SOLUCIÓN DEL CASO EUROMOTORS

Por Edwin Santos Vidal

GRUPO
EUROMOTORS

1. Planteamiento del problema

- Se presentó un dataset de 5000 clientes quienes respondieron una encuesta sobre algunas de sus preferencias y comportamientos.
- Se planteó que, a partir de esta encuesta, se pueda saber cómo son los clientes de autos de lujo y cómo hacer una campaña de marketing efectiva.

2. Herramienta de trabajo

- El análisis se realizó en el lenguaje de programación R, ya que es muy versátil para el manejo estadístico. Todo el código estará colgado en rpubs por si desearan revisar el proceso de solución.

Ve a la solución

3. Procedimiento de solución

Los pasos que se siguieron fueron los siguientes:

i. *Limpieza de la data*
ii. *Exploración de la data*
iii. *Construcción de un modelo de regresión*
iv. *Revisión de resultados*

i. Limpieza de la data

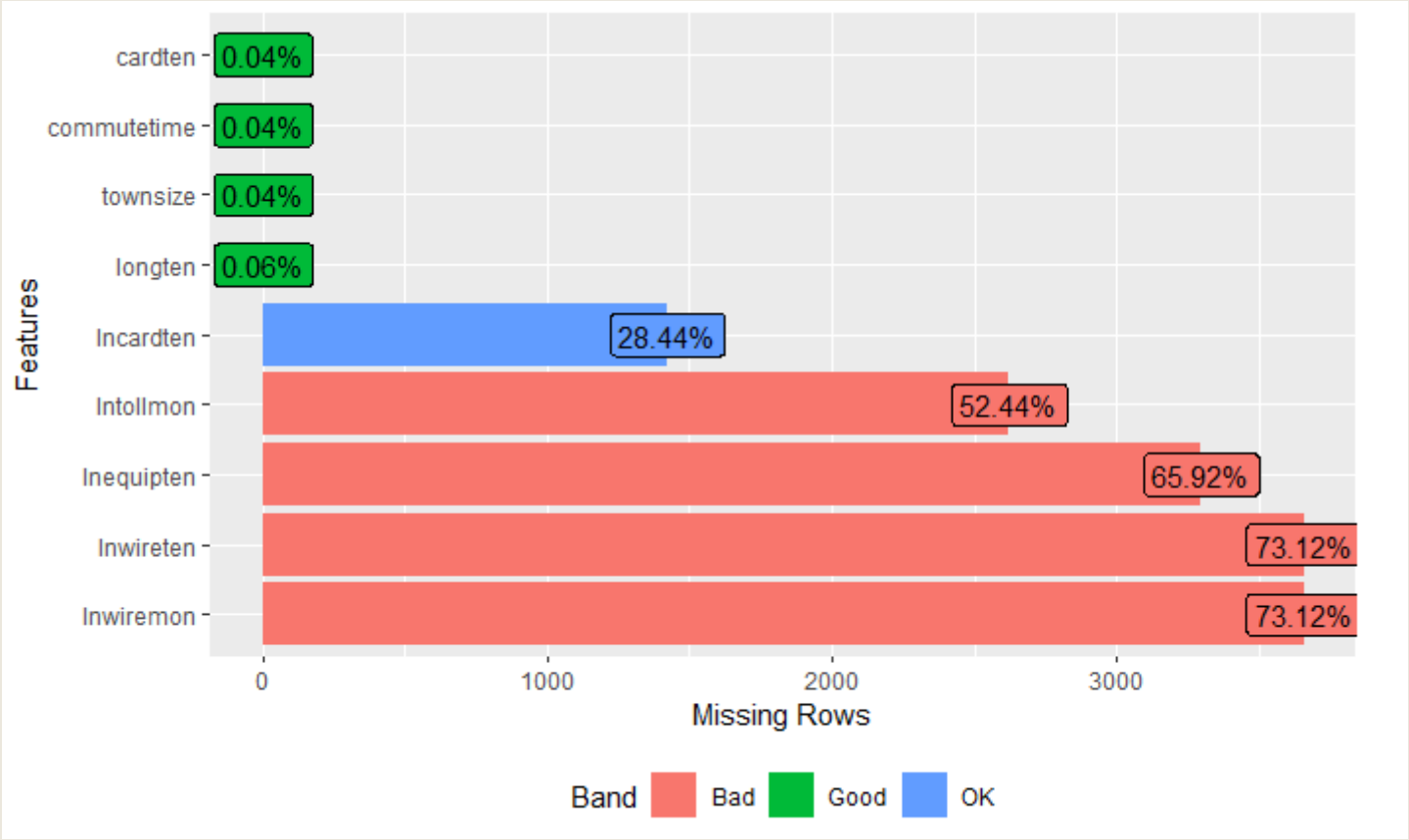
- Primero se buscó si habían duplicados en el dataset y se encontró que no hubo ninguno

