



UNIVERSIDAD EL BOSQUE

FACULTAD DE CIENCIAS

Modelos ML para regresión y clasificación

Angie Caterine Sarmiento Gonzalez [1233507154]

ESTADÍSTICA | MINERÍA DE DATOS

22 de octubre de 2020

ACTIVIDAD

Construir y validar el modelo de regresión más potente para predecir el precio de venta Price de un automóvil nuevo con base en las variables predictoras $X1=KM$ (kilometraje), $X2=Age$ (años de uso) y $X3=Weight$ (peso)

PASOS A SEGUIR

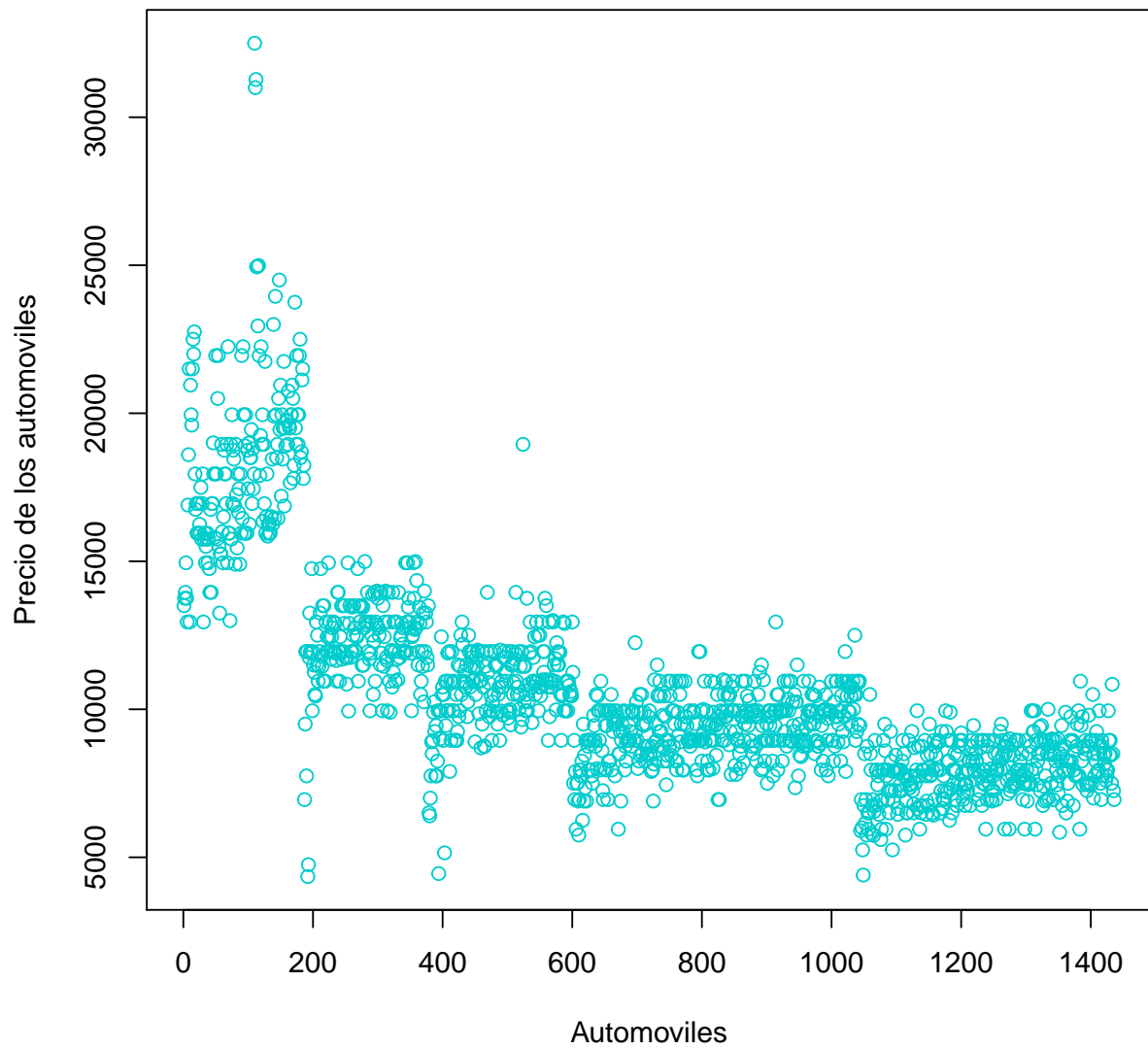
1. Seleccione en un mismo data frame las variables de interés.

