TA2 - Enzo Cozza - Agustín Fernández

Ejercicio 1

Contexto y significado

1. CRIM: ratio de crimen per cápita por ciudad.
2. ZN: proporción de tierra residencial para lotes sobre 25000 pies cuadrados.
3. INDUS: proporción de hectáreas para negocios no minoristas por ciudad.
4. CHAS: 1 si limita con el río Charles River, 0 si no.
5. NOX: concentración de óxido nítrico (partes cada 10 millones).
6. RM: promedio del número de cuartos por vivienda.
7. AGE: proporción de casas habitadas construidas después de 1940.
8. DIS: distancia a cinco centros de trabajo de Boston.
9. RAD: índice de accesibilidad a autopistas radiales.
10. TAX: valor del impuesto a la vivienda cada $10000.
11. PTRATIO: ratio alumnos-maestros en la ciudad.
12. B: proporción de gente negra en la ciudad.
13. LSTAT: status más bajo de la población.
14. MEDV: valor mediano de casas (en miles).

Tipo y rangos

1. Real – 0 a 9.967.
2. Real – 0 a 100.
3. Real – 0 a 27.740.
4. Binomial – 0 ó 1.
5. Real – 0.385 a 7.313.
6. Real – 3.561 a 100.
7. Real – 1.137 a 100.
8. Real – 1.130 a 24.
9. Entero – 1 a 666.
10. Entero – 20 a 711.
11. Real – 2.600 a 396.900.
12. Real – 0.320 a 396.900.
13. Real – 1.730 a 34.410.
14. Real – 6.3 a 50.

Distribuciones y outliers

1. Exponencial.
2. Exponencial.
3. No sigue una distribución conocida.
4. -
5. Sesgada a la derecha.
6. Normal.
7. Exponencial.

8. Sesgada a la derecha.
9. No sigue una distribución conocida.
10. No sigue una distribución conocida.
11. No sigue una distribución conocida.
12. No sigue una distribución conocida.
13. Sesgada a la derecha.
14. Normal.

Variable de salida

La variable de salida es MEDV (la media del valor de la casa en miles de dólares).

Ejercicio 2

- ¿Por qué aplicamos *"Shuffle"* luego de recuperar el dataset?

Para crear una copia mezclada del conjunto de datos, por si estos estaban ordenados de alguna manera particular.

- ¿Cómo funciona el operador *"filter examples range"*?

Este operador se encarga de filtrar un dataset por el rango de filas que le sea pasado por parámetro. Se pasa una cota inferior y una superior llamadas 'first example' y 'last example', y devuelve solamente las filas que se encuentran en ese rango (incluyendo los bordes).

- ¿Qué parámetros podemos variar en el operador *"Linear Regression"*?

Se pueden variar feature selection, eliminate colinear features, min tolerance, use bias y ridge.

> ¿Qué hace "feature selection"? ¿cómo?
> Indica el método de selección de característica a utilizar en la regresión. Lo hace mediante las opciones: none, M5 prime, greedy, T-Test, Iterative T-Test.
>
> ¿Cómo afectan "eliminate colinear features" y "use bias"?
> "Eliminate colinear features" es un algoritmo que elimina variables correlacionadas entre sí durante la regresión, mientras que "use bias" es un parámetro que indica si un valor de intercepción debería ser calculado o no.

Ejercicio 3

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code ↑
INDUS	-0.022	0.067	-0.017	0.632	-0.334	0.738	
CHAS	1.242	1.143	0.031	0.987	1.087	0.278	
AGE	-0.030	0.015	-0.092	0.803	-2.021	0.044	**
CRIM	-0.087	0.034	-0.090	0.879	-2.596	0.010	***
ZN	0.042	0.015	0.109	0.815	2.754	0.006	***
NOX	-11.747	4.407	-0.147	0.752	-2.665	0.008	***
B	0.010	0.003	0.102	0.902	3.188	0.002	***
RM	4.873	0.464	0.396	0.596	10.506	0	****
DIS	-1.384	0.233	-0.309	0.786	-5.952	0.000	****
RAD	0.270	0.072	0.248	0.756	3.734	0.000	****
TAX	-0.014	0.004	-0.256	0.716	-3.543	0.000	****
PTRATIO	-0.948	0.151	-0.225	0.818	-6.298	0.000	****
LSTAT	-0.360	0.063	-0.275	0.440	-5.758	0.000	****
(Intercept)	27.266	5.659	?	?	4.818	0.000	****