Hide

```
n_occur<-data.frame(table(db$custid))
n_occur[n_occur$Freq>1,]
```

0 rows

- Lo siguiente que se hizo fue buscar si es que había columnas con valores nulos, encontrándose lo siguiente:



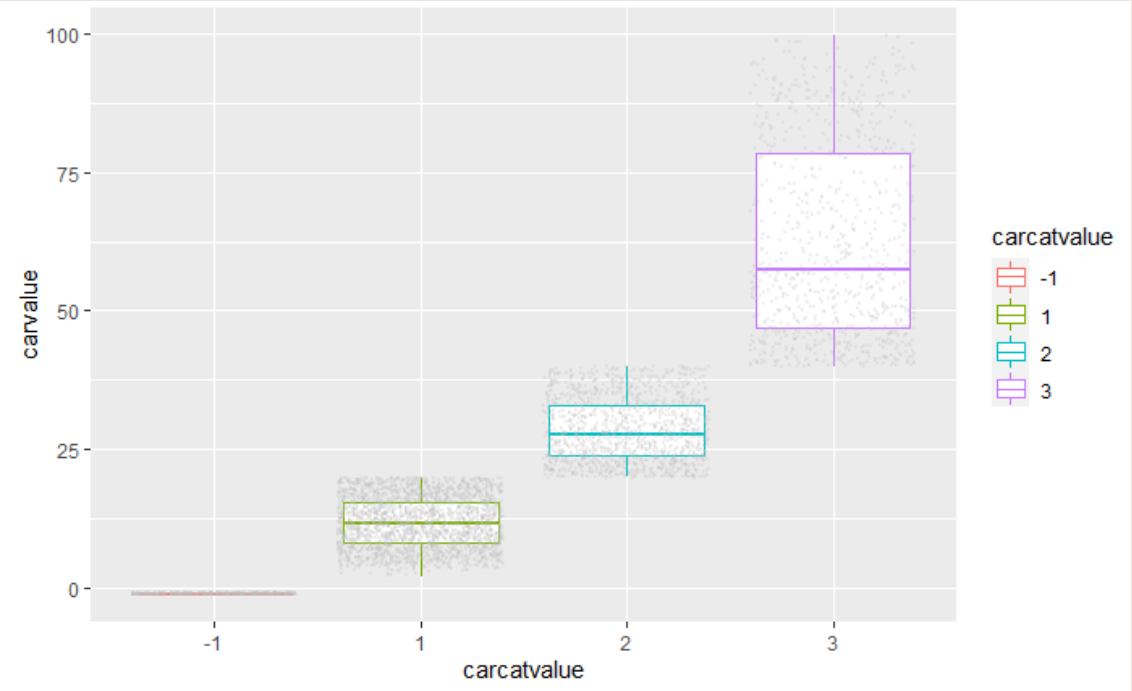
- Como se puede ver, hay una gran cantidad de valores nulos en las columnas lncardten, intollmon, inequipten, inwireten e lnwiremon, por lo que se pasan a descartar. Las otras columnas se imputaron mediante kNN.

ii. Exploración de los datos

- Para la exploración de datos se hicieron varios gráficos, los cuales servirán como punto de partida para el planteamiento del modelo.

