evaluar, un modelo lineal clásico (OLS) y para mejorar la predicción y evitar posibles sobreajustes, usamos además dos variantes: Ridge y Lasso. Ambos agregan un término de penalización. Ridge añade una penalización L2 (cuadrada) a la magnitud de los coeficientes, mientras Lasso usa una penalización L1 (valor absoluto). Para elegir la fuerza de esa penalización (el hiperparámetro λ) utilizamos validación cruzada en datos de entrenamiento. Entonces seleccionamos dos candidatos: λ_{min}, que minimiza el MSE medio, y además λ_{1se}, que es el valor más grande de λ cuyo error sigue estando dentro de una desviación estándar del mínimo, pues permite escoger el modelo más sencillo (menor sobreajuste) sin sacarificar en gran medida la exactitud.

Tabla II
Comparación de modelos predictivos

Modelo	MSE	RMSE	R^2
LM Básico	0.2840	0.5329	0.0057
LM Todas	0.1985	0.4455	0.3050
LM Interacciones	0.1983	0.4453	0.3058
Ridge (λ_{min})	0.1984	0.4455 ($\lambda = 0.025197$)	0.3051
Ridge (λ_{1se})	0.1991	0.4462 ($\lambda = 0.058207$)	0.3030
Lasso (λ_{min})	0.1983	0.4453 ($\lambda = 0.000311$)	0.3058
Lasso (λ_{1se})	0.1987	0.4457 ($\lambda = 0.008847$)	0.3043

La Tabla II muestra primero que el LM Básico rinde claramente peor, lo que confirma que el gasto es fuertemente condicional en las otras variables. Al incluir covariables (tiempo en el sitio transformado, sesiones previas, dispositivo, usuario recurrente y sistema operativo), el desempeño mejora sustancialmente y LM Interacciones apenas mejora un poco más. Cuando pasamos a Ridge y Lasso, los resultados quedan prácticamente empatados con LM Interacciones (MSE/RMSE bastante similares). Estas relaciones se deben a que con pocas variables y relaciones bastante lineales, OLS ya capta todo el efecto y la regularización no mejora mucho la predicción fuera de muestra.

3. Árboles de Clasificación y Regresión (CART)

Si nos basamos en lo anterior, sabemos las regresiones lineales (OLS, Ridge y Lasso) ofrecían desempeños muy

parecidos. Por lo tanto, si queremos evaluar posibles mejoras en el MSE/RMSE es pertinente evaluar patrones no lineales e interacciones que modelos lineales no identifican de manera apropiada. En este orden de ideas, vamos a comprobar si un árbol de regresión (CART) mejora la predicción fuera de muestra y, de paso, ofrece reglas que nos ayuden a entender subgrupos de usuarios con niveles de gasto distinto.

1) Profundidad por Iteraciones:

4. Recomendación preliminar

En este caso, vamos a entrenar un CART de regresión, usando validación cruzada y un rango de profundidades $(1, \ldots, 10)$, en donde iterativamente vamos a generar un arbol para cada profundidad dentro del rango y luego calculamos el error de validación promedio; escogemos la profundidad con mejor desempeño y en cada iteración elige la variable que más reduce el error dentro de las ramas. Puesto que, la profundidad en un arbol controla la complejidad del modelo; arboles más profundos capturan más detalles, pero pueden sobreajustar; arboles menos profundos generalizan mejor pero pueden quedarse cortos. Con validación cruzada buscamos un punto intermedio que mejor predice fuera de muestra. Con nuestros datos, el optimo fue 4. En la siguiente figura se muestra este arbol con sus subgrupos de decisión.

