Trabajo Práctico 2

Alessio Bocco (boccoalessio@gmail.com)

20 de September de 2020

Contents

El presente trabajo práctico analiza los resultados de la campaña de marketing online realizada por Rocket Fuel y TaskaBella.

1 Análisis exploratorio

El dataset cuenta con usuarios que participaron del estudio. Del total, 564577 han sido expuestos a la campaña de marketing mientras que 23524 pertenecen al grupo de control. Las variables que serán consideradas en el análisis son:

- Tasa de conversión: variable binaria, si el usuario compró la cartera o no.
- Cantidad de impresiones: variable cuantitativa, numero total de impresiones a las que eran expuestos los usuarios.
- Día de la semana
- Hora del día

Las variables se resumen en la siguiente tabla.

N

Mean

SD

Min

Q1

Median

Q3

Max

test

588101

0.96

0.20

0

1

1

1

1

converted

588101

0.03

0.16

0

0

0

0

1

tot_impr

588101

24.82

43.72

1

4

13

27

2065

mode_impr_day

588101

4.03

2.00

1

2

4

6

7

mode_impr_hour

588101

14.47

4.83

0

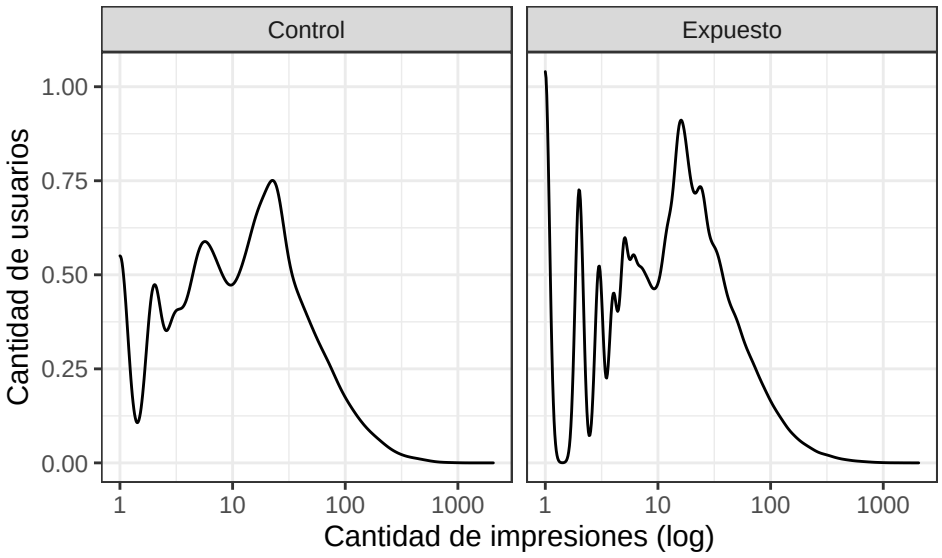
11

14

18

23

Los dos grupos, control y expuesto, fueron expuestos a avisos publiciatarios diferentes pero como muestra la Figura @ref(fig:densidad_impresiones), ambos grupos tuvieron una distribución de impresiones bastante similar.



El efecto del tiempo sobre la tasa de conversión se evalúa con información sobre el día de la semana y hora del día en la que ocurre la conversión. La Tabla @ref(tab:conversion_dia_tabla) muestra la tasa de conversión por cada día de la semana.

Día

Tasa de conversión (%)

Lunes

3.28

Martes

2.98

Miércoles

2.49

Jueves

2.16

Viernes

2.22

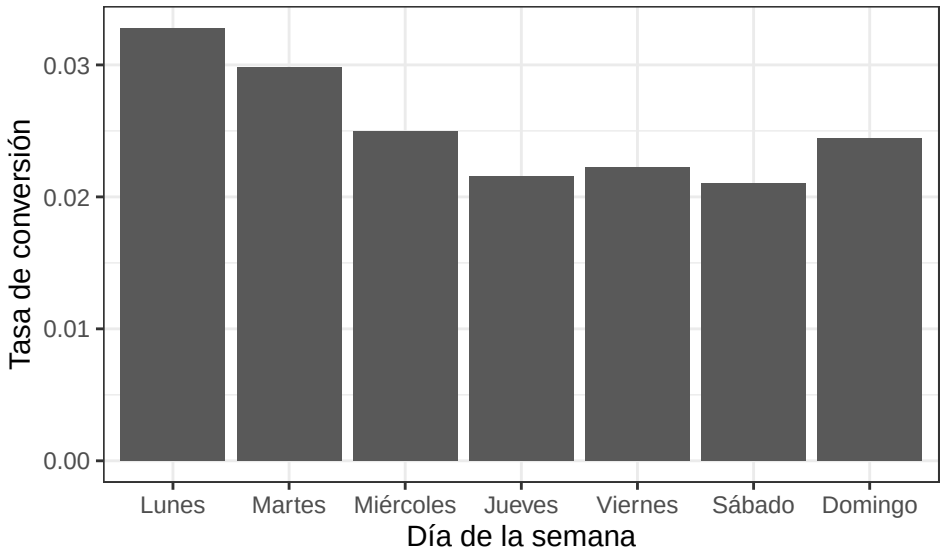
Sábado

2.11

Domingo

2.45

Se observa que la tasa de conversión es similar entre días a excepción del domingo. Con respecto a la hora del día, la Figura @ref(tab:conversion_hora_figura).



Se observa una estacionalidad en las tasas de conversión con picos en las horas de la tarde/noche.