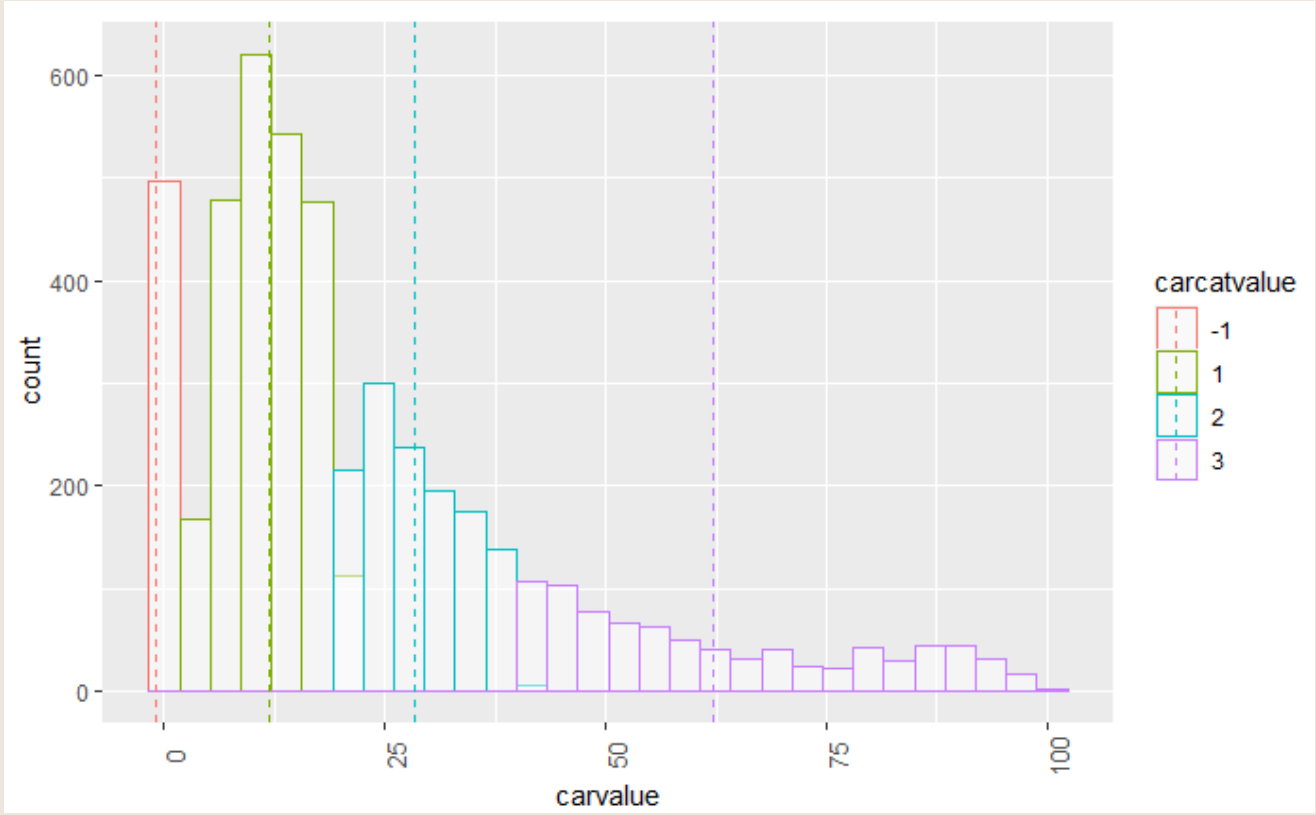
Valor del Auto VS Categoría de Valor

- Podemos ver que los datos para la categoría de lujo están bastante más dispersos que en las otras categorías. Además, su mediana tiende a la baja.



