Universidad Autónoma de Yucatán

Maestría en Ciencias de la Computación

Métodos Estadísticos de Machine Learning

Proyecto 1

Autor: Mario Herrera Almira

11 de febrero del 2023

En la página web https://datasciencedojo.com/blog/datasets-data-science-skills/?utm_campaign=DSD%20blogs%202022&utm_content=223545409&utm_medium=social&utm_source=twitter&hss_channel=tw-1318985240 se presenta un repositorio de datos, que se pueden utilizar con el objetivo de mejorar las habilidades en el área de Ciencia de Datos.

Para este proyecto van a tomar la base de datos 23 y proponer un modelo de regresión lineal para el problema propuesto.

Se debe escribir un reporte que incluya al menos la siguiente información:

1. Un análisis descriptivo de los datos, que incluya:
    a. Identificación de la variable dependiente y la independiente (5 puntos)
    b. Identificar la relación que existe entre la variable dependiente y cada variable independiente (15 puntos)
2. El modelo de regresión estimado: su ecuación y el método utilizado para llegar al modelo (10 puntos)
3. Determinar cuáles variables son significativas en presencia de las otras en el modelo estimado (10 puntos)
4. Evaluar el modelo: Presenta el valor del R^2 y su interpretación. Presenta el valor del MSE. (10 puntos)

Respuesta

Para resolver este problema solamente se tiene en cuenta la base de datos de “Hours”, la base de datos “Day” no se tiene en cuenta porque contiene la misma información que “Hours” pero con una columna menos que es la de “hr”, esta columna en mi opinión si es importante tenerla en cuenta porque la hora del día si puede ser influyente sobre la cantidad de bicicletas que se pueden rentar, ya que hay horarios del día en los que normalmente las personas no rentan bicicletas.

1-)

a-)

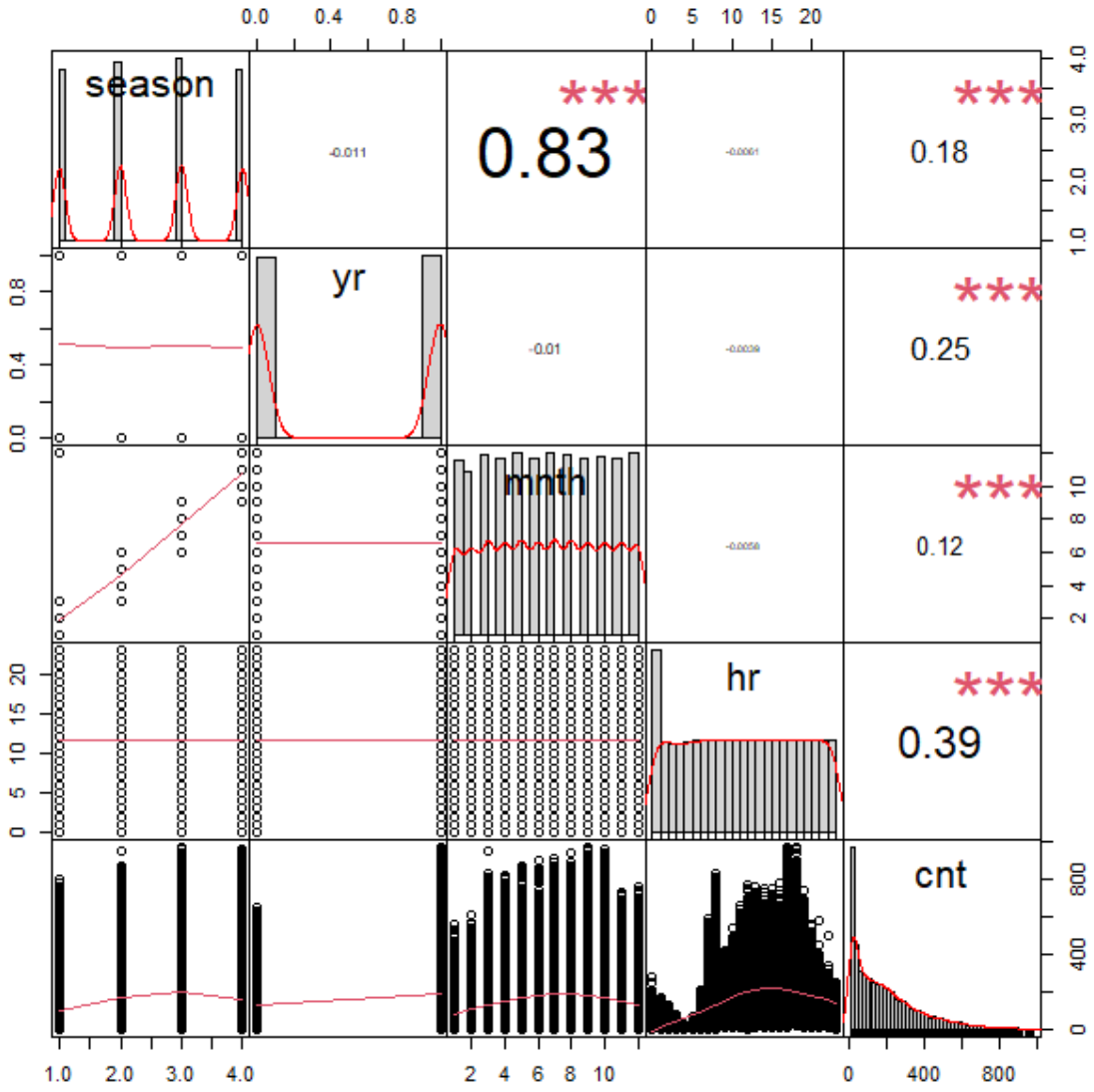
La **variable dependiente** es la cantidad de bicicletas rentadas (cnt) ya que es el valor que se quiere redecir en este ejercicio utilizando regresión lineal y además es la única que puede depender de cómo se comporten el resto de las variables.