2 Efectividad de la campaña

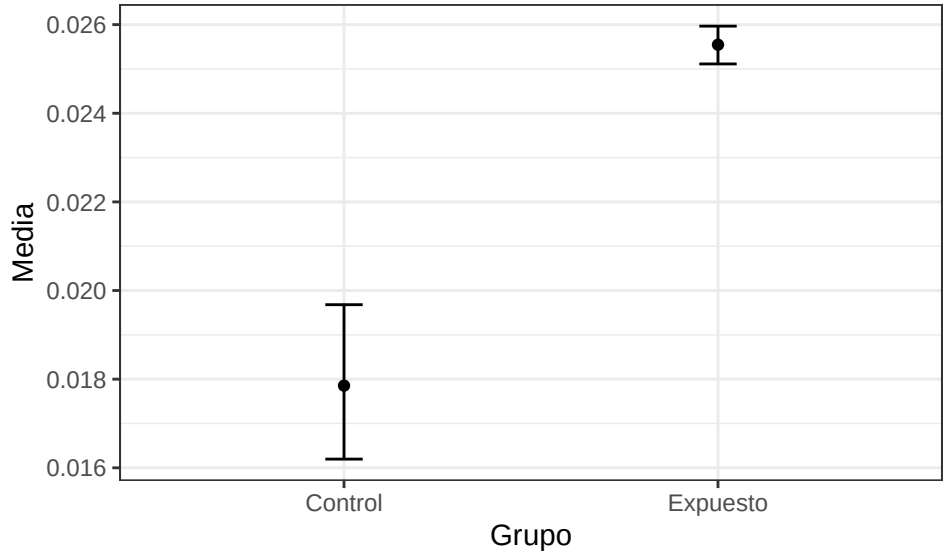
2.1 Diferencia de medias

Para evaluar la efectividad de la campaña de marketing es necesario conocer si existe una diferencia significativa entre la tasa de conversión de los usuarios expuestos y del grupo de control. Se realizó un A/B testing

para cuantificar el efecto de la campaña sobre los usuarios. La prueba compara las tasas de conversión de ambos grupos con para determinar si son iguales o no.

El test reveló una diferencia estadísticamente significativa en la tasa de conversión entre los usuarios expuestos (M = 0.0178541, SD = 0.1324239) y aquellos del grupo de control (M = 0.0255466, SD = 0.1577783), $t(5.88099 \times 10^5)$ = -7.370406, $p < .001$, d = 0.0490456. A partir de lo anterior se concluye que la campaña ha sido efectiva.

Aplicando la técnica de bootstrap es posible conocer la distribución de la diferencia entre las medias de ambos grupos. La Figura @ref(fig:conversion_hora_dia_grafico) muestra la distribución de la diferencia de medias (linea negra) y el intervalo superior de confianza del 95% (línea roja)



Se observa que la distribución toma valores negativos lo que implica que la tasa de conversión de los usuarios expuestos a la campaña ha sido superior a la tasa del grupo de control con un 95% de confianza.

A modo de resumen numérico se presentan los siguientes resultados:

- Cantidad de usuarios expuestos que compraron la cartera: 14423 (2.5547%).
- Cantidad de usuarios en control que compraron la cartera: 420 (1.7854%).
- Usuarios expuestos convertidos por la campaña: 4343

Otra alternativa para la evaluación de la efectividad consiste en utilizar modelos estadísticos como regresiones lineales y logísticas.

2.2 Regresión lineal

El modelo regresión modela la probabilidad de que un usuario expuesto a la campaña compre la cartera. El modelo estadístico es el siguiente:

$$\text{converted} = \beta_0 + \beta_1(\text{user\_group}_{\text{Expuesto}}) + \epsilon \tag{1}$$

El modelo ajustado por MCO es:

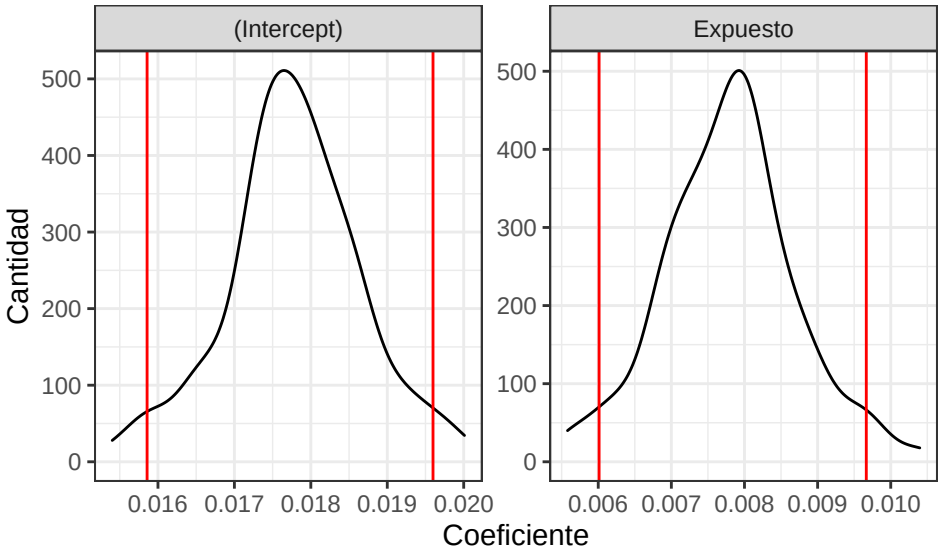
$$\hat{\text{converted}} = 0.018 + 0.008(\text{user\_group}_{\text{Expuesto}}) \tag{2}$$

El coeficiente del modelo indica que estar expuesto a la campanna aumenta la probabilidad de compra en un 0.8%. Estos resultados están en línea con el análisis numérico mostrado anteriormente.

El modelo lineal es estadísticamente significativo.

	Model 1
(Intercept)	0.018
	[0.016, 0.020]
	s.e. = 0.001
	t = 17.459
	p = 0.000
user_groupExpuesto	0.008
	[0.006, 0.010]
	s.e. = 0.001
	t = 7.370
	p = 0.000
Num.Obs.	588 101
R2	0.000
R2 Adj.	0.000
AIC	−509 964.9
BIC	−509 931.0
Log.Lik.	254 985.436
F	54.323
RMSE	0.16

Aplicando también bootstrap sobre el modelo lineal es posible conocer la variabilidad del parametro β_1 que corresponde a la probabilidad de compra. La Figura @ref(fig:histograma_coeficientes_basico) muestra la distribución de los coeficientes del modelo lineal.