Frecuencia de categoría de valor

- Podemos ver que a medida que crece el valor del auto, decrece la cantidad de personas que lo tienen.



Comparación de datos personales de personas que tienen autos de lujo contra las que no las tienen

- Se observa que las personas que tienen autos de lujo son mayores en promedio (54 contra 45 años). Adicionalmente, sus años de educación, al igual que los de su consorte, son un poco mayores en promedio.

lujo_si <fctr>	mean(age) <dbl>	median(age) <int>	mean(ed) <dbl>	mean(spoused) <dbl>
0	45.58059	43	14.36368	6.000721
1	54.21266	55	15.43489	6.670251

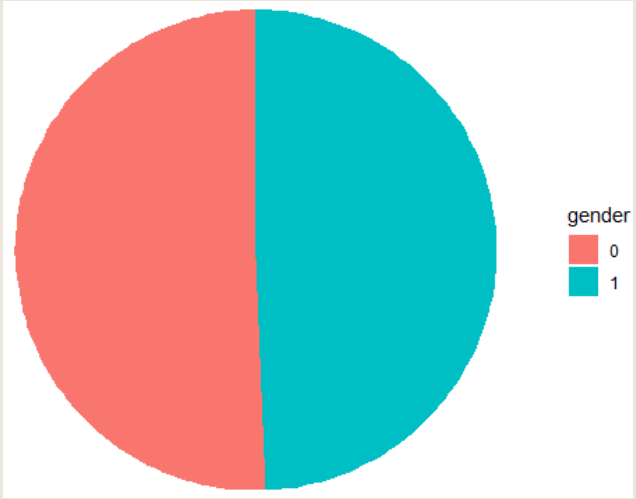
2 rows

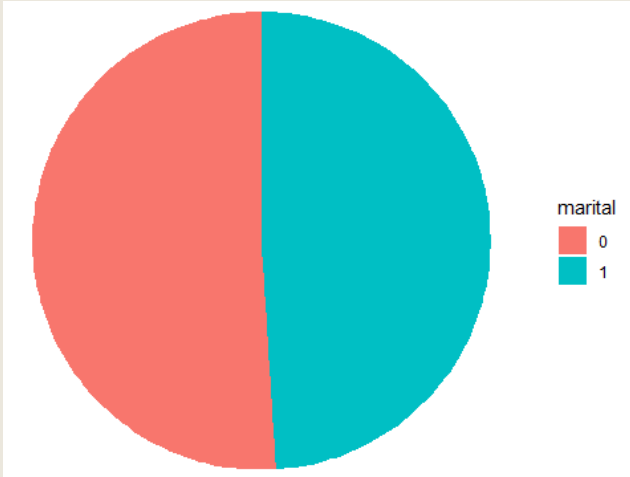
	jobcat <fctr>	0 <int>	1 <int>
1	1	1138	250
2	2	1520	120
3	3	478	142
4	4	170	42
5	5	346	106
6	6	511	177

6 rows

Distribución de género y estado civil de personas con autos de lujo

- Se observa que no hay una gran diferencia en cuanto al género y al estado civil de las personas que tienen autos de lujo





Datos de empleo e ingresos

- Otro rasgo característico es que tienen unos sueldos muy altos (más de 136 000 al año) y están en el mismo trabajo desde hace más de 16 años de media (employ). El débito en sus tarjetas de crédito también es mayor, así como su gasto.
- Adicionalmente, los tipos de tarjeta 2 y 4 están correlacionados negativamente con tener un auto de lujo.

lujo_si <fctr>	mean(income) <dbl>	mean(employ) <dbl>	mean(creddebt) <dbl>	mean(cardspent) <dbl>
0	38.32933	8.44271	1.275004	308.4913
1	136.47909	16.13501	4.753626	480.0039