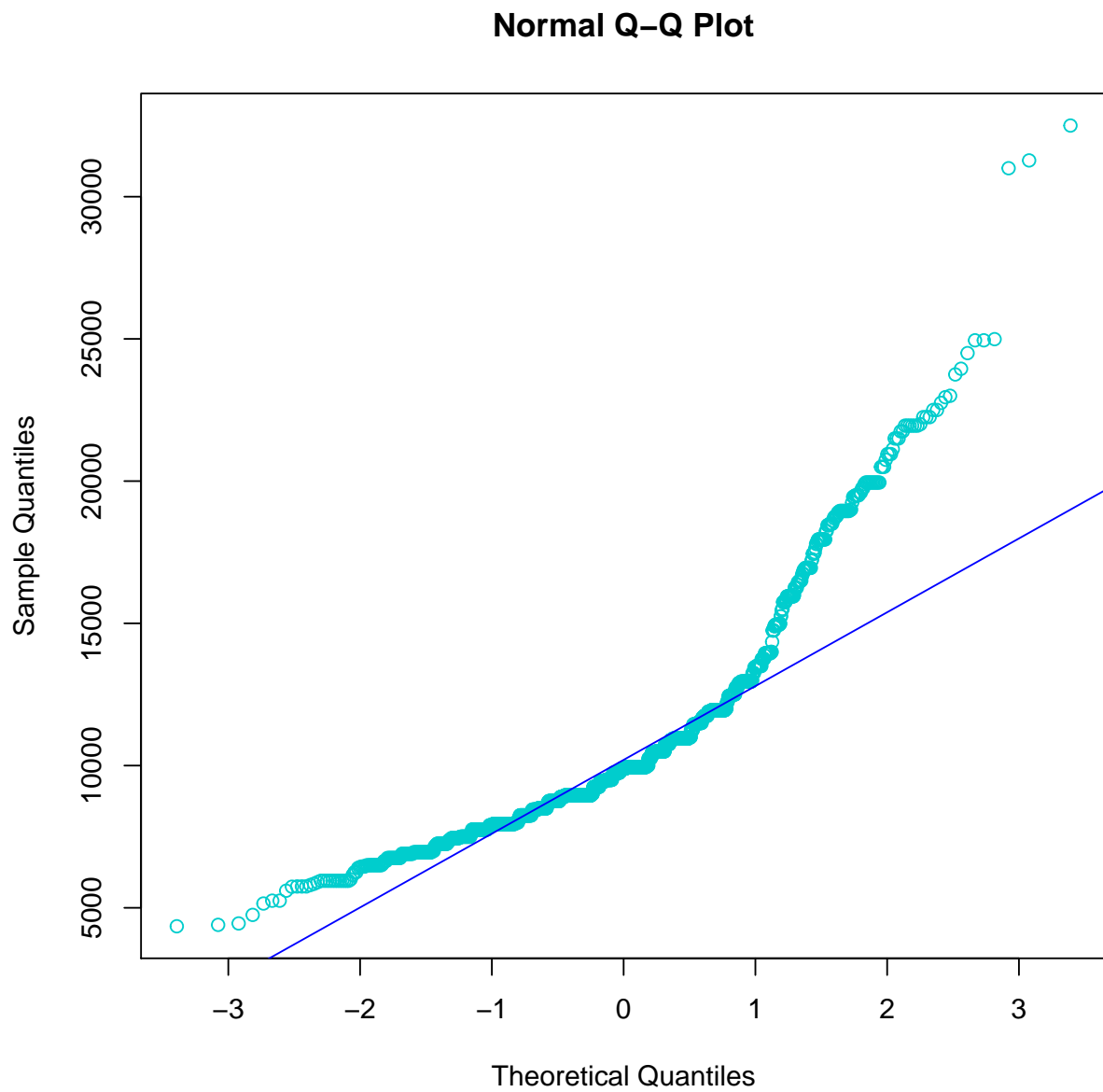
```
## # A tibble: 6 x 4
##      KM      Age Weight Price
##    <dbl> <dbl>   <dbl> <dbl>
## 1 46986    23    1165 13500
## 2 72937    23    1165 13750
## 3 41711    24    1165 13950
## 4 48000    26    1165 14950
## 5 38500    30    1170 13750
## 6 61000    32    1170 12950
```