2.3 Regresión logística

Una mejor alternativa al modelo lineal es la regresión logística. Este tipo de modelos se ajusta mejor a situaciones binarias (compra vs no compra) como el caso en estudio. El modelo estadístico es el siguiente:

$$\log\left[\frac{P(\text{converted}=1)}{1-P(\text{converted}=1)}\right] = \beta_0 + \beta_1(\text{user\_group}_{\text{Expuesto}}) \tag{3}$$

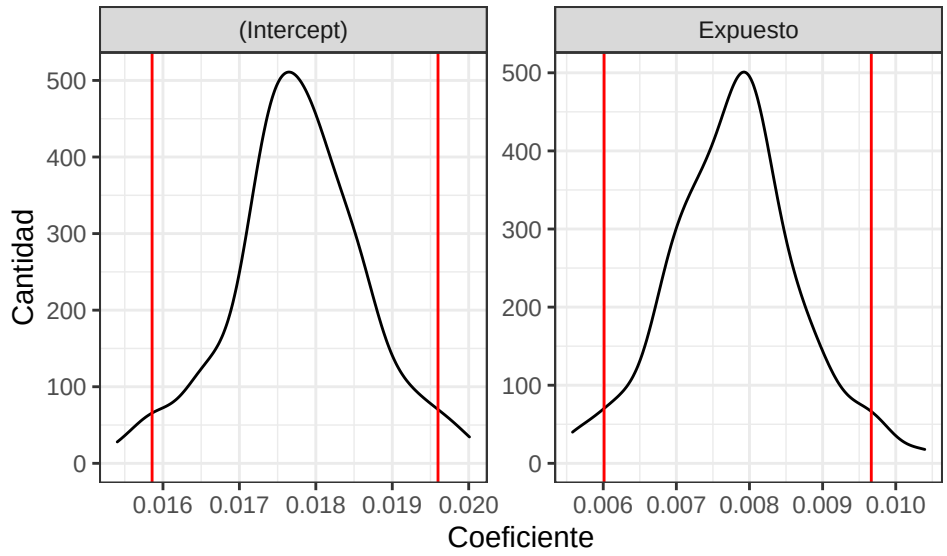
El coeficiente del modelo logístico no tiene una interpretación directa en cuanto a magnitud pero si por su signo. El modelo ajustado es:

$$\log\left[\frac{P(\widehat{\text{converted}}=1)}{1-P(\widehat{\text{converted}}=1)}\right] = -4.008 + 0.366(\text{user\_group}_{\text{Expuesto}}) \tag{4}$$

Un valor positivo de β_1 indica que estar expuesto a la campaña aumenta las probabilidades de comprar la cartera. Al exponenciar dicho valor se obtiene que estar expuesto a la campaña aumenta las probabilidades de compra 1.44 veces.

	Model 1
(Intercept)	−4.008
	[−4.106, −3.912]
	s.e. = 0.049
	t = −81.404
	p = 0.000
user_groupExpuesto	0.366
	[0.270, 0.466]
	s.e. = 0.050
	t = 7.330
	p = 0.000
Num.Obs.	588 101
AIC	138 477.8
BIC	138 500.4
Log.Lik.	−69 236.914
F	53.734
RMSE	0.49

Al aplicar también la 'tecnica del bootstrap sobre este modelo se puede observar en la Figura @ref(fig:histograma_coeficientes_logistico) que los valores de β_1 son positivos lo que indican la efectividad de la campaña.



3 Rentabilidad de la campaña

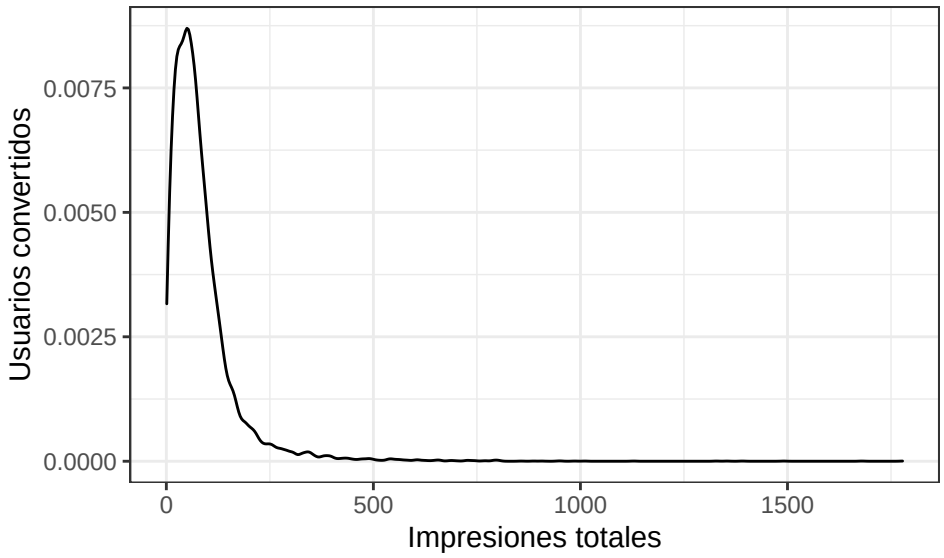
La rentabilidad de la campaña para TaskaBell podría pensarse como los usuarios expuestos que compraron las carteras. Al multiplicar la cantidad de usuarios convertidos por el valor que la firma le asigna a cada uno de ellos se obtiene un valor de USD 173719.3. Los costos de la campaña de marketing se estima a través de la cantidad de impresiones que fueron utilizadas. Si mil impresiones tienen un valor de 9 USD, la totalidad de impresiones usadas equivalen a USD 131374.6. El retorno de la inversión en la campaña de marketing, ROI, se estima como la diferencia entre el margen obtenido gracias a la campaña sobre los costos. La fórmula del ROI es la siguiente:

$$ROI = \frac{Ingresos - Costo}{Costos}$$

La rentabilidad de la campaña fue de 42344.65 USD mientras que el ROI fue de 32.232%. Basándose en ambas métricas la campaña de marketing fue rentable. El análisis económico se concluye con el costo de oportunidad de la campaña. El costo de oportunidad se entiende como aquellos usuarios que se encontraban en el grupo de control y que no fueron expuesto a las impresiones. Es decir, los clientes potenciales que se han perdido. Al tomar la fracción de usuarios del control que podrían haberse convertido con la campaña se obtiene que 181 usuarios, por un valor de 7238 USD podrían haber comprado las carteras.