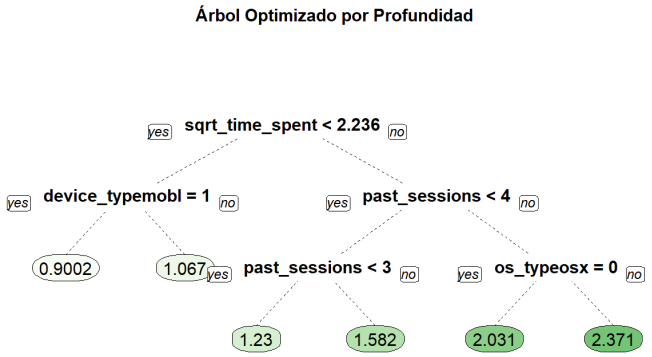

Fig. 1. Árbol de regresión con profundidad 4

Viendo el arbol, se ve que el primer nodo aparece la raíz del tiempo con un umbral cercano a 2.2. Al estar en la parte superior, indica que para predecir el gasto, el modelo primero separa a los usuarios según su tiempo en el sitio. Además, la relación no es lineal: a mayor tiempo en el sitio, mayor es el efecto sobre el gasto. Si vemos la rama izquierda, para los usuarios de poco tiempo en el sitio, el tipo de dispositivo ayuda mucho a distinguir el gasto esperado. Eso quiere decir, que los usuarios de paso rápido no se comportan igual en móvil que en desktop/Tablet. Por otro lado, la rama derecha, exhibe que para las personas que pasan mas tiempo en sitio, el historial de sesiones marca claras diferencias en el gasto. Para los usuarios con el mayor número de sesiones previas (> 4), conviene distinguir entre sistema operativo (os_type), y es evidente que conviene optimizar la experiencia para dispositivo con Os, pues tienen el mayor gasto esperado en este nodo.

1) Pruning (poda) para profundidad: A diferencia del método anterior, en donde controlando ex ante su tamaño, vamos a dejar el árbol completo y luego lo vamos a ir simplificando mediante pruning (poda) para evitar sobreajuste. Para ello definimos la ecuacion de costo de complejidad: $C_\alpha(T) = \text{Error}(T) + \alpha|T|$.

Donde *Error*(T) resume el error cuadrático en las hojas y $|T|$ es cuántas hojas usamos.

El parámetro α (cp) es la penalización por complejidad, que funciona como un umbral mínimo de mejora en el error; cada posible división se acepta solo si reduce el error de predicción en más de cp; de lo contrario, esa rama se elimina. La idea, es ir podando con un cp fijo hasta encontrar el arbol que optimiza el costo de complejidad.

No obstante, para determinar este hiperparametro(cp) utilizamos validación cruzada 10-fold y con un rango de valores (de 0.00001 a 0.0001). Probamos todos esos valores de cp, y nos quedamos con el que mejor predice fuera de muestra.

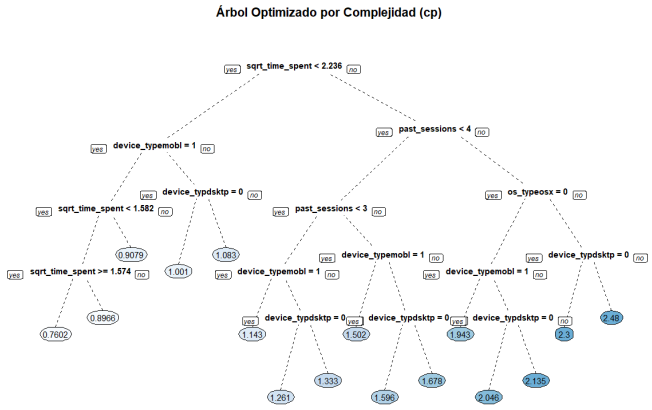

Fig. 2. Árbol de regresión con pruning por complejidad

Como en el árbol previo, la primera división está en la variable sqrt_time_spent, confirmando que el tiempo en el sitio es el predictor más relevante. A partir de ahí, el árbol sigue manteniendo la jerarquía global de manera general (Tiempo – Sesiones - Dispositivos/OS), lo cual implica consistencia entre los métodos para plantear las políticas. Sin embargo, a diferencia del método anterior, el pruning por complejidad recorta las divisiones que no mejoran la predicción y permite mayor complejidad donde si es necesario. De ahí que, el arbol queda mas profundo(Depth=6) donde la adición de esta variable si mejora la predicción del modelo. De ahí que, a gente que pasa más tiempo en el sitio y acumula más sesiones (≥ 4), no solo conviene distinguir entre sistema operativo (os_type) sino que también importa el tipo de dispositivo (device_type); dentro de

este grupo también conviene distinguir según móvil o desktop ,lo cual es crucial porque en esas ramas están los usuarios con mayor gasto esperado Pero si el grupo, debería enfocarse en un solo grupo, debería ser los usuarios que pasen mas tiempo en sitio y acumulan mas de 4 sesiones y que tengas sistema operativo OSX (predicción mas alta de gasto 2.48)