Las **variables independientes** son todas las demás sin contar a instant, dteday, casual ni registered. Instant no se tiene en cuenta porque esta variable es solo un valor de identificación para cada uno de los registros, no tiene ninguna influencia en el comportamiento de los datos. Dteday también es una variable que no aporta información sobre el comportamiento d ellos datos ya que solo registra la fecha en la que se hizo la operación, además que el año y el mes ya se encuentran en columnas independientes. Las variables casual y registered no son factores medio

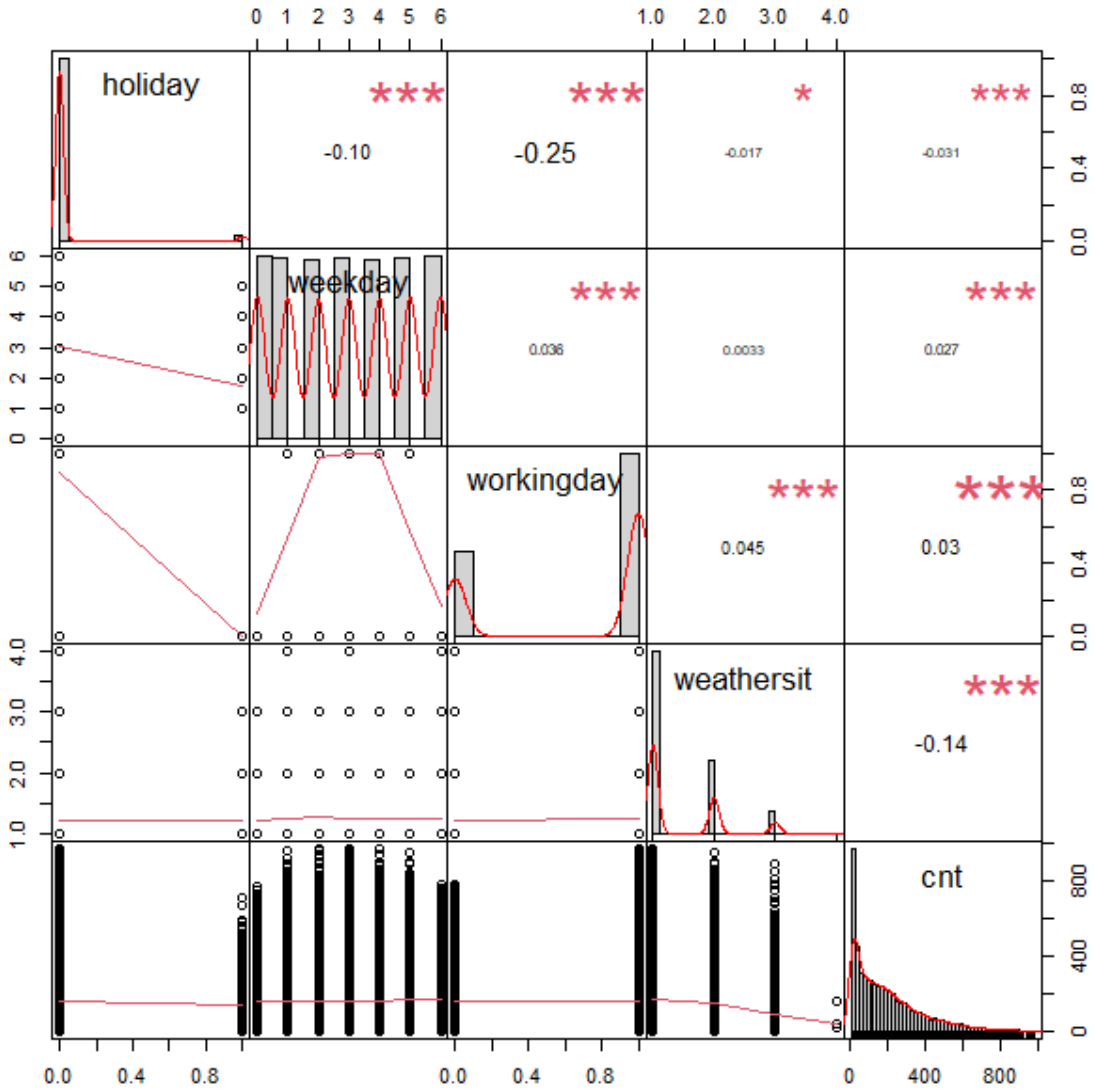
ambientales que influyan en la la cantidad de bicicletas que se rentan por lo que no se tienen en cuenta tampoco para el modelo.