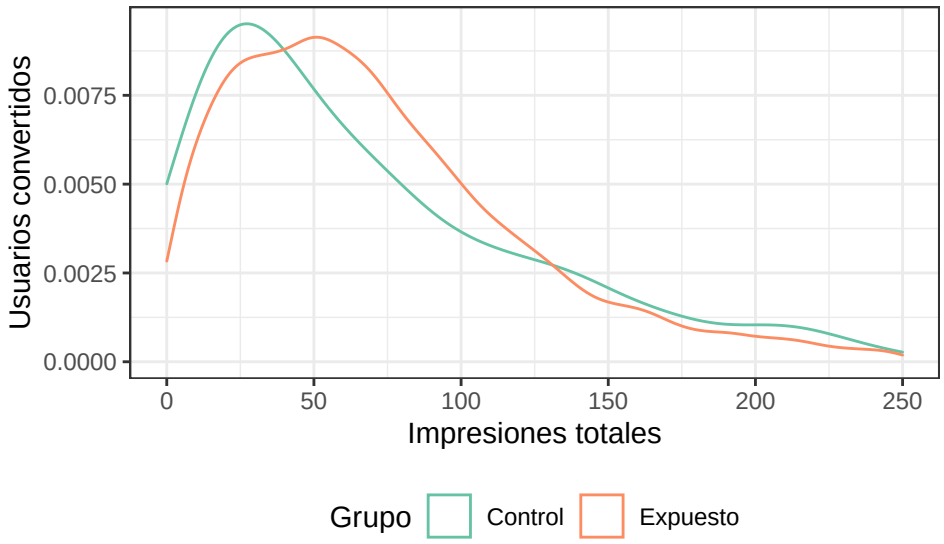
4 Análisis de efectividad de la campaña

La efectividad de la campaña se evalúa al comparar la cantidad de impresiones usadas con la cantidad de usuarios convertidos. La distribución de la cantidad de impresiones por usuarios se muestra en la Figura @ref(fig:impresiones_densidad).



La distribución del número total de impresiones que fueron necesarias para convertir usuarios es muy asimétrica con una cola pesada en el extremo derecho de la distribución. Sin embargo, es posible identificar un codo alrededor de 250 impresiones. Por encima de este valor la cantidad de usuario convertidos disminuye notablemente. Habiendo dicho esto, el análisis se centra en el rango de 1-250 impresiones que son las que tuvieron el mayor impacto durante la campaña. La Figura @ref(fig:densidad_conversion_impresiones_250) muestra los usuario convertidos por cantidad de impresiones para los grupos expuesto (línea roja) y control (línea verde).

```
## Warning: Removed 577 rows containing non-finite values (stat_density).
```



La distribución es bastante similar entre grupos y que la gran mayoría de las conversiones tiene lugar en el rango 1-150 impresiones. Para indagar aún más sobre la dinámica de la conversión se dividió a la variable *total de impresiones* en clases discretas separadas por intervalos de 25 impresiones, totalizando 11 clases. Las principales métricas para cada una de las clases se muestra en la Tabla **??**