Tabla III
COMPARACIÓN DE MODELOS PREDICTIVOS

Modelo	MSE	RMSE	R^2
LM Básico	0.2840	0.5329	0.0057
LM Todas	0.1985	0.4455	0.3050
LM Interacciones	0.1983	0.4453	0.3058
Ridge (λ_{min})	0.1984	0.4455 ($\lambda = 0.025197$)	0.3051
Ridge (λ_{1se})	0.1991	0.4462 ($\lambda = 0.058207$)	0.3030
Lasso (λ_{min})	0.1983	0.4453 ($\lambda = 0.000311$)	0.3058
Lasso (λ_{1se})	0.1987	0.4457 ($\lambda = 0.008847$)	0.3043
Árbol (Profundidad)	0.1256	0.3544 (maxdepth = 5)	0.5602
Árbol (Complejidad)	0.1228	0.3504 (cp = 0.000060)	0.5701

2) Analisis conjunto: Viendo la tabla comparativa, el mejor desempeño predictivo lo entrega el árbol con poda por complejidad (cp alrededor de 0.000060) con MSE = 0.1228, RMSE = 0.3504 seguido muy de cerca por el árbol optimizado por profundidad (maxdepth = 5) con MSE = 0.1256 y RMSE = 0.3544 . Esta caída del error fuera de muestra a comparación de los modelos lineales, confirma que en nuestros datos había no linealidades y umbrales (tiempo en el sitio, 3 y 4 sesiones, OS/dispositivo) que la forma lineal no capturaba; el árbol, en cambio, las traduce en reglas por segmentos y por eso predice mejor.

III. DATOS EXPERIMENTALES: ¿FUNCIONA FACILITAR EL REGISTRO?

En esta segunda parte evaluamos el efecto causal de facilitar el registro (sign_up) sobre el gasto de los usuarios de CheMarket Inc., aprovechando el diseño experimental de la base de datos. Nuestro objetivo es estimar el impacto promedio y explorar heterogeneidad de efectos entre distintos perfiles de usuario.

1. Datos y preparación

Partimos de la misma base experimental usada en el Taller 1, con 10,000 observaciones y variables: sign_up (tratamiento), time_spent, past_sessions, device_type, os_type, is_returning_user y revenue. Transformamos la variable de resultado a logaritmo natural (ln_revenue) para estabilizar la varianza.

2. Efecto promedio del tratamiento

Primero estimamos el efecto medio de sign_up mediante regresiones OLS:

- Modelo Base: solo incluye la variable de tratamiento.
- Modelo Completo: agrega todos los controles (tiempo, sesiones, dispositivo, sistema operativo, usuario recurrente).
- Modelos con interacciones: permiten que el efecto de sign_up varíe con las características tecnológicas y de comportamiento.

Esta tabla se encuentra al final.

El efecto promedio de sign_up resulta muy pequeño y no significativo en los modelos básicos (0.013 en el modelo base y 0.001 en el completo). Sin embargo, al incluir interacciones, el coeficiente principal se eleva (0.236–0.394) pero acompañado de efectos diferenciales. En particular, los usuarios que acceden desde dispositivo móvil o emplean el sistema Windows/Other presentan interacciones negativas y significativas, lo que atenúa el impacto general. En contraste, ser un usuario recurrente amplifica notablemente el efecto (0.167). Esta heterogeneidad se refleja también en la capacidad explicativa del modelo, que aumenta de un R^2 cercano a 0.0 hasta aproximadamente 0.35, confirmando que las diferencias entre grupos son clave para entender la respuesta observada.

3. Heterogeneidad: Árbol Causal

Para identificar subgrupos de usuarios con efectos distintos, aplicamos un honest causal tree. El árbol divide la población según umbrales en time_spent y past_sessions, y según combinaciones de device_type y os_type.

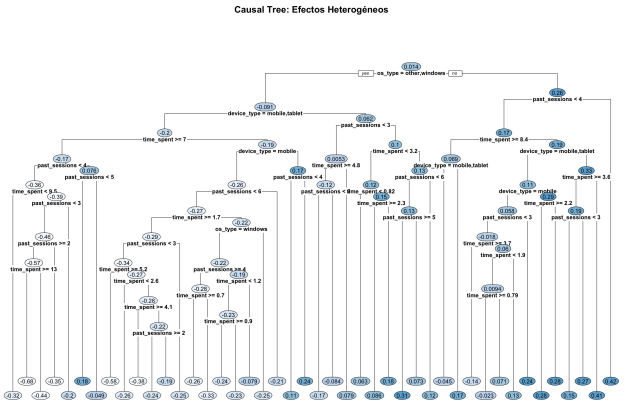

Fig. 3. Árbol Causal de Efectos Heterogéneos del Registro

Principales hallazgos del árbol:

- Los nodos iniciales separan a los usuarios por tiempo en sesión y número de sesiones previas, confirmando su importancia.
- Se observan hojas con efectos positivos de hasta ≈ 0.4 (usuarios con tiempo largo y varias sesiones, sobre todo en OS X) y otras con efectos negativos (usuarios móviles con pocas sesiones).
- Esto sugiere que conviene focalizar el incentivo de registro en usuarios con mayor engagement y dispositivos de escritorio/OS X.