LinearRegression

```
- 0.087 * CRIM
+ 0.042 * ZN
- 0.022 * INDUS
+ 1.242 * CHAS
- 11.747 * NOX
+ 4.873 * RM
- 0.030 * AGE
- 1.384 * DIS
+ 0.270 * RAD
- 0.014 * TAX
- 0.948 * PTRATIO
+ 0.010 * B
- 0.360 * LSTAT
+ 27.266
```

¿Qué atributos son más y menos significativos?
Los atributos menos significativos según los resultados obtenidos son: INDUS y CHAS (los de menos de 2 estrellas).

Después de aplicar ‘feature selection’ -> ‘greedy’:

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code ↑
AGE	-0.029	0.015	-0.091	0.803	-2.000	0.046	**
CRIM	-0.088	0.033	-0.091	0.880	-2.641	0.009	***
ZN	0.042	0.015	0.108	0.815	2.742	0.006	***
NOX	-11.414	4.218	-0.143	0.754	-2.706	0.007	***
B	0.010	0.003	0.103	0.903	3.228	0.001	***
RM	4.918	0.460	0.400	0.602	10.698	0	****
DIS	-1.364	0.227	-0.304	0.783	-5.994	0.000	****
RAD	0.281	0.069	0.257	0.749	4.058	0.000	****
TAX	-0.015	0.004	-0.270	0.728	-4.166	0.000	****
PTRATIO	-0.972	0.147	-0.231	0.825	-6.591	0.000	****
LSTAT	-0.368	0.062	-0.281	0.444	-5.933	0.000	****
(Intercept)	27.223	5.636	?	?	4.830	0.000	****

LinearRegression

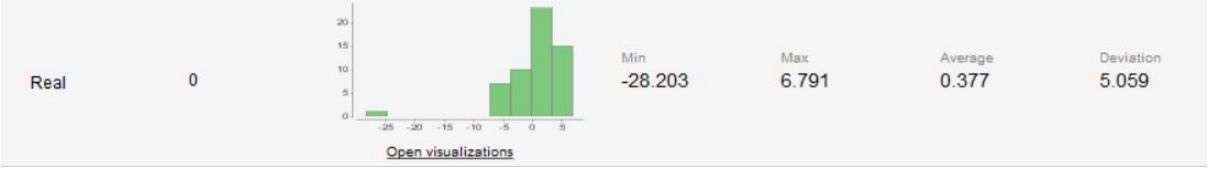
```
- 0.088 * CRIM
+ 0.042 * ZN
- 11.414 * NOX
+ 4.918 * RM
- 0.029 * AGE
- 1.364 * DIS
+ 0.281 * RAD
- 0.015 * TAX
- 0.972 * PTRATIO
+ 0.010 * B
- 0.368 * LSTAT
+ 27.223
```