Clases

Usuarios

Conversión

Total impresiones

Tasa de conversión

Impresiones por conversión

0-25

430022

2578

4008079

0.6

1555

26-50

89013

3124

3158395

3.5

1011

51-75

31334

3054

1920966

9.7

629

76-100

14668

2188

1270878

14.9

581

101-125

7802

1381

872680

17.7

632

126-150

4585

799

627297

17.4

785

151-175

2841

499

460736

17.6

923

176-200

1884

293

352699

15.6

1204

201-225

1371

220

290980

16.0

1323

226-250

965

130

229037

13.5

1762

251+

3616

577

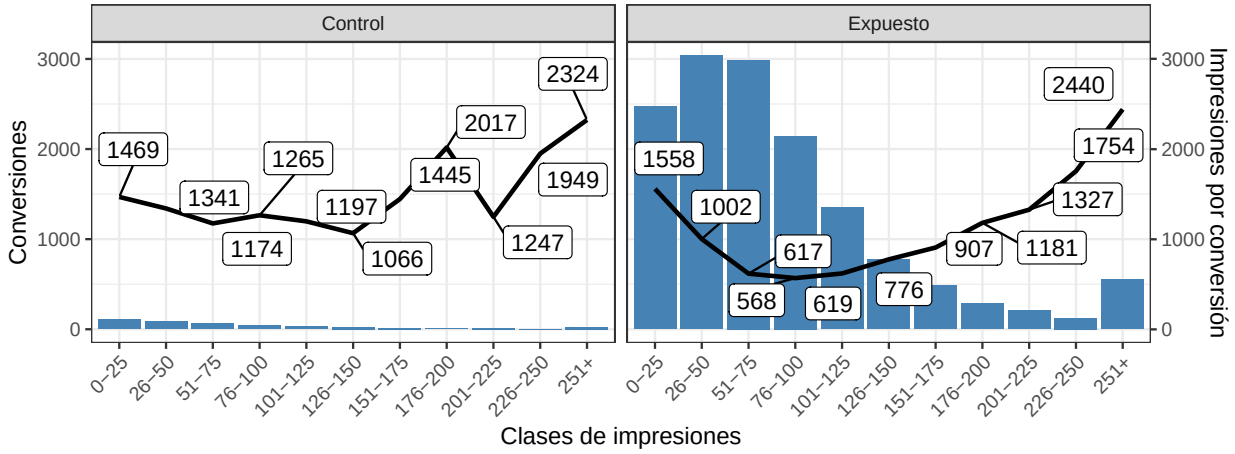
1405435

16.0

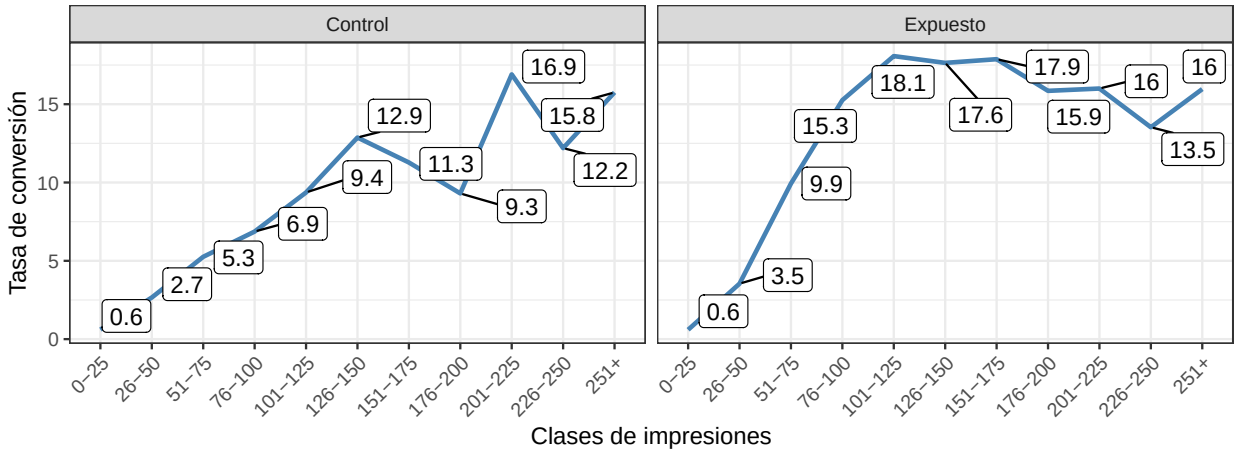
2436

La información mostrada en la Tabla se muestra también en la Figura , dónde las barras corresponden a la cantidad de conversiones por clase de número de impresiones, mientras que la línea negra corresponde a la cantidad de impresiones a las que deben ser expuestos esos usuarios convertirlos. Entre 1 y 100 impresiones la cantidad de usuarios convertidos es máxima y tiende a decrecer por encima de dicho umbral mostrando algunos síntomas de saturación. Por encima de las 100 impresiones por usuario la cantidad de impresiones por conversión aumenta en gran medida, la campaña pierda efectividad. Además de lo anterior, al incrementarse la cantidad de impresiones los costos lo hacen proporcionalmente.

`histograma_conversiones_totales`



En línea con lo anterior, la Figura @ref(fig:evolucion_tasa_conversion) muestra la dinámica de la tasa de conversión a medida que aumentan las impresiones por usuario.



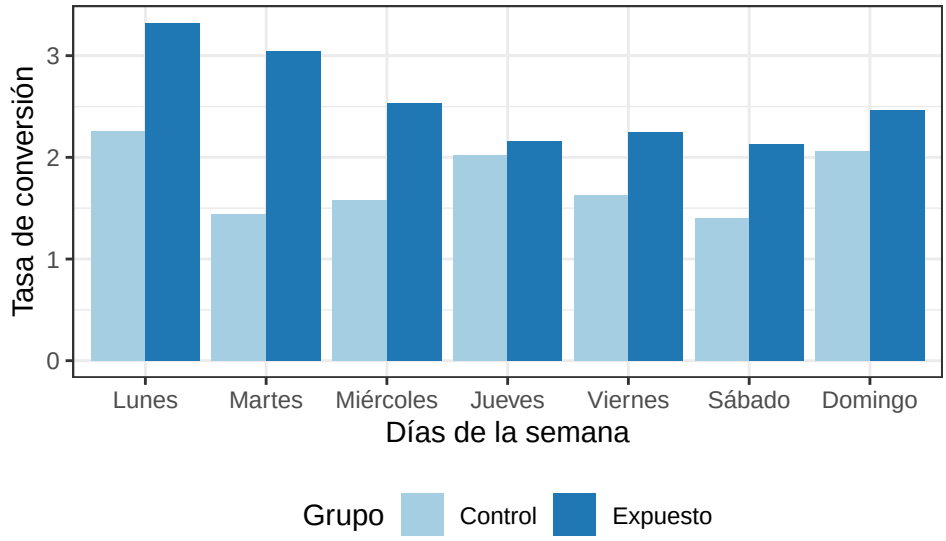
La tasa de conversión aumenta sostenidamente hasta alcanzar un máximo alrededor de las 100 impresiones por usuario. Esta cantidad de impresiones serían las óptimas para tener la mayor efectividad. Por encima de las 100-150 impresiones pareciera generarse un cierto hastío porque la tasa de conversión disminuye.

5 Efecto del momento de exposición

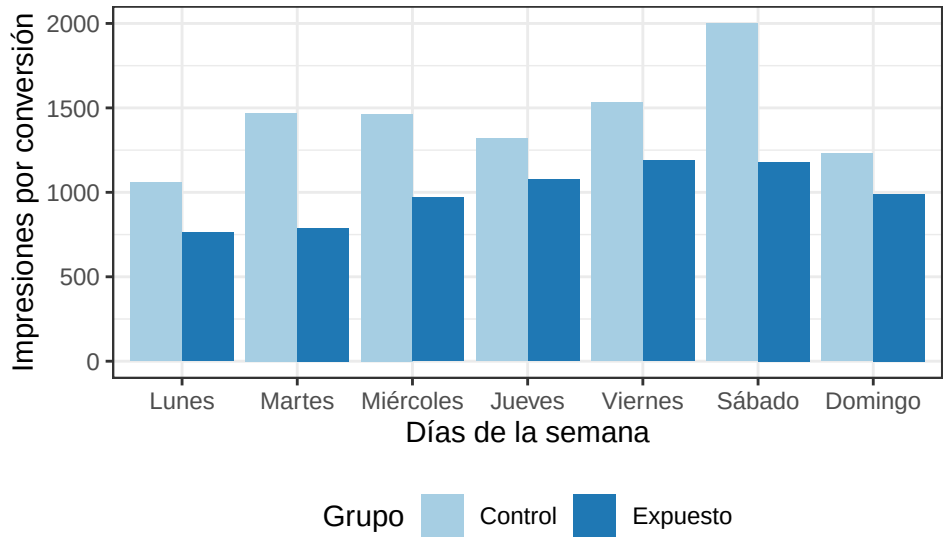
5.1 Día de la semana

La Figura @ref(fig:efecto_dia_semana) muestra la tasa de conversión por cada día de la semana para los dos grupos, Expuesto y Control. Se puede observar como la tasa de conversión para el grupo expuesto es

superior para el comienzo de la semana, días lunes y martes, y luego tiende a decrecer hacia el fin de semana para repuntar un poco el día domingo. El grupo control no muestra un comportamiento claro y es más variable.



