Juan Jose Elgueta Rodriguez

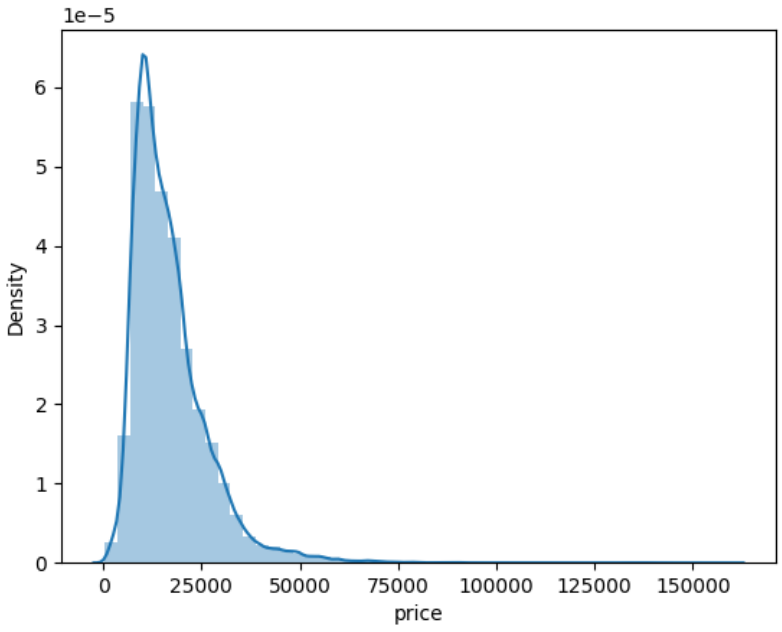
Proyecto 3

introducciòn

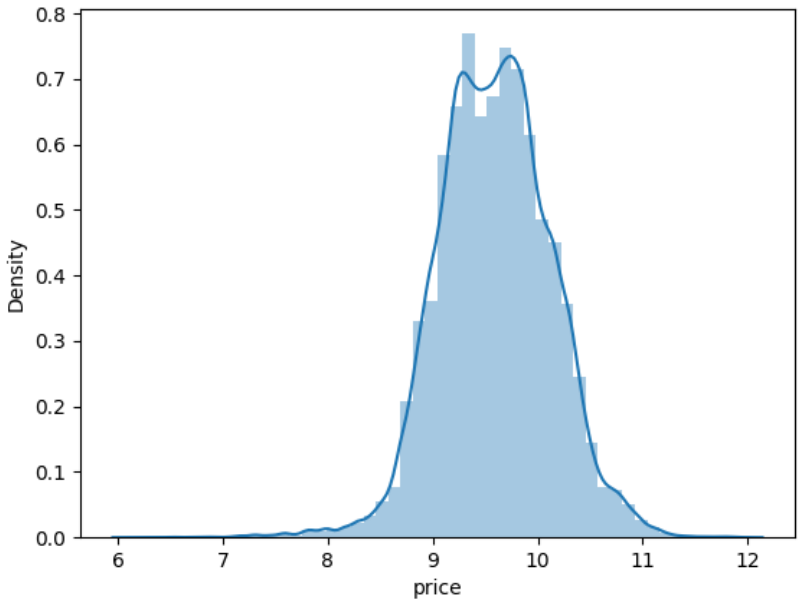
Cuando pensamos en machine learning a veces automáticamente pensamos en alguna u otra aplicacion sofisicada como Natural Language Processing o Computer Vision, y a pesar de que estas aplicaciones son populares y tremendamente útiles, se nos olvida que hay otras aplicaciones de machine learning que tal vez no son tan elevadas pero pueden ser aún más útiles para nuestro dia a dia, un gran ejemplo de esto es la regresión linear. ¿Cuántos fenómenos que ocurren en nuestro dia a dia no pueden ser explicados por esta poderosa arma? El propósito de este informe es aplicar conocimientos básicos de machine learning para crear un modelo predictivo que explique cómo varía el precio de carros usados según los features que tenemos disponibles en el dataset que será explicado a continuación.
Intentaremos contestar la siguiente pregunta: ¿es suficiente un modelo de regresión lineal para explicar de manera exacta la variación en el precio de autos usados según las variables en el dataset proporcionado?

Data

La data proporcionada consiste en 108540 observaciones de carros usados en predios, esta contiene el precio del vehículo y 9 features: model, year, transmission, milage, fuelType, tax, mpg, engineSize y make. Estos features son autoexplicativos y por lo tanto no los discutiremos en profundidad.
Nuestra variable objetivo tiene la siguiente distribución:



Podemos observar que esta forma se asemeja a una distribución log-normal y dado que está distribución regularmente se acopla a variables del tipo financiero como los precios de activos podemos asumir que esta es la distribución de la variable. Para hacernos la tarea de predecir esta variable de forma más fácil vamos a transformarla a una distribución normal, para esto solo necesitamos aplicar logaritmo a los precios. La distribución de los precios después de la transformación será:



Ahora esta distribución parece ser más normal.