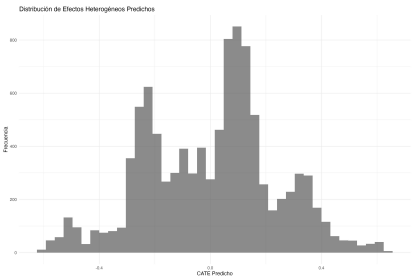

Fig. 4. Distribución de CATEs del Registro

4. Bosque Causal y CATEs

Para robustecer los resultados usamos un causal forest (grf) con 10000 árboles, que permite estimar el efecto de tratamiento condicional (CATE) para cada observación.

Los CATEs confirman una distribución amplia: hay usuarios con efectos cercanos a cero, otros claramente positivos (> 0.3) y un grupo minoritario con efectos negativos (< -0.2).

5. Visualización de patrones de heterogeneidad

A continuación mostramos algunos cruces clave:

6. Discusión y recomendaciones

La evidencia experimental muestra que el efecto promedio de facilitar el registro sobre el gasto es bajo y poco significativo, pero esconde una heterogeneidad marcada. Los usuarios que combinan sistema operativo OS X, dispositivo de escritorio o tablet, varias sesiones previas y tiempos prolongados en el sitio presentan CATEs claramente positivos, mientras que usuarios móviles con pocas sesiones y sistemas Windows u "other" exhiben efectos nulos o incluso negativos. En conjunto, estos resultados sugieren que una política uniforme de incentivos al registro sería ineficiente; en su lugar, conviene focalizar la intervención en los segmentos de alto potencial, donde el beneficio económico de promover el registro es más alto y más seguro.

Tabla IV
EFECTOS DEL REGISTRO EN EL INGRESO DE CHEMARKET

	Dependent variable: Log(Gasto Mensual)			
	Modelo Base (1)	Modelo Completo (2)	Modelo Efectos Heterogéneos (1) (3)	Modelo
Tratamiento (Registro)	0.013	0.001	0.394**	
	(0.012)	(0.010)	(0.021)	
Tiempo	0.048**	0.048*	0.048*	
	(0.001)	(0.001)	(0.001)	
Sesiones Pasadas	0.099**	0.099*	0.099*	
	(0.003)	(0.003)	(0.003)	
Dispositivo: Móvil	-0.326**	-0.171*	-0.171**	
	(0.010)	(0.014)	(0.014)	
Dispositivo: Tablet	-0.072**	-0.076*	-0.076**	
	(0.018)	(0.024)	(0.024)	
Sistema: Otro	-0.186***	-0.001	-0.0005	
	(0.018)	(0.025)	(0.025)	
Sistema: Windows	-0.190***	-0.021	-0.021	
	(0.011)	(0.015)	(0.015)	
Usuario Recurrente	-0.098**	-0.095*	-0.179**	
	(0.024)	(0.024)	(0.032)	
Registro: Móvil		-0.308**	-0.309**	
		(0.020)	(0.020)	
Registro: Tablet		0.011	0.011	
		(0.034)	(0.034)	
Registro: Sistema Otro		-0.372**	-0.375**	
		(0.035)	(0.035)	
Registro: Sistema Windows		-0.339**	-0.340**	
		(0.022)	(0.022)	
Registro: Usuario Recurrente			0.167***	
			(0.043)	
Constante	1.222**	1.084*	0.884*	
	(0.008)	(0.025)	(0.026)	
Observations	10,000	10,000	10,000	
R^2	0.0001	0.318	0.352	
Adjusted R^2	0.00003	0.317	0.352	
Residual Std. Error	0.599 (df = 9998)	0.495 (df = 9991)	0.482 (df = 9987)	
F Statistic	1.253 (df = 1; 9998)	581.899** (df = 8; 9991)	452.654* (df = 12; 9987)	41

Note: Errores estándar en paréntesis. * p<0.01, ** p<0.05, * p<0.1.