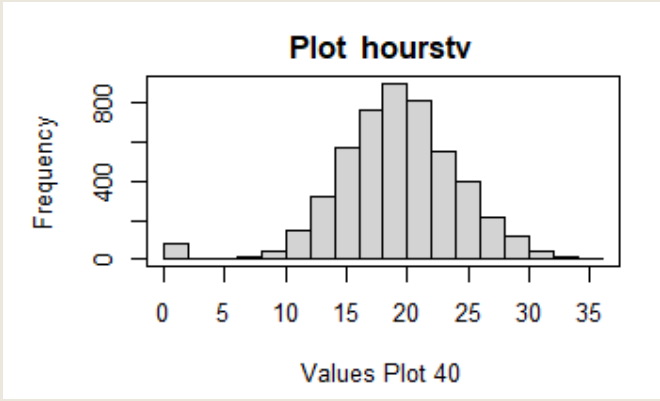
2 rows

	cardtype <fctr>	0 <int>	1 <int>
1	1	1015	227
2	2	1055	186
3	3	1034	223
4	4	1059	201

4 rows

Datos de empleo e ingresos

- Tanto personas con autos lujosos como las que no ven un alto promedio de horas por semana.
- Se puede observar claramente que la mayoría de personas con autos lujosos están suscritas al periódico.



lujo_si <fctr>	mean(hourstv) <dbl>
0	19.58059
1	19.96535

2 rows

	news <fctr>	0 <int>	1 <int>
1	0	2308	329
2	1	1855	508

2 rows

iii. Planteamiento del modelo

- Se planteó realizar un modelo de **regresión logística** para predecir en base a las variables seleccionadas la posibilidad de que una persona posea un auto de lujo. Esto permitirá comprobar que las variables seleccionadas sean las más adecuadas para el caso planteado. Para esto se dividieron los datos en train y test

```
#Se dividen los datos para entrenamiento (75% de train y 25% de test)

set.seed(88)
split=sample.split(db$lujo_si,SplitRatio = 0.75)


#Crear el training y testing data sets

dt=subset(db,split==TRUE)  #Train
de=subset(db,split==FALSE) #Test
```

- Después de probar con muchas variables se seleccionaron las siguientes:

Atributo	Descripción	Tipo de dato
Cardspent	Cantidad cargada a la tarjeta primaria el mes pasado	Numérico
Cardtype	Tipo de tarjeta de crédito principal	Categorico
ebill	Facturación electrónica	Categorico
pets_reptiles	Número de reptiles poseídos	Numérico
forward	Desvío de llamadas	Categorico
income	Ingresos familiares en miles	Numérico
jobcat	Categoría laboral	Categorico
spoused	Años de educación del consorte	Numérico
tollfree	Servicio de llamadas gratuitas	Categorico