Para evaluar la efectividad de la campaña se considera la variación del total de impresiones por conversión para día de la semana. La @ref(fig:efecto_dia_semana_efectividad) muestra la cantidad de impresiones que fueron necesarias para lograr la compra de la cartera por cada día de la semana para ambos grupos. Valores más bajos de esta variable indican una mayor efectividad de la campaña de marketing.

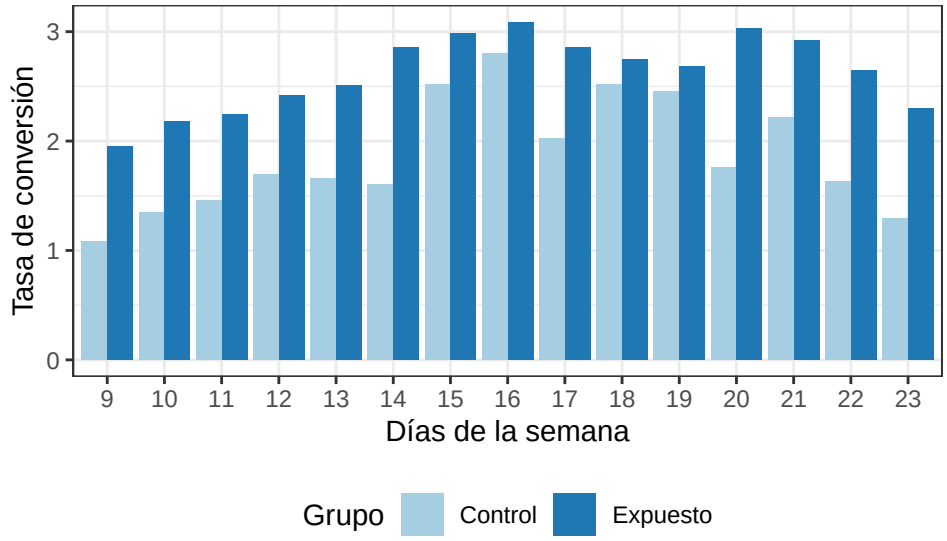


Nuevamente, los días lunes y martes fueron los días más efectivos de la semana, con una efectividad media de alrededor de 800 impresiones por conversión. En los días viernes y sábado la efectividad disminuyó a cerca de 1200 impresiones por conversión.

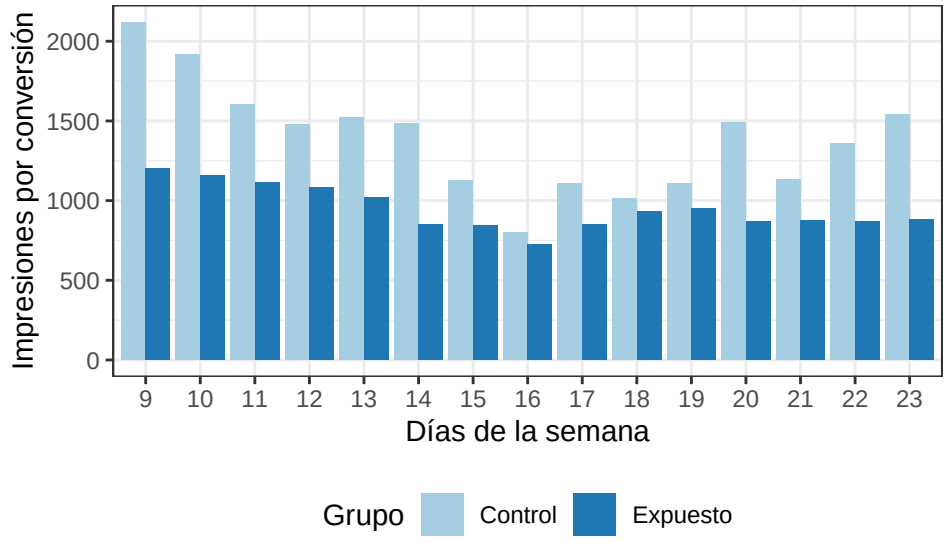
5.2 Hora del día

Siguiendo la misma lógica anterior se estimaron las mismas métricas por cada hora del día. La Figura **??** muestra la tasa de conversión por hora del día para cada uno de los grupos. Las mayores tasas de conversión

tienen lugar durante la tarde entre las 15:00 y las 21:00. En esta franja horaria la tasa se mantiene por encima del 2.5% con un pico de más del 3% a las 16:00. Contrariamente, aquellos usuarios que fueron expuestos durante la madrugada fueron los que menor tasa de conversión mostraron.



Para complementar el análisis es calculó la cantidad de impresiones por conversión para cada hora del día. La Figura @ref(fig:efecto_hora_efectividad) muestra la evolución de dicha métrica para cada uno de los grupos. La campaña fue menos eficiente durante la madrugada, en consonancia con lo dicho anteriormente, ya que se necesitaron una mayor cantidad de avisos por cada conversión. Lo contrario ocurrió durante la tarde donde fueron necesarios menos de 1000 impresiones por cada conversión.



6 Análisis de regresión

El modelo lineal utilizado intenta explicar el efecto del total de impresiones a las que es expuesto un usuario y el momento (día de la semana y hora del día) sobre la probabilidad de que éste compre la cartera. Basado en el análisis descripto en secciones anteriores se ajusto un primer modelo considerando las interacciones

existentes entre el total de impresiones y el día de la semana y entre el día de la semana y la hora del día. Esto se debe a que los usuarios no son indiferentes al número de avisos diarios a los que son expuestos durante los distintos días de la semana y también a que no es lo mismo observar un anuncio un día de mayor sensibilidad como los lunes en comparación como un día domingo. Además de las interacciones planteadas se considero la relación no lineal entre la conversión y el total de impresiones. Como se mostró en el análisis anterior la relación puede ser modelada como una función cuadrática.

