Practical Application Assignment 11

What Drives the Price of a Used-Car?

The purpose of this analysis is to determine the factors that affect the price of used cars. The dataset used for this analysis is obtained from Kaggle, which includes information on various features of used cars such as make, model, year, mileage, engine size, and others. A linear regression analysis was conducted on the dataset to identify the factors that influence the price of used cars.

Business Understanding

The goal of the study is to identify a set of factors that drive the price of a car upwards or downwards and a model that can predict the price of a vehicle based on these factors. It is being done for a client who owns a used car dealership.

Data Understanding

The dataset contains 426,880 rows and 18 features for each row. The features are:

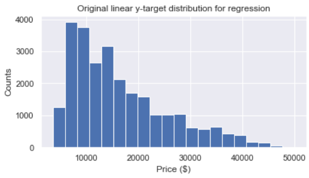
1. ID: Unique identifier
2. VIN: Vehicle identification number
3. Region:
4. Price: Price in USD, not adjusted for inflation.
5. Year: The year in which the car was manufactured
6. Manufacturer: 43 unique auto brands that manufacture the automobiles that are a part of this dataset.
7. Model: The model of the car.
8. Condition: The condition of the car; excellent, good, fair, like new, salvage, new.
9. Cylinders: The number of cylinders in the car engine ranging from 3 to 12. Also has the 'other' category too.
10. Fuel: There were five types of fuel, 'diesel', 'gas', 'electric', 'hybrid' and 'other'.
11. Odometer: The distance travelled by the car after it was first bought.
12. Title Status: The cars have 6 types of statues; 'clean', 'lien', 'rebuilt', 'salvage' , 'parts only' and 'missing'.
13. Transmission: There are 3 types of transmission; 'automatic', manual and other.
14. Drive: There are 3 types of drive transmissions; '4WD, 'FWD' and 'RWD'. (Four-wheel drive, forward-wheel drive and rear-wheel drive.)
15. Size: Size of the vehicle
16. Type: This feature identifies if a vehicle is a SUV or a mini-van. There 13 unique values in this feature.
17. Paint_Color: This feature identifies the color of the car. There 12 unique values in this feature.
18. State: The state is political territory and is represented in short form in the data set. Like "fl" is used for the state of Florida.

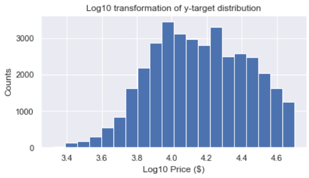
The price column is the target variable that will be predicted by the model. The rest are features some of which will used to predict price using a variety of linear regression models.

Data Preparation

We begin inspecting and eliminating columns are useful in predicting price. id and VIN are unique identifiers of each record and have no correlation with price. region is related to another feature, state. They are dropped from the feature set.

Price, year, and odometer are cleaned by picking specific ranges that make sense and dropping all records with NULL or outlier values. Price is transformed to log_price (to get closer to a normal distribution). Running a correlation matrix on the numerical variables shows better correlation with log_price than price.



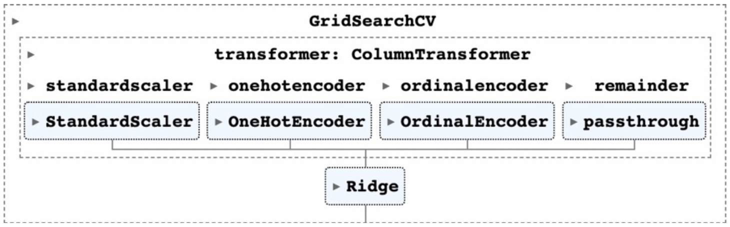


Modeling

The final features chosen for modeling are 'year', 'condition', 'cylinders', 'fuel', 'odometer', 'title_status', 'transmission', 'drive', 'size', 'type'. A correlation heatmap showed that model, odometer, year have the highest correlations with log_price, followed by type, transmission, and manufacturer.

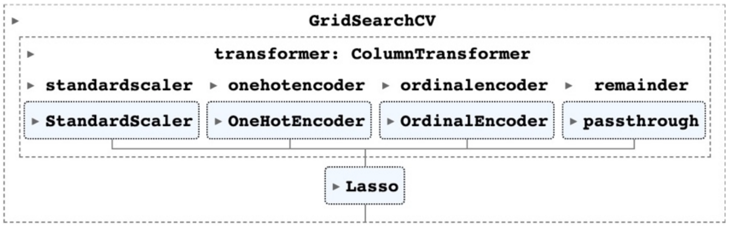
The following linear regression models were explored:

1. **Ridge Regression:** Was run with GridSearchCV and selected optimal hyperparameter alpha of 3.23. The top features selected are a function or some higher order or interaction term including year, odometer, model, type, drive, and condition.



```
▸ GridSearchCV
▸ transformer: ColumnTransformer
▸ standardscaler    ▸ onehotencoder    ▸ ordinalencoder    ▸ remainder
▸ StandardScaler    ▸ OneHotEncoder    ▸ OrdinalEncoder    ▸ passthrough
▸ Ridge
```

2. **Lasso Regression**: Was run with GridSearchCV and selected optimal hyperparameter alpha of 0.003.