Exploración de los datos

Se realiza una visualización de la distribución de la variable de respuesta para encontrar el modelo más óptimo que prediga los datos.

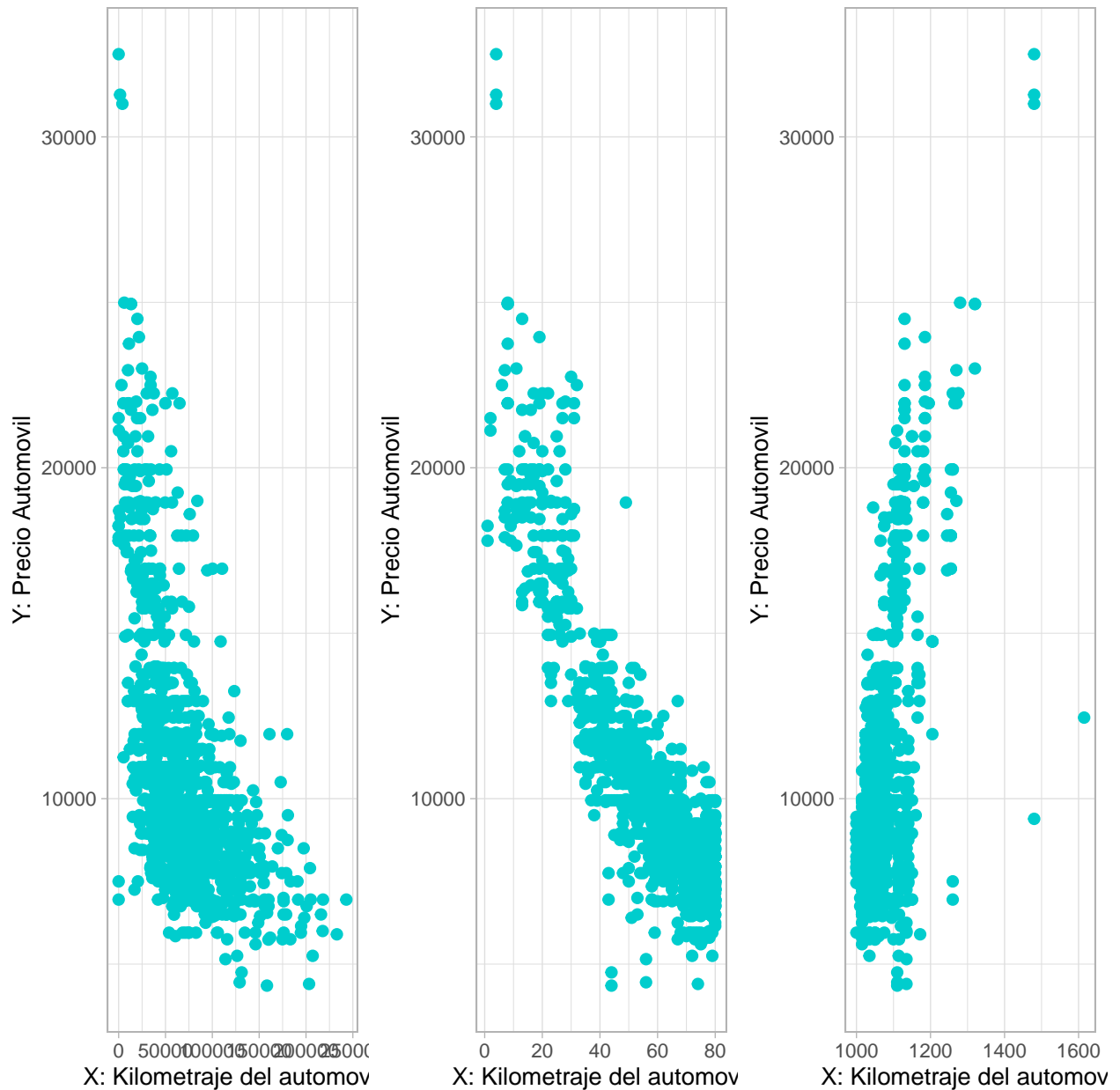
Distribución de la variable de respuesta