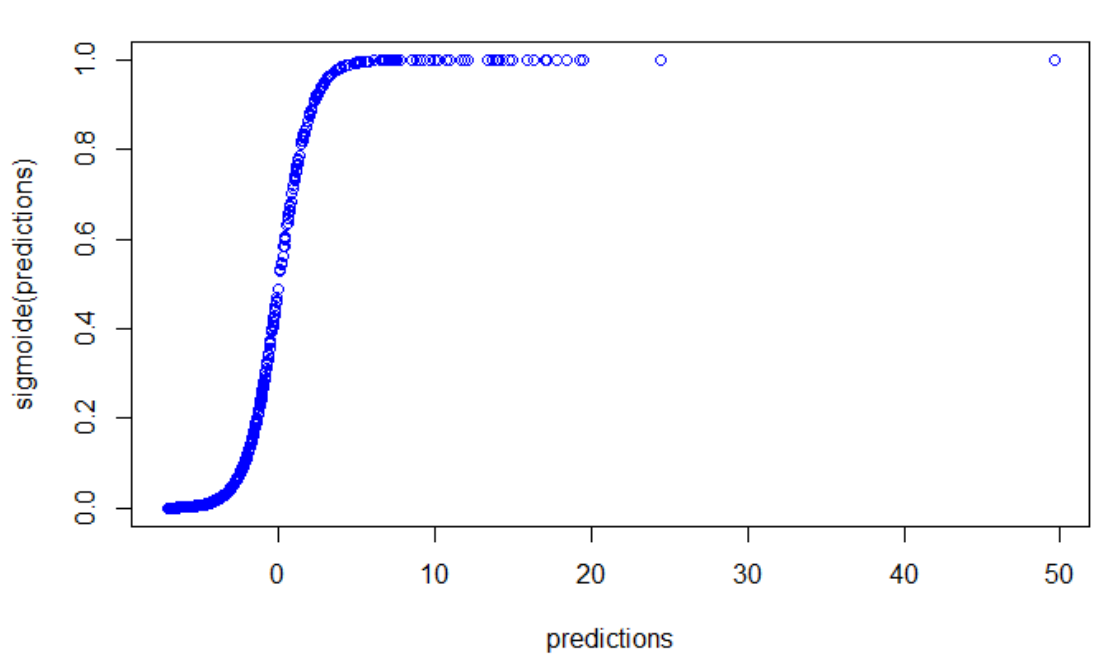
- Con lo que se obtuvo el siguiente modelo:

```
Call:
glm(formula = lujo_si ~ income + tollfree + cardtype + forward +
    ebill + cardspent + spoused + forward + jobcat + pets_reptiles,
    family = binomial(), data = dt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7045  -0.2479  -0.1289  -0.0767   2.3865
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.5825378  0.3171776 -20.753  < 2e-16 ***
income         0.0724357  0.0030602  23.670  < 2e-16 ***
tollfree1     -0.6437073  0.1912504  -3.366 0.000763 ***
cardtype2     -0.4858296  0.2064148  -2.354 0.018590 *
cardtype3     -0.1285183  0.1970437  -0.652 0.514251
cardtype4     -0.5206139  0.2084453  -2.498 0.012504 *
forward1       0.4726924  0.1908615   2.477 0.013263 *
ebill1        -0.4204455  0.1577652  -2.665 0.007699 **
cardspent      0.0005434  0.0002685   2.024 0.042966 *
spoused        0.0173578  0.0091474   1.898 0.057755 .
jobcat2       -0.0814042  0.2051047  -0.397 0.691448
jobcat3        0.1716952  0.2328781   0.737 0.460955