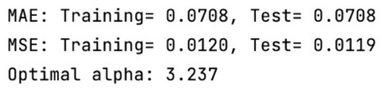


```
▸ GridSearchCV
▸ transformer: ColumnTransformer
▸ standardscaler    ▸ onehotencoder    ▸ ordinalencoder    ▸ remainder
▸ StandardScaler    ▸ OneHotEncoder    ▸ OrdinalEncoder    ▸ passthrough
▸ Lasso
```

Evaluation

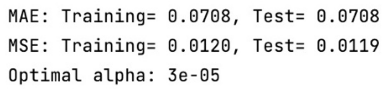
The Validation Mean Squared Errors (MSEs) for the models run above were:

a) Ridge Regression:

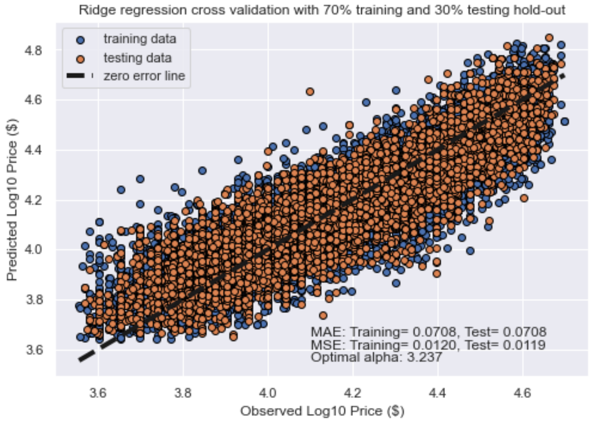


```
MAE: Training= 0.0708, Test= 0.0708
MSE: Training= 0.0120, Test= 0.0119
Optimal alpha: 3.237
```

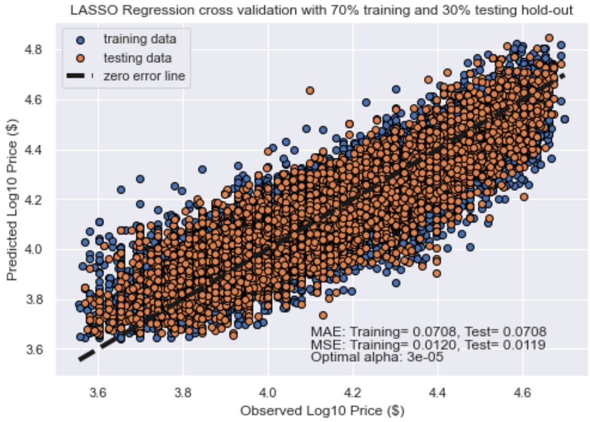
b) Lasso Regression:



```
MAE: Training= 0.0708, Test= 0.0708
MSE: Training= 0.0120, Test= 0.0119
Optimal alpha: 3e-05
```

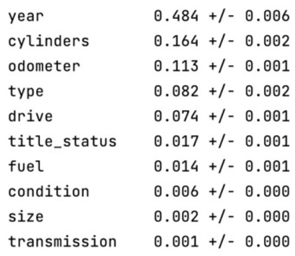


Ridge regression cross validation with 70% training and 30% testing hold-out



LASSO Regression cross validation with 70% training and 30% testing hold-out

Validation MSEs are similar for both Ridge and Lasso models but since Ridge is computationally more efficient, it is picked as the winner. The feature importance for this model, as computed by Permutation importance was:



```
year            0.484 +/- 0.006
cylinders       0.164 +/- 0.002
odometer        0.113 +/- 0.001
type            0.082 +/- 0.002
drive           0.074 +/- 0.001
title_status    0.017 +/- 0.001
fuel            0.014 +/- 0.001
condition       0.006 +/- 0.000
size            0.002 +/- 0.000
transmission    0.001 +/- 0.000
```

Deployment

From permutation importance of the best model, the following are the top variables (in order) that determine the price of a used car and can be informed to the client/used car dealer. a) Year, b) Cylinders, c) Odometer, d) Type, e) Drive, f) Title Status and g) Fuel.

Key Findings and Recommendations

1. The price of vehicles decrease with increasing age and higher odometer readings. Once the vehicle is ten years and older, its price falls-down to $10,000 or less. It is recommended to manage the age and mileage of inventory held carefully.

2. The regression model indicates that specialized vehicles (e.g., off-road, truck) and powerful vehicles (cylinders > 6), have a higher price value while economy vehicles with lower power have a lower resale value. Consumers want utility (off-road, trucks, four-wheel-drive) and powerful (10,8,6 cylinder) vehicles. As a result, vehicles with these features are valued more.

Next Steps

a) We recommend further analysis on the following features: a) region, b) model and make and c) color. These features were not used in the regression. If trends exist, then future work could include adding these.

b) Revisit the dataset in a few years to see if there is an increase in value for: a) fuel efficient vehicles due to gas inflation AND b) electric vehicles.