Al considerar el modelo con interacciones la cantidad de coeficientes del modelo fue muy importante y, en general, no fueron significativos. Esto da la pauta que si bien el análisis exploratorio dio pistas al respecto el supuesto de interacción no se verificó con el modelo lineal. Esto puede deberse a que el modelo lineal no fue lo suficientemente flexible como para capturar el comportamiento mencionado.

Dado que los coeficientes de la interacción no fueron significativos se eliminaron del modelo y se ajustó nuevamente sólo considerando el total de impresiones y las variables dummy para el día de la semana y la hora del día. El modelo estadístico lineal probabilístico es el siguiente:

$$\begin{aligned}\text{converted} = {} & \beta_0 + \beta_1(\text{tot\_impr}) + \beta_2(\text{tot\_impr\^{}2}) + \beta_3(\text{dia}_{\text{Martes}}) + \\ & \beta_4(\text{dia}_{\text{Miércoles}}) + \beta_5(\text{dia}_{\text{Jueves}}) + \beta_6(\text{dia}_{\text{Viernes}}) + \beta_7(\text{dia}_{\text{Sábado}}) + \\ & \beta_8(\text{dia}_{\text{Domingo}}) + \beta_9(\text{hora}_1) + \beta_{10}(\text{hora}_2) + \beta_{11}(\text{hora}_3) + \\ & \beta_{12}(\text{hora}_4) + \beta_{13}(\text{hora}_5) + \beta_{14}(\text{hora}_6) + \beta_{15}(\text{hora}_7) + \\ & \beta_{16}(\text{hora}_8) + \beta_{17}(\text{hora}_9) + \beta_{18}(\text{hora}_{10}) + \beta_{19}(\text{hora}_{11}) + \\ & \beta_{20}(\text{hora}_{12}) + \beta_{21}(\text{hora}_{13}) + \beta_{22}(\text{hora}_{14}) + \beta_{23}(\text{hora}_{15}) + \\ & \beta_{24}(\text{hora}_{16}) + \beta_{25}(\text{hora}_{17}) + \beta_{26}(\text{hora}_{18}) + \beta_{27}(\text{hora}_{19}) + \\ & \beta_{28}(\text{hora}_{20}) + \beta_{29}(\text{hora}_{21}) + \beta_{30}(\text{hora}_{22}) + \beta_{31}(\text{hora}_{23}) + \\ & \epsilon\end{aligned} \tag{5}$$

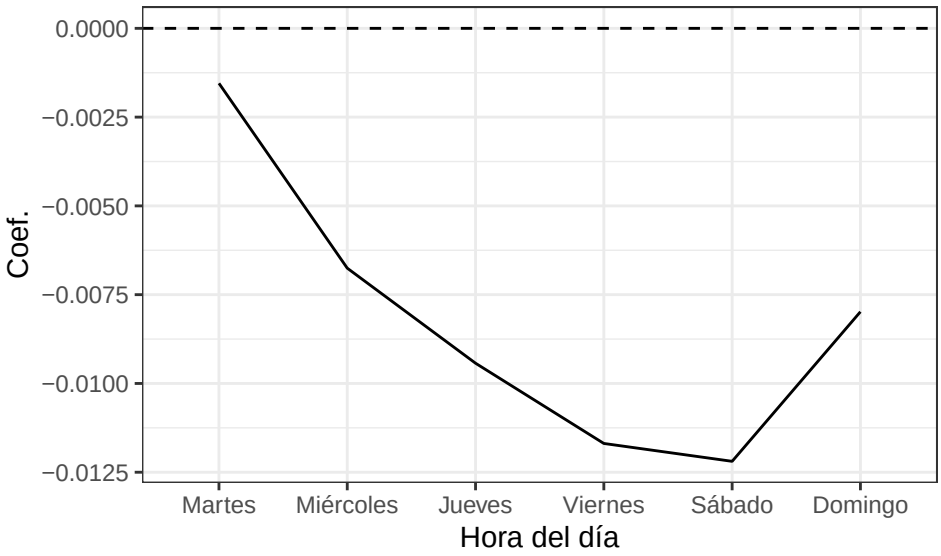
El modelo ajustado por MCO es:

$$\begin{aligned}\hat{\text{converted}} = {} & -0.000108 + 0.001239(\text{tot\_impr}) - 1e-06(\text{tot\_impr\^{}2}) - 0.001548(\text{dia}_{\text{Martes}}) - \\ & 0.006753(\text{dia}_{\text{Miércoles}}) - 0.00943(\text{dia}_{\text{Jueves}}) - 0.01169(\text{dia}_{\text{Viernes}}) - 0.012193(\text{dia}_{\text{Sábado}}) - \\ & 0.007981(\text{dia}_{\text{Domingo}}) - 0.007966(\text{hora}_1) - 0.012649(\text{hora}_2) - 0.007434(\text{hora}_3) - \\ & 0.013051(\text{hora}_4) - 0.015769(\text{hora}_5) - 0.013787(\text{hora}_6) - 0.007349(\text{hora}_7) + \\ & 0.000196(\text{hora}_8) - 0.00021(\text{hora}_9) + 0.000806(\text{hora}_{10}) + 0.00169(\text{hora}_{11}) + \\ & 0.002084(\text{hora}_{12}) + 0.003287(\text{hora}_{13}) + 0.008047(\text{hora}_{14}) + 0.008786(\text{hora}_{15}) + \\ & 0.012626(\text{hora}_{16}) + 0.008309(\text{hora}_{17}) + 0.006357(\text{hora}_{18}) + 0.005722(\text{hora}_{19}) + \\ & 0.007817(\text{hora}_{20}) + 0.007533(\text{hora}_{21}) + 0.007657(\text{hora}_{22}) + 0.007056(\text{hora}_{23})\end{aligned} \tag{6}$$

El modelo lineal es estadísticamente significativo.