jobcat4        0.0304426  0.3767898   0.081 0.935605
jobcat5        0.1700695  0.2705825   0.629 0.529656
jobcat6        0.4527598  0.2212254   2.047 0.040697 *
pets_reptiles -0.4833516  0.2793629  -1.730 0.083596 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

iv. Revisión de resultados

- Se realiza una gráfica sigmoide para revisar cómo es que se están comportando las predicciones.



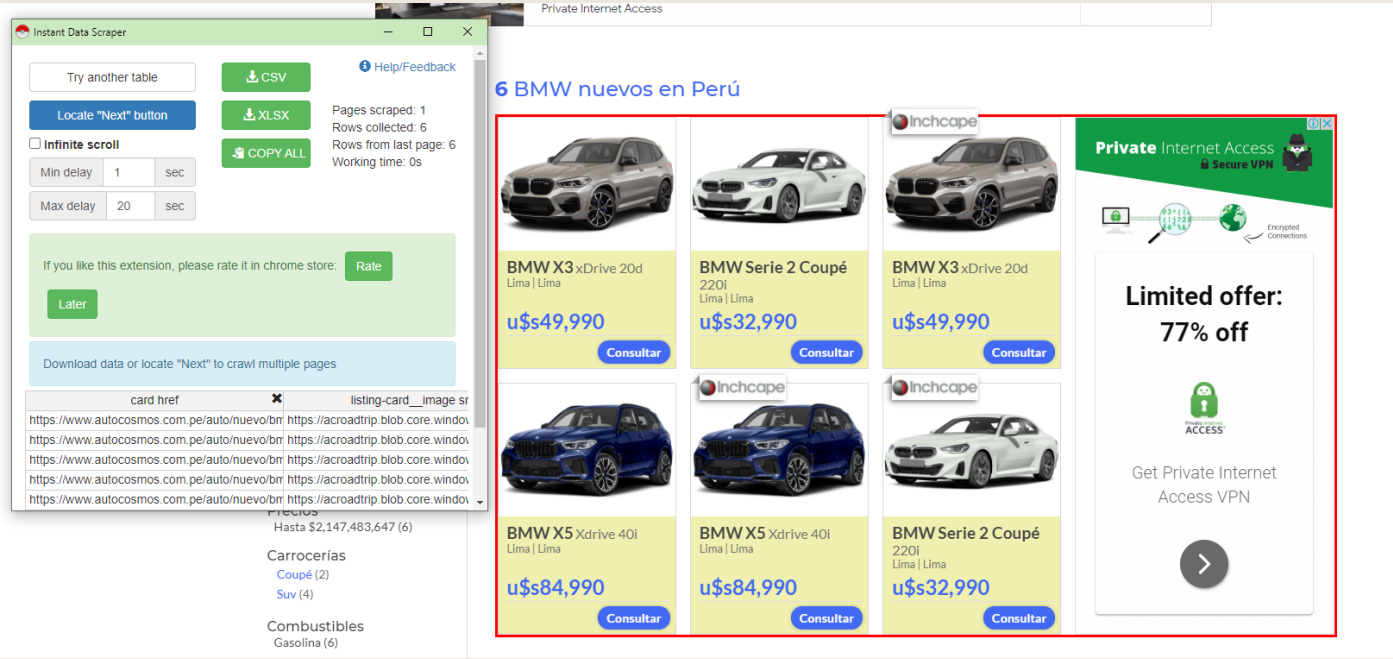
- Entonces se puede observar los resultados en la matriz de confusión

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1023   64
         1   18  145

               Accuracy : 0.9344
                 95% CI : (0.9192, 0.9475)
    No Information Rate : 0.8328
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7417

 Mcnemar's Test P-Value : 6.715e-07

            Sensitivity : 0.9827
            Specificity : 0.6938
         Pos Pred Value : 0.9411
         Neg Pred Value : 0.8896
             Prevalence : 0.8328
         Detection Rate : 0.8184
   Detection Prevalence : 0.8696
      Balanced Accuracy : 0.8382

       'Positive' Class : 0
```