Además de nuestra variable objetivo, dentro de los features tenemos dos columnas con valores vacíos los cuales imputaremos luego.

```
model              0
year               0
price              0
transmission       0
mileage            0
fuelType           0
tax             9353
mpg             9353
engineSize         0
make               0
dtype: int64
```

Juan Jose Elgueta Rodriguez

Metodos

Outliers

Dentro de nuestro dataset encontramos algunos outliers que valen la pena mencionar:

a) El primero en nuestra variable ‘year’ en donde encontramos un carro del año 2060, esta observación la eliminaremos del dataset ya que no proporciona ninguna información importante. La observación es un Ford Fiesta, tenemos 6557 otros Ford Fiesta así que no perderemos nada con solo eliminarlo.
b) El segundo es un grupo de vehículos con un mpg mayor a 400, lo cual es imposible. Al explorar la data podemos encontrar que son todos los BMW I3. Podemos asumir que este es un error en la data ya que el valor es inverosímil. No queremos perder a todas las observaciones de un modelo de auto así que lo que hacemos es sustituir ese valor por la media de ese valor en todos los autos que tienen un tipo de combustible que sea: "Electric","Hybrid" u “Other”

Valores vacios

Tenemos valores vacíos en las columnas de mpg y tax. Al analizar la data podemos ver que estos valores pertenecen solo a 2 modelos de autos: Mercedes C Class y Ford Focus. Como tenemos varias otras observaciones para esos modelos entonces crearemos un imputador a la medida para imputar según la media de una variable categórica.

Modelo

El modelo que utilizaremos es ridge regression. Esta es una regresión lineal con regularización L2. Se escogió este modelo ya que queremos un modelo que generalice bien y además tenemos varios features que pueden ser buenos estimadores.

Selección de parámetros

Se utilizó grid search para encontrar que alpha de regularización sería la mejor para el modelo.

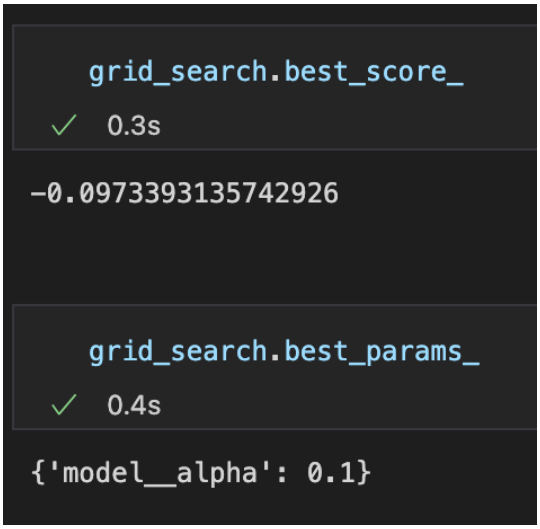
```
ridge_pipeline = Pipeline(steps=[
    ('impute', ByCategoryImputer( category_col='model',target_cols=['mpg'],strategy='mean')),
    ('preprocess', full_processor),
    ('model', Ridge())
])

param_grid = {
    'model__alpha': [1e-3, 1e-2, 1e-1, 1e0]
}
```

```
grid_search = GridSearchCV(
    ridge_pipeline,
    param_grid,
    cv=5,
    scoring='neg_mean_absolute_error'
    )
```

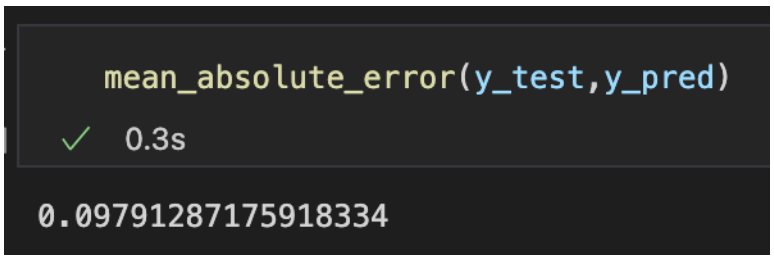
La métrica que vamos a optimizar es MAE con esta métrica mientras más cercano sea a 0 el valor significa que nuestro modelo es un mejor predictor.

Resultados

Los resultados de grid search fueron los siguientes:



Al calcular el MAE para la porción de test del dataset obtenemos los siguientes resultados:



Estos resultados significan que nuestro modelo es bastante buen predictor tanto para la porción de entrenamiento como para la de test.

Juan Jose Elgueta Rodriguez

Conclusiones

Regresando a la pregunta de la introducción ¿es suficiente un modelo de regresión lineal para explicar de manera exacta la variación en el precio de autos usados según las variables en el dataset proporcionado? Podemos contestar que sí. Nuestro modelo de regresión lineal regularizada alcanzó un MAE de 0.0973393135742926 en la porción de entrenamiento del dataset y 0.09791287175918334 en la porción de test.
En el futuro podríamos buscar si algún otro modelo de regresión lineal podría darnos incluso mejores resultados.