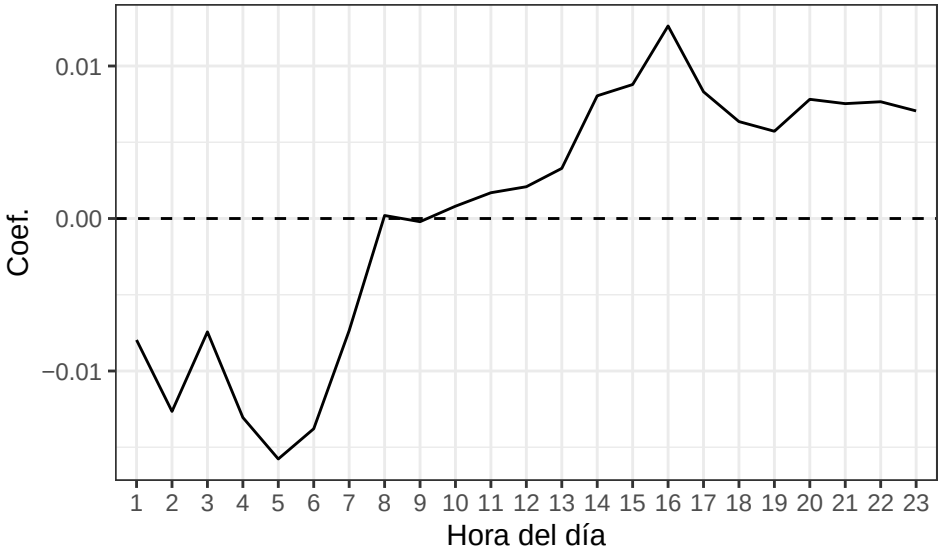
	Model 1
(Intercept)	0.000
	[−0.004, 0.004]
	p = 0.959
tot_impr	0.001
	[0.001, 0.001]
	p = 0.000
I(tot_impr^2)	0.000
	[0.000, 0.000]
	p = 0.000
diaMartes	−0.002
	[−0.003, 0.000]
	p = 0.039
diaMiércoles	−0.007
	[−0.008, −0.005]
	p = 0.000
diaJueves	−0.009
	[−0.011, −0.008]
	p = 0.000
diaViernes	−0.012
	[−0.013, −0.010]
	p = 0.000
diaSábado	−0.012
	[−0.014, −0.011]
	p = 0.000
diaDomingo	−0.008
	[−0.009, −0.007]
	p = 0.000
hora1	−0.008
	[−0.014, −0.002]
	p = 0.008
hora2	−0.013
	[−0.018, −0.007]
	p = 0.000
hora3	−0.007
	[−0.014, 0.000]
	p = 0.038
hora4	−0.013
	[−0.025, −0.001]
	p = 0.030
hora5	−0.016
	[−0.027, −0.004]
	p = 0.007
hora6	−0.014
	[−0.021, −0.006]
	p = 0.000
hora7	−0.007
	[−0.013, −0.002]
	p = 0.008
hora8	0.000
	[−0.004, 0.005]
	p = 0.933
hora9	0.000
	[−0.005, 0.004]
	p = 0.925
hora10	0.001
	[−0.003, 0.005]
	p = 0.712
hora11	0.002
	[−0.003, 0.006]

Para analizar el efecto del día de la semana sobre la conversión se graficaron los coeficientes del modelo lineal. La Figura @ref(fig:coeficiente_dia_semana) muestra los coeficientes para cada uno de los días de la semana a excepción del día lunes que es considerado la base.



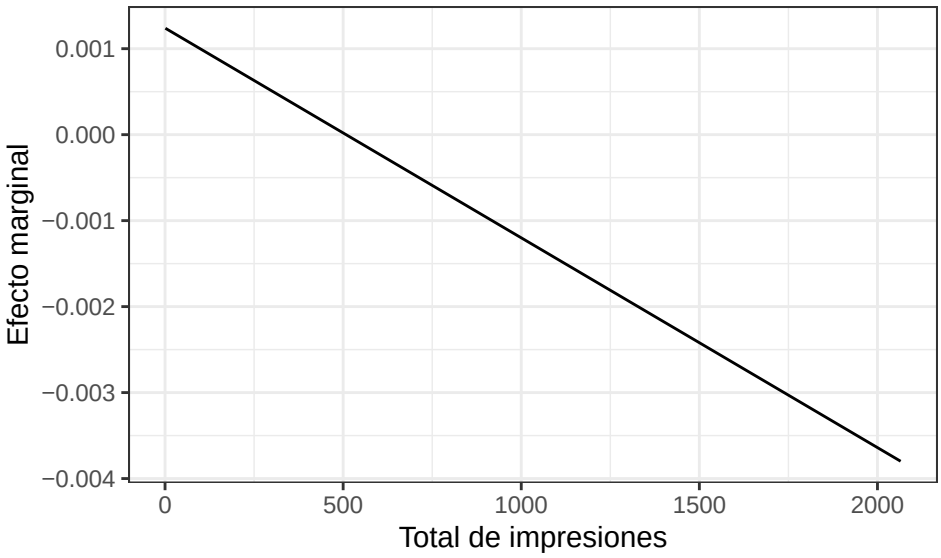
Dado que el lunes es el día con mayor probabilidad de conversión todos los coeficientes son negativos. Como se mostró anteriormenete, la tasa de conversión disminuía hacia el fin de semana con un ligero repunto el día domingo tal como se muestra en la Figura.

Con las horas del día se realizo el mismo análisis y se muestra en la Figura @ref(fig:coeficiente_hora_dia)



Este resultado es consecuente con los mostrado con valores negativos para la madrugada, los momentos con menores probabilidades de conversión, y luego creciendo para alcanzar el el pico a las 16 horas, con una probabilidad mayor al 1% de comprar la cartera con respecto a la base de la medianoche.

La variable total de impresiones fue evaluada mediante el efecto marginal calculado a partir de la derivada primera de la función cuadrática que modela su comportamiento. El resultado se muestra en la Figura @ref(fig:efecto_maringal_impresiones)


0.001
0.000
−0.001
−0.002
−0.003
−0.004
Efecto marginal
0
500
1000
1500
2000
Total de impresiones