Más de 93% de eficiencia de predicción del modelo

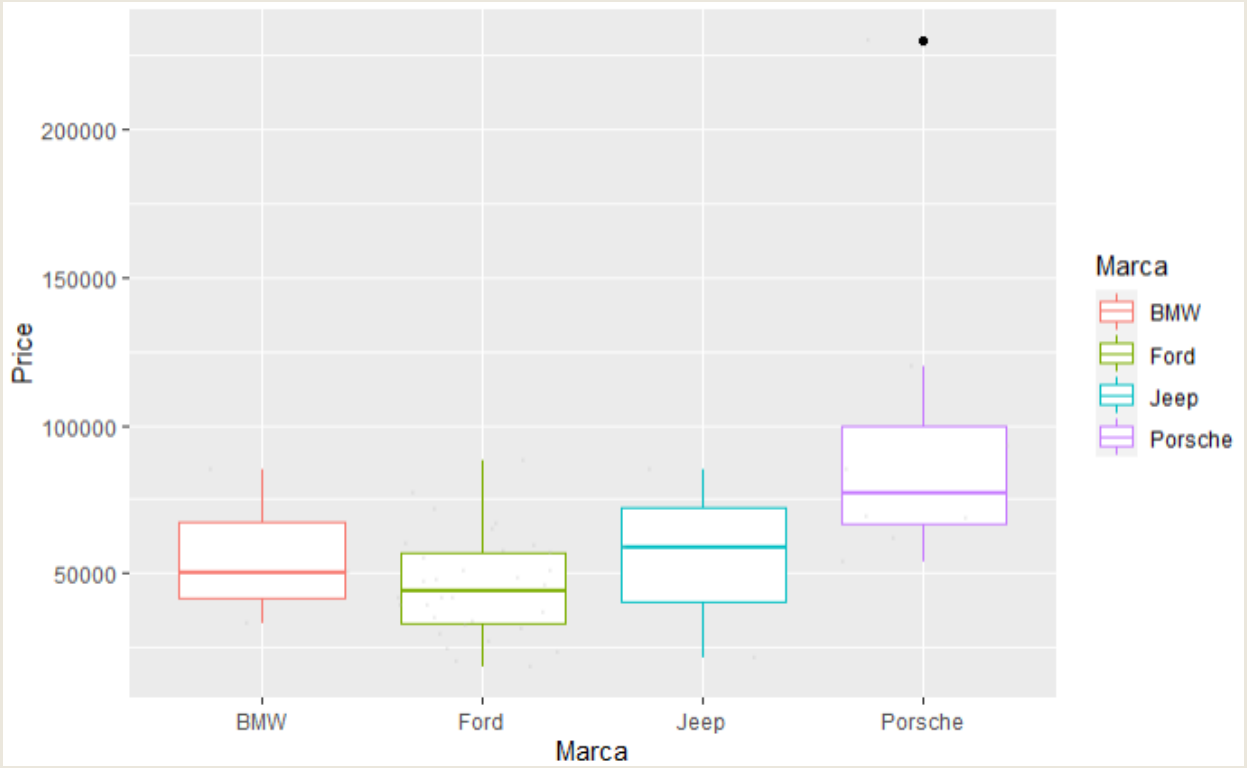
4. Comparación de precios de autos con los presentados en el caso

- Para la comparación de precios se hizo un scrape de páginas peruanas como neoauto o autocosmos de solo marcas de lujo.



Comparación de autos de lujo del caso vs datos reales

- Se puede ver que la media y mediana están muy cercanas entre ambos datasets.
- Además el rango es también muy similar (dejando de lado los outliers).



scrape

mean(Price) <dbl>	median(Price) <dbl>	min(Price) <dbl>	max(Price) <dbl>
55976.52	50490	18490	230000

1 row

dataset

mean(carvalue) <dbl>	median(carvalue) <dbl>	min(carvalue) <dbl>	max(carvalue) <dbl>
62.27706	57.5	40	99.6

1 row

5. Respuestas directas a las preguntas planteadas inicialmente

- 10 variables más importantes.
  - *Las 10 variables más importantes según lo revisado son: cardspent, cardtype, ebill, pets_reptiles, forward, income, jobcat, spoused y tollfree.*
- ¿Son los precios similares a la realidad?
  - *Sí, son bastante similares a la realidad, según lo revisado en el scrape.*

- Variables relevantes para la campaña de marketing
  - *Las variables más relevantes son las que están relacionadas con el nivel de ingresos y la cantidad de gasto de sus tarjetas; no obstante, para plantear una campaña de marketing, también es necesario tener en cuenta los datos generales del público objetivo como la edad, sexo, etc. Asimismo, sería importante saber los canales de información que más utilizan diariamente, para así poder dirigir mejor la publicidad. En este caso, según lo revisado, los canales de información más relevantes son la tv y el periódico, por lo que convendría realizar la publicidad por ahí.*
- ¿Cómo confirmarías que la propuesta es la adecuada?
  - *En este caso sería revisando el nivel de ventas vs el gasto en la publicidad (ROAS). Si este incrementa, entonces la propuesta fue la adecuada.*

¡Muchas gracias!