b-)

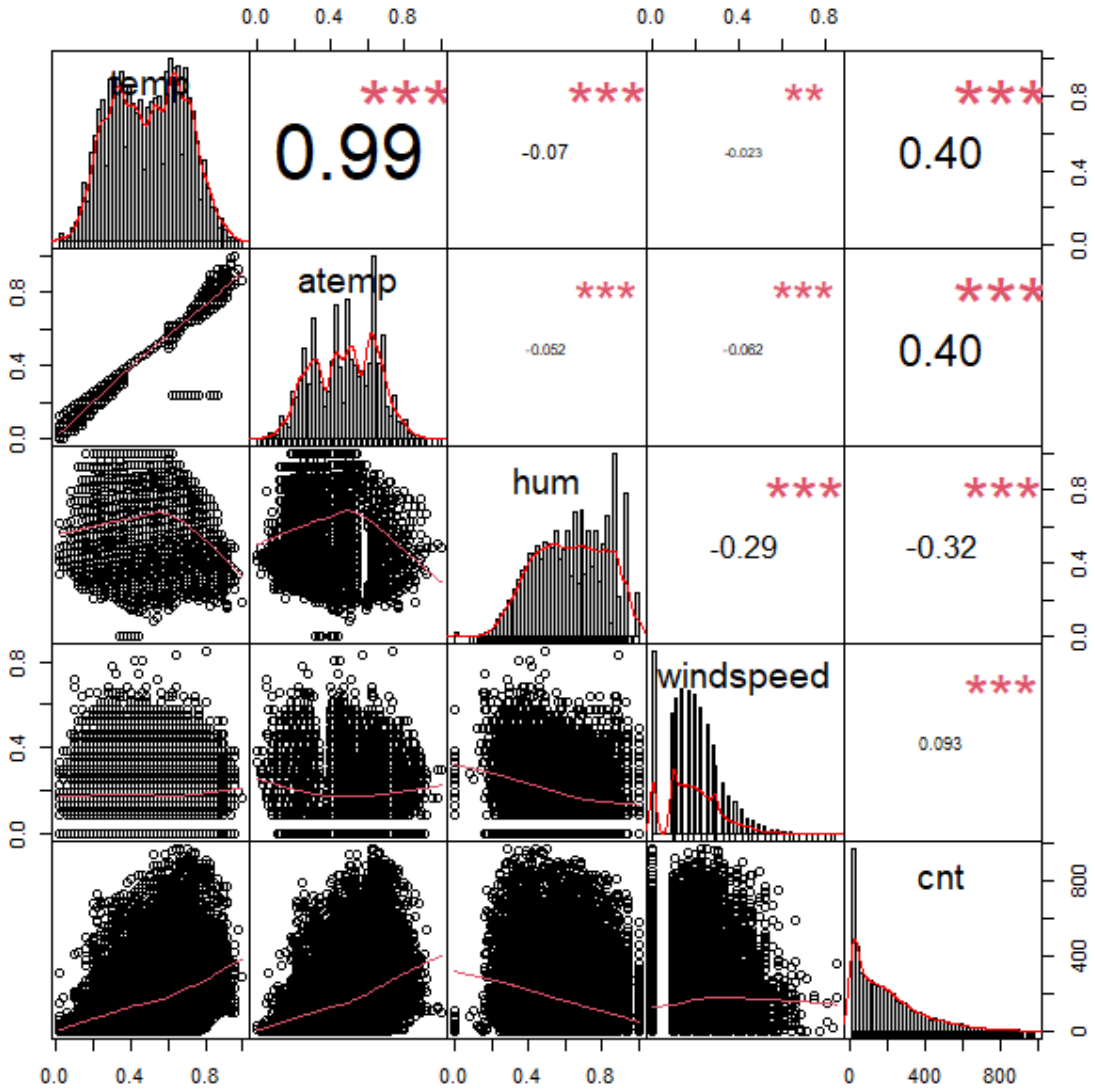
En la tabla siguiente se puede observar que la variable dependiente “cnt” posee una relación relativamente fuerte con la variable independiente “hr” ya que poseen una correlación de 0.39 y el histograma de estas dos variables muestra que la cantidad de bicicletas que se rentan si depende en parte de la hora del día. El resto de las variables en esta tabla (season, yr y mnth) no presentan una gran correlación con la variable dependiente.



