

Práctica 2

Predicciones sobre precios de autos mediante regresión lineal
21/06/2022

Alberto Isaac Pico Lara

Con el objetivo de encontrar el mejor modelo de regresión, se aplicaron las siguientes metodologías de regresión a los tres conjuntos de datos sobre precios de automóviles para las marcas Audi, BMW y Toyota.

- Regresión lineal simple
- Lasso
- LARS
- Regresión de cresta
- ElasticNet

Para encontrar el mejor ajuste, se realizó una selección de características utilizando el método de correlación con el objetivo, en este caso el precio, del resto de variables continuas para lo que se obtuvo que en los casos de las marcas premium sin tanta variabilidad, las variables con mayor correlación respecto al objetivo son

```
bmw[vars_c].corrwith(bmw['price'])  
✓ 0.2s  
mileage      -0.605397  
tax           0.263484  
mpg          -0.204882  
engineSize   0.460196  
age          -0.623756  
dtype: float64
```

```
audi[vars_c].corrwith(audi['price'])  
✓ 0.1s  
mileage      -0.535357  
tax           0.356157  
mpg          -0.600334  
engineSize   0.591262  
age          -0.592581  
dtype: float64
```

Con un criterio de tomar sólo aquellas variables con una $|correlación| > 0.4$ Nos quedamos con el valor del odómetro, el tamaño del motor y la edad y para Audi, adicionalmente las millas por galón.

Para el caso de la marca Toyota, las cosas fueron muy distintas, debido a la variabilidad de sus modelos la correlación con la variable objetivo no es muy fuerte.

Un análisis preliminar de este fenómeno nos puede indicar el hecho de que Toyota como marca generalista que cuenta con vehículos deportivos de alto rendimiento, pero a la vez vehículos básicos; el valor del odómetro no es tan decisivo como se podría pensar, pero sí lo es el tamaño del motor, ya que llevando a cabo una exploración básica y una búsqueda en internet, los modelos de Toyota con motor más grande suelen ser más costosos, además del factor de las motorizaciones híbridas en las que Toyota lleva más de 20 años siendo líder en su implementación, en cuyo caso es una relación inversa, ya que son vehículos con tamaño de motor contenido, pero por la tecnología involucrada en su construcción suben de precio y se deprecian menos. Esto puede explicarse debido a las leyes aplicables en materia vehicular que imponen menos restricciones a la circulación de vehículos menos contaminantes y por ende, pierden menos valor a través del tiempo.

```
ty[vars_c].corrwith(ty['price'])
✓ 0.2s
```

mileage	-0.297936
tax	0.214498
mpg	-0.039910
engineSize	0.498595
age	-0.420959
dtype:	float64

DATO CURIOSO: Por lo anterior, se puede pensar que si uno de los criterios para elegir un automóvil es el valor de reventa futuro, la mejor opción a mediano y largo plazo sería un Toyota con tecnología híbrida (no todos son tan feos como Prius)

Entrando en materia de modelos de regresión para predecir el precio. Tras un análisis de residuales y cumplimiento de supuestos para regresión como normalidad, homocedasticidad, linealidad e independencia.

Se obtuvo que para el caso de Audi el modelo LARS es el que mejor ajusta, para BMW un modelo de ElasticNet es el indicado y finalmente para Toyota un modelo LARS.

Con las millas y los años como variables de interés para medir la depreciación de un vehículo se presentan a continuación los resultados de los modelos con mejor ajuste para cada marca.

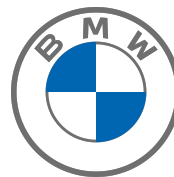
```
lars
Por cada kilómetro, se pierden 0.038 dólares
Por cada año que pasa, se pierden 1058.031 dólares
*****
Se pierde $1 USD cada 26.163 kilómetros
Se pierde $1 USD cada 0.00095 años o bien... 8.322 horas
```



```
lars
Por cada kilómetro, se pierden 0.093 dólares
Por cada año que pasa, se pierden 2142.175 dólares
*****
Se pierde $1 USD cada 10.696 kilómetros
Se pierde $1 USD cada 0.00047 años o bien... 4.1172 horas
```



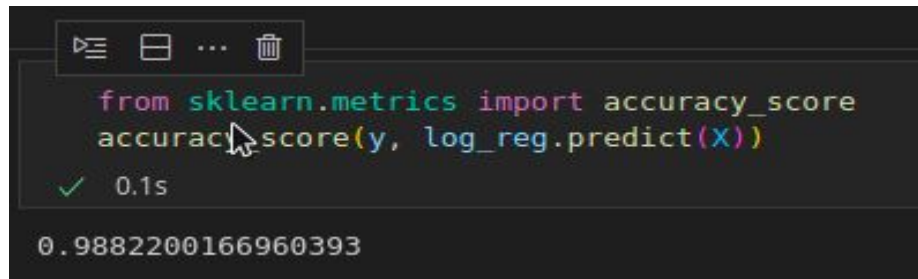
```
elasticNet
Por cada kilómetro, se pierden 0.147 dólares
Por cada año que pasa, se pierden 1769.081 dólares
*****
Se pierde $1 USD cada 6.822 kilómetros
Se pierde $1 USD cada 0.00057 años o bien... 4.9932 horas
```



Encontrando modelos híbridos en la marca BMW mediante una regresión logística...

Se realizó un modelo de regresión logística para los datos de BMW utilizando un filtro de baja varianza como criterio para selección de variables. Dicho filtro nos arrojó como resultado que se pueden utilizar como variables el precio, el impuesto (TAX), las millas por galón y el tamaño del motor.

El modelo obtuvo un accuracy score de

A screenshot of a Jupyter Notebook interface. The top bar shows icons for running, saving, and deleting cells. The code cell contains the following Python code:

```
from sklearn.metrics import accuracy_score  
accuracy_score(y, log_reg.predict(X))
```

 Below the code, there is a green checkmark icon and the text "0.1s", indicating the execution time. At the bottom of the cell, the output is displayed as "0.9882200166960393".

```
from sklearn.metrics import accuracy_score  
accuracy_score(y, log_reg.predict(X))  
✓ 0.1s  
0.9882200166960393
```

Conclusiones. De las tres marcas que se analizaron para probar distintos modelos lineales, Toyota es la marca más compleja de analizar debido a la ya mencionada variabilidad posible en su gama de modelos y representa una excelente opción a compra para quienes buscan un vehículo fiable, duradero y con un buen valor de reventa. También se encontraron otros hallazgos al respecto de la relación entre el tamaño del motor y el valor del odómetro, es interesante debido a la física de materiales. Entre más pequeño es un motor, más esfuerzo realiza para mover el vehículo, entre más millas recorridas, más desgastado y menor precio de reventa, excepto para los automóviles híbridos.

Conclusiones parte 2. Todos los modelos de regresión parten de la misma base de la regresión lineal y el ajuste por mínimos cuadrados. Que si bien es el primer modelo que se aprende en la materia de Estadística II también es confiable y su simpleza (respecto a otros modelos) de interpretabilidad de resultados e implementación los convierten en la mejor opción para toma rápida de decisiones y encontrar la respuesta a algunas preguntas de negocio e incluso para temas de econometría.

Las variaciones que se pueden realizar sobre el modelo para mejorar los resultados sin modificar la interpretabilidad también son herramientas de gran utilidad, llegando a ser incluso herramientas bastante robustas como lo es la regresión logística que de manera personal he tenido que utilizar en el ámbito profesional para detección de eventos de brecha de seguridad.