Universidad de los Andes
Facultad de Economía

Ciencia de Datos y Econometría Aplicada
ECON-3785
Créditos: 3
2025-2

Taller 1:
Impulsando los ingresos de *CheMarket Inc.*

Bienvenidos a CheMarket Inc.

Han sido contratados como analistas de datos por *CheMarket Inc.*, una empresa de comercio electrónico líder en su sector. La compañía busca aumentar sus ingresos significativamente y ha recurrido a ustedes para entender mejor qué impulsa el gasto de los usuarios en la plataforma.

Como economistas y científicos de datos, su tarea es aplicar técnicas de análisis estadístico y machine learning para responder preguntas clave que ayuden a tomar decisiones estratégicas.

Su análisis se divide en dos partes:

- **Parte A:** Análisis exploratorio y predictivo con datos observacionales.
- **Parte B:** Análisis causal usando datos de un experimento aleatorizado.

Encontrarán dos conjuntos de datos, uno para cada parte del taller. Ambos están en formato `.Rds`

- `Parte_A.Rds` para la Parte A.
- `Parte_B.Rds` para la Parte B.

Entrega final:

- Un **informe escrito** de máximo **6 páginas** (sin contar anexos). Deben responder las preguntas planteadas, interpretar resultados, justificar sus decisiones y anticipar preguntas razonables que podrían surgir por parte de la mesa directiva.
- Dos sets de **slides para presentación**:
  - Slides Parte A (máx. 4)
  - Slides Parte B (máx. 4)
- La presentación debe durar 15 minutos, y deben enfocarse en resultados y recomendaciones. No incluyan código ni detalles técnicos en las slides.

Formato de entrega:

- `informe_equipo_XX.pdf` (reemplazar XX por el número del equipo con dos dígitos)
- `slides_parteA_equipo_XX.pdf`
- `slides_parteB_equipo_XX.pdf`

Universidad de los Andes
Facultad de Economía

Ciencia de Datos y Econometría Aplicada
ECON-3785
Créditos: 3
2025-2

Parte A: ¿Qué impulsa las ventas?

La dirección de *CheMarket Inc.* cree que registrarse en la plataforma (`sign_up`) lleva a los usuarios a gastar más. Ustedes tienen acceso a datos históricos sobre comportamiento de usuarios, y deben evaluar si esta hipótesis se sostiene en la evidencia.

El informe será leído por la mesa directiva, que espera argumentos claros, visualizaciones bien diseñadas y recomendaciones accionables. Les interesa saber si deben invertir recursos en estrategias para aumentar el registro. Su responsabilidad como analistas es presentar evidencia sólida, con el respaldo técnico justo y comprensible.

¿Qué espera leer la mesa directiva?

- **Una descripción general del comportamiento de los usuarios.** ¿Quiénes gastan más?, ¿quiénes se registran?, ¿hay diferencias visibles según el tipo de dispositivo, el tiempo en el sitio o el número de visitas previas?
- **Una estimación del efecto de registrarse sobre el gasto.** Un modelo de regresión lineal como `log(Revenue) ~sign_up` permite medir diferencias promedio en gasto entre usuarios registrados y no registrados. La dirección necesita saber si esta diferencia es significativa, grande y consistente al controlar por otras variables.
- **Una evaluación de la capacidad predictiva.** ¿Qué tan bien pueden predecir el gasto de un usuario? ¿Qué variables ayudan más a hacerlo? Separar los datos en conjuntos de entrenamiento y prueba, y reportar el MSE (error cuadrático medio) en el conjunto de prueba, les dará evidencia clara.
- **Una reflexión sobre causalidad.** La empresa quiere saber si registrarse *causa* un mayor gasto, no solo si están correlacionados. ¿Qué sesgos pueden estar afectando esa relación? ¿Creen que los usuarios que se registran ya tenían intención de gastar más? ¿Faltan variables clave?
- **Una recomendación.** ¿Vale la pena impulsar el registro? ¿Qué tan confiados están en su recomendación? ¿Conviene hacer un experimento para obtener evidencia más creíble?