Parece ser que hay evidencia para rechazar normalidad de la variable precio.

Distribución del precio de los datos explicado por cada una de las variables



Hay una tendencia de disminución de precio al comparar precio del auto con las variables Age y kM

2. Construya una conjunto de entrenamiento (75 %) y otro de prueba (25 %). Tome la semilla 12345

```
# Seleccionar siempre la misma partición
set.seed(12345)

# Muestra aleatoria del 75% de las filas del
#conjunto "datos" para el conjunto de entrenamiento

train.filas<-sample(x=row.names(datos),size = dim(datos)[1]*0.75)

# CONJUNTO DE ENTRENAMIENTO (selección de columnas)
```

```

train.set<-datos[train.filas,]
train1<-train.set %>% mutate_if(is.numeric,scale)
dim(train.set)

## [1] 1077    4

# CONJUNTO DE PRUEBA
test.filas<-setdiff(x = row.names(datos),train.filas)
test.set<-datos[test.filas,]
test.set1<- test.set %>% mutate_if(is.numeric,scale)
dim(test.set)

## [1] 359    4

```

3. Entrene los cinco modelos con base en el conjunto de entrenamiento y almacene los correspondientes precios predichos para los automóviles de dicho conjunto.

Modelo 1: un modelo de regresión lineal múltiple:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

```

modelo1<-lm(Price~KM+Age+Weight,data=train.set)

```

$$\widehat{Price} = -3189.31 - 0.02 \cdot KM - 117.76 \cdot Age + 20.71 \cdot Weight$$

Modelo 2: un modelo de regresión múltiple de grado 3:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_3^3 + \epsilon$$

```

modelo2<-lm(Price~KM+(Age^2)+(Weight^3),data=train.set)

```

$$\widehat{Price} = -3189.31 - 0.02 \cdot KM - 117.76 \cdot Age^2 + 20.71 \cdot Weight^3$$

Modelo 3: Un modelo ajustado por algoritmo kNN con $k = 10$ vecinos más próximos.

```

modelo3<-FNN::knn.reg(train = train.set,y =train.set$Price,k = 10)

```

Modelo 4: Un modelo ajustado por algoritmo kNN con $k = 10$ vecinos más próximos sobre las variables normalizadas Z_1 , Z_2 y Z_3 .

```
modelo4<-FNN::knn.reg(train=train1 ,y =train1$Price,k = 10)
```

Modelo 5: Un modelo loglineal

```
modelo5 = lm(Price~KM+Age+Weight,data=train.set)
```

```
## # A tibble: 1,077 x 9
##       KM    Age Weight Price Price_pred1 Price_pred2 Price_pred3 Price_pred4
##   <dbl> <dbl> <dbl> <dbl>         <dbl>         <dbl>         <dbl>         <dbl>
## 1  21684    19   1185 23950      18581.       18581.       20584         3.00
## 2  62636    22   1255 17950      18680.       18680.       15575         2.28
## 3  88807    68   1050  8500       8382.        8382.        8505        -0.646
## 4  86714    68   1035  8950       8122.        8122.        8525        -0.582
## 5  81930    76   1070  7750       8021.        8021.       7844.        -0.825
## 6 110287    68   1050  9500       7859.        7859.       9115        -0.330
## 7  69103    68   1035  9750       8551.        8551.       9530        -0.236
## 8 204250    68   1115  7900       6917.        6917.       6305        -1.14
## 9  29650    55   1025  9950      10835.       10835.      10170        -0.170
##10 57000    80   1000  7750       6708.        6708.       8255        -0.793
## # ... with 1,067 more rows, and 1 more variable: Price_pred5 <dbl>
```

4. Estime (y almacene) los correspondientes errores cuadráticos medios de entrenamiento MSEE de los cinco modelos ¿Cuál modelo ajustó mejor al conjunto de entrenamiento?

```
##       MSE_m3 MSE_m1 MSE_m2 MSE_m5 MSE_m4
## 1 1118598 1957205 1957205 1957205 129063302
```

El modelo que mejor se ajustó fue el modelo no paramétrico KNN con k=10, ya que su error cuadrático medio es el más bajo.