En la tabla siguiente se puede observar que ninguna de las variables presentadas tiene una correlación fuerte con la variable dependiente.



En la tabla siguiente se puede observar que hay dos variables que poseen una fuerte correlación con “cnt”, estas son “temp” y “atemp”, con correlaciones de 0.40 cada una. El resto de las variables en esta tabla no poseen una fuerte correlación con “cnt”. En los dos histogramas se puede observar que la relación de estas variables con “cnt” es creciente.



Utilizando esta información se puede concluir que las variables que más relacionadas se encuentran con el comportamiento de “cnt” son: “hr”, “temp” y “atemp”. Por lo que estas variables deberían ser parte del modelo final ya que explican gran parte del comportamiento de la variable que se desea predecir (“cnt”).

2-)

El método utilizado para llegar al modelo que le da solución a este problema fue el de “Selección hacia adelante” utilizando como variables predictoras todas aquellas presentes en el modelo menos las cuatro que fueron descartas al inicio por las razones mencionadas (instant, dteday, casual y registered).

En el punto 3 de este ejercicio se explica cuáles fueron las variables que se eliminaron por no tener una influencia significativa en presencia de las demás.

Luego de eliminar dichas variables la ecuación propuesta para resolver el problema es la siguiente:

$$cnt = -29.24 + (322.03 * atemp) + (7.63 * hr) + (80.97 * yr) + (-203.98 * hum) + (19.98 * season) + (43.47 * windspeed) + (2.12 * weekday)$$

3-)

Luego de crear el modelo una primera vez utilizando la selección hacia adelante con todas las variables predictoras se determinó que había cuatro variables que no eran necesarias ya que aportaban poca información en presencia de las demás. Este es el caso para las variables: “temp”, “holiday”, “weathersit” y “workingday”. Al eliminar estas variables el valor de R cuadrado ajustado solamente varía de 0.3885 a 0.3878, y el estadístico F aumentó de 1005 a 1573, por estos motivos se consideró apropiado retirar estas cuatro variables del modelo final.

En las siguientes imágenes se puede observar como estaba el modelo antes y después de eliminar las variables que no aportaban mucha información.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -25.7569     7.0568  -3.650 0.000263 ***
temp          78.1520    36.9559   2.115 0.034467 *
hr             7.6705     0.1648  46.540  < 2e-16 ***
yr            81.0869     2.1643  37.465  < 2e-16 ***
hum         -198.1911     6.8754 -28.826  < 2e-16 ***
season        19.8767     1.0480  18.967  < 2e-16 ***
atemp        233.1652    41.5125   5.617 1.98e-08 ***
holiday      -21.8837     6.6878  -3.272 0.001069 **
windspeed     41.5660     9.6281   4.317 1.59e-05 ***
weekday        1.8781     0.5404   3.475 0.000512 ***
weathersit    -3.4319     1.9045  -1.802 0.071560 .
workingday     3.9396     2.3954   1.645 0.100059
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 141.8 on 17367 degrees of freedom
Multiple R-squared:  0.3889,    Adjusted R-squared:  0.3885
F-statistic:  1005 on 11 and 17367 DF,  p-value: < 2.2e-16
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -29.2494     6.7064  -4.361 1.30e-05 ***
atemp        322.0260     6.7121  47.977  < 2e-16 ***
hr             7.6324     0.1638  46.590  < 2e-16 ***
yr            80.9696     2.1645  37.409  < 2e-16 ***
hum         -203.9841     6.1159 -33.353  < 2e-16 ***
season        19.9784     1.0472  19.077  < 2e-16 ***
windspeed     43.4682     9.2932   4.677 2.93e-06 ***
weekday        2.1247     0.5372   3.955 7.69e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 141.9 on 17371 degrees of freedom
Multiple R-squared:  0.388,     Adjusted R-squared:  0.3878
F-statistic:  1573 on 7 and 17371 DF,  p-value: < 2.2e-16
```

4-)

El valor final del R cuadrado ajustado para este modelo propuesto es de 0.3878 lo que es igual a un 38.78%, esto significa que el modelo es capaz de predecir el 38.78% de la variabilidad de “cnt” que es la variable dependiente. Por lo que se puede decir que el poder de predicción de este modelo es un poco bajo.

El valor del error cuadrático medio es de 20133.98, este valor es elevado debido a que este modelo no es capaz de predecir con un alto nivel de confianza a la variable “cnt”,