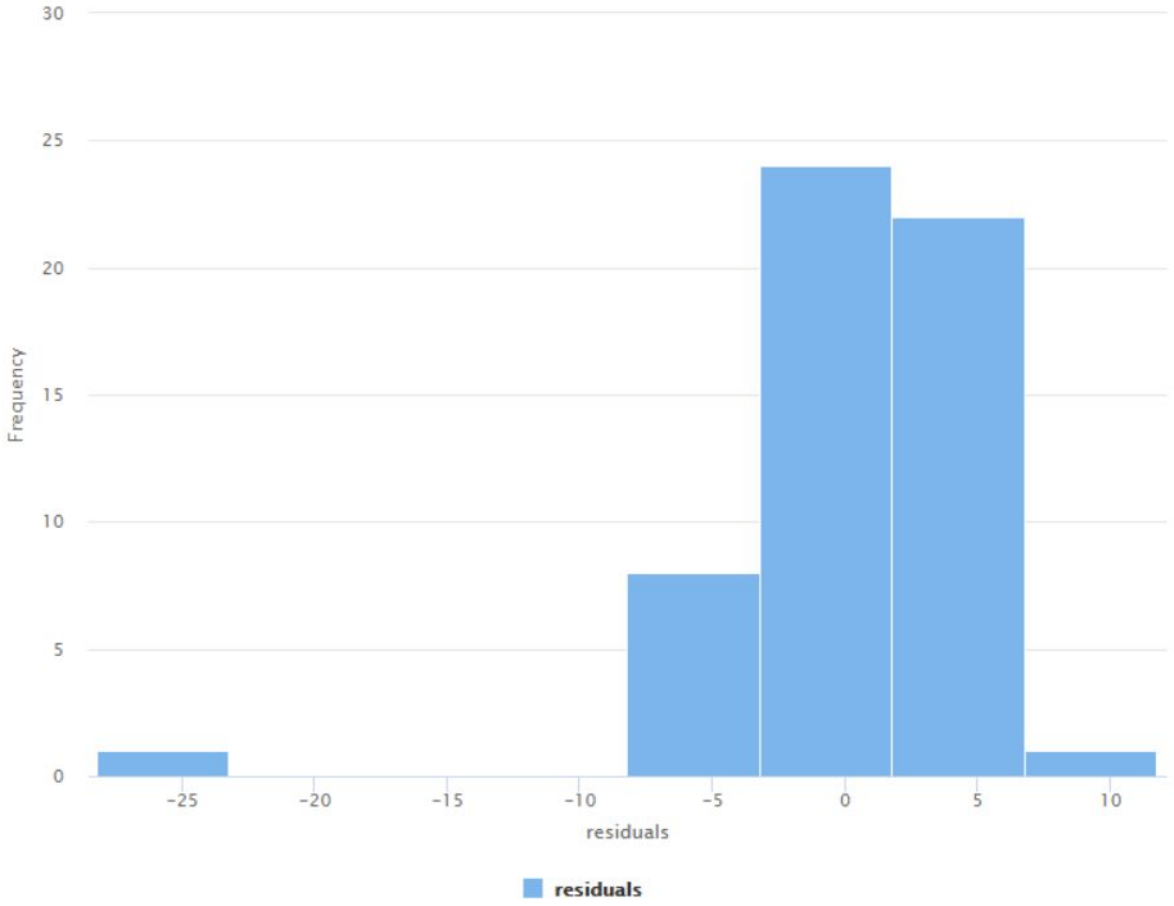
Los atributos eliminados fueron INDUS y CHAS. Respecto a los coeficientes, casi todos sufrieron un reajuste en el entorno de las centésimas o milésimas. Logrando, de esta manera, una mejora en el modelo.

Valor de R^2
Valor con el feature selection: 0.655.
Valor sin el feature selection: 0.666.

Valor del error medio cuadrático
Valor con el feature selection: 33.544 +/- 106.173.
Valor sin el feature selection: 32.408 +/- 101.926.

Ejercicio 4





De las estadísticas podemos concluir que la mayoría de los residuos tienden a estar cercanos a 0, es decir, las predicciones se acercan valor real de la casa. Hay excepciones, como se ve en las dos columnas de los extremos, que se podrían tratar como outliers del problema. Por otra parte, teniendo en cuenta la media, vemos que la misma se aproxima bastante a 0, sin embargo la desviación estándar es bastante elevada (aproximadamente 5), lo que indica que los datos se encuentran dispersos.