5. Evalúe los modelos entrenados en el paso 4 utilizando el conjunto de prueba y almacene los correspondientes precios predichos para los automóviles de dicho conjunto.

```
## # A tibble: 359 x 9
##       KM    Age Weight Price Price_pred1 Price_pred2 Price_pred3 Price_pred4
##   <dbl> <dbl> <dbl> <dbl>         <dbl>         <dbl>         <dbl>         <dbl>
## 1  72937    23   1165 13750      16448.       16448.       12135         0.825
## 2  41711    24   1165 13950      17091.       17091.       14147         1.08
## 3  75889    30   1245 18600      17209.       17209.       14225         2.15
## 4  31461    25   1185 20950      17637.       17637.       19895         2.70
## 5  18739    28   1185 22000      17593.       17593.       20475         2.82
## 6  34000    30   1185 22750      16986.       16986.       20025         2.92
## 7  64359    30   1105 16950      14591.       14591.       15725         1.43
## 8  43905    29   1170 16950      16552.       16552.       15846.         1.88
## 9  56349    28   1120 15950      15332.       15332.       14010         1.45
##10  41000    22   1100 15500      15998.       15998.       14774.         1.42
## # ... with 349 more rows, and 1 more variable: Price_pred5 <dbl>
```

6. Estime (y almacene) los correspondientes errores cuadráticos medios de prueba MSEP de los cinco modelos.

```
##      MSEP_m3 MSEP_m1 MSEP_m2 MSEP_m5  MSEP_m4
## 1 867725.3 2130161 2130161 2130161 125969090
```

7. Compare visualmente los MSEE y MSEP de los cinco modelos. A su criterio ¿Cuál modelo escogería para predecir el precio de nuevos autos? Justifique

```
## Error in data.frame(dato = as.numeric(c(tablaMSEE, tablaMSEP)), Modelo =
rep(c("Modelo1", : arguments imply differing number of rows: 10, 8
## Error in eval(expr, envir, enclos): objeto 'tabla' no encontrado
## Error in ggplot(tabla, aes(x = dato, y = Modelo, fill = Tipo)): objeto 'tabla'
no encontrado
## Error in eval(expr, envir, enclos): objeto 'tabla' no encontrado
```

```
## Error in eval(lhs, parent, parent): objeto 'tabla' no encontrado
```

Se escoge el modelo 3 ya que tiene el menor MSEE

8. Con base en el modelo que seleccionó en el punto 7, prediga el precio que tendrán los siguientes tres automóviles con perfiles:

Automovil	KM	Age	Weight
1	60.000	30	1.300
2	22.000	25	1.500
3	3.000	4	1.070

Tabla 1: Características de los nuevos autos

Estimación de precios para los nuevos autos mediante el modelo KNN con k=10

```
nuevo<- tibble(KM=c(60.000,22.000,3.000),Age=c(30,25,4),Weight=c(1.300,1.500,1.070))

knn.reg(train = train.set[-4], # Sólo predictoras de train.set
        test = nuevo,        # Sólo predictoras de test.set
        y = train.set$Price,  # Sólo respuesta de train.set
        k = 10)

## Prediction:
## [1] 17738.5 17738.5 17738.5
```

Automovil	KM	Age	Weight	Precio Estimado
1	60.000	30	1.300	17738.5
2	22.000	25	1.500	17738.5
3	3.000	4	1.070	17738.5

Tabla 2: Predicción del precio de los nuevos autos

REFERENCIAS

- [1] Ramos David, *Evaluación de modelos para regresión: Ejemplo*, (2020).