Variables disponibles:

- `time_spent`: tiempo en el sitio durante la sesión
- `past_sessions`: número de sesiones anteriores
- `device_type`: tipo de dispositivo (mobile, desktop, tablet)
- `is_returning_user`: si el usuario ya había visitado antes

Universidad de los Andes
Facultad de Economía

Ciencia de Datos y Econometría Aplicada
ECON-3785
Créditos: 3
2025-2

- `sign_up`: si se registró o no
- `Revenue`: cuánto gastó el usuario en esa sesión

Preguntas que su informe debe responder:

- ¿Cuánto más gastan los usuarios registrados, en promedio?
- ¿La diferencia persiste al controlar por otras variables?
- ¿Qué tan bien predice su modelo el gasto individual?
- ¿Qué variables son más importantes para explicar el gasto?
- ¿Se puede afirmar que el registro causa mayor gasto? ¿Por qué sí o por qué no?
- ¿Es razonable recomendar una estrategia de promoción del registro?
- ¿Vale la pena hacer un experimento para obtener evidencia causal?

Herramientas que pueden ayudar:

- **Estadíticas descriptivas** (tablas y gráficos) que describan la base de datos y patrones interesantes entre las variables.
- **Modelos de regresión lineal** con y sin controles para estimar el efecto de `sign_up`.
- **Modelos de regresión logística** que estiman la probabilidad de `sign_up`
- **Métricas de desempeño predictivo** como el MSE en el conjunto de prueba para evaluar qué tan bien predicen el gasto.Para ello, separen la muestra en entrenamiento (70 %) y prueba (30 %), utilizando la semilla `2025`.

Parte B: ¿Funciona hacer más fácil el registro?

Después de analizar los datos históricos, la empresa decidió actuar: implementó un cambio en el diseño del sitio para facilitar el proceso de registro. Esta intervención fue asignada aleatoriamente a algunos usuarios (`easier_signup`), lo que permite evaluar su efecto usando un experimento aleatorio controlado.

Su tarea ahora es analizar los datos del experimento. La dirección quiere saber si este cambio logró su objetivo y si generó consecuencias adicionales sobre el comportamiento de los usuarios.

¿Qué espera leer la mesa directiva?

Universidad de los Andes
Facultad de Economía

Ciencia de Datos y Econometría Aplicada
ECON-3785
Créditos: 3
2025-2

- **Una verificación de que el experimento fue bien implementado.** ¿Los grupos asignados aleatoriamente eran comparables al inicio? ¿Hubo balance entre ellos en las principales características observables?
- **Una estimación clara del efecto del tratamiento.** La pregunta clave es: ¿el registro (`sign_up`) generó un aumento en el gasto?
- **Una estimación clara del efecto del tratamiento.** ¿El nuevo diseño aumentó la tasa de registro? ¿En cuánto? ¿Es un cambio significativo o marginal?
- **Una evaluación del impacto posterior.** ¿Registrarse más fácilmente llevó a un mayor uso o gasto en el sitio? ¿Cómo se puede medir ese efecto?
- **Una discusión sobre qué se puede concluir.** ¿Pueden atribuir estos efectos al cambio en el diseño? ¿Qué supuestos hacen falta? ¿Qué tan generalizables son los resultados?
- **Una recomendación.** ¿Conviene escalar esta intervención a todos los usuarios? ¿Qué información adicional se necesitaría para estar más seguros?

Variables nuevas relevantes para esta parte:

- `easier_signup`: indicador de asignación al grupo con registro facilitado (tratamiento)
- `sign_up`: si el usuario se registró o no
- `Revenue`: cuánto gastó el usuario

Herramientas técnicas útiles para este análisis:

- **Comparación de medias y tests de balance** entre grupos tratados y control. Esto permite verificar que la aleatorización se implementó correctamente.
- **Resultado de la intervención**. ¿Facilitar el registro indujo más registros (`sign_up`)? ¿Como fue la adopción?
- **Análisis del impacto en uso o gasto.** ¿Se puede observar un efecto del tratamiento sobre `Revenue`? ¿O sobre otras variables que midan el comportamiento posterior?

Preguntas que podrían anticipar o responder en el informe:

- ¿Funcionó la asignación aleatoria? ¿Los grupos eran comparables antes del tratamiento?

Universidad de los Andes
Facultad de Economía

Ciencia de Datos y Econometría Aplicada
ECON-3785
Créditos: 3
2025-2

- ¿El nuevo diseño aumentó la probabilidad de que un usuario se registre?
- ¿Qué efecto tuvo sobre el gasto posterior?
- ¿Es razonable interpretar estos efectos como causales?
- ¿Qué limitaciones tiene el experimento?
- ¿Recomiendan mantener o escalar esta intervención?