[**Supervised Machine Learning: Regression**](https://www.coursera.org/learn/supervised-machine-learning-regression/home/welcome)

**Peer To Peer Assignment**

# Main objective: Car Price Prediction

We are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels.

**Brief description of the data set and a summary of its attributes**

I have data set consisting of 205 data points and 26 columns representing features.

There are some features that are not important for target variable , I will simply drop them and there are also some categorical feature, that I will convert to numerical form by Encoding method.

**car\_ID** ID of every car

**CarName** Name of Car

**fueltype** Type of Fuel

**doornumber** Total number of door

**carbody** Body of Car weather Sedan or Hatchback etc

**enginelocation** Location of engine in car

**wheelbase** Distance between rear and front wheel

**carlength** Length of Car

**carwidth** Width of Car

**carheight** Height of Car

**curbweight** Weight of Car without any passenger or item

**enginetype** Type of engine

**cylindernumber** Total cylinder in Car

**enginesize** Size of Engine

**boreratio** Combustion Performance of Lean Burn Heavy-Duty Gaseous Engine

**stroke** A phase of the engine's cycle

**horsepower** Power of Car

**peakrpm** Revolution per minute

**citympg** City mileage per gallon

**highwaympg** Highway mileage per gallon

Table

Description automatically generated**price** Price of car

A picture containing graphical user interface

Description automatically generated

A picture containing scatter chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

A black and white checkered surface

Description automatically generated with low confidence

Graphical user interface, text, application, email

Description automatically generated

**Summary of data exploration and Actions taken for Data Cleaning and Feature Engineering**

1. **FINDINGS**

After reading data by pandas read\_csv function, I applied

* 1. **isnull()** to see is their any missing in the dataset. But I found their no missing value in my dataset.
  2. **Info()** function to see data type of different feature of my dataset and I found there are many feature that are important for my target variable and their data type is object.
  3. **Skew() and hist()** function to check shewness of data and their was some columns that are right or left skewed. e.g compressionratio and wheel etc.
  4. **Pairplot()** to see correlation and also to find is their any need to use **Polynomial feature** for higher degree. i.e 2,3,4 etc.
  5. **Heatmap()** to see correlation in more depth by printing their corresponding value of relation with each other.

1. **ACTION:**
2. For column that are important for by target variable and their data type was Object, I applied **get\_dummies()** function of pandas to convert them into numeric type.
3. For removing skewness of different columns, I applied **log1p** transformation function.
4. Although I found there are some columns are correlated with each other but when I applied **Variation Inflation Factor** technique to remove correlation among them, I found that it have negative impact on my r2\_score because in this technique some columns are drop for eliminating correlation. So then I trained my model without removing correlation.
5. Also their was some columns that are not important for our target, So I simply drop them.
6. Similarly, I used **Standard Scaler** and **Polynomial feature** technique on dataset before giving it to model for training.

**Summary of Three Different Linear Regression Model**

1. **Simple Linear Regression Model**

Graphical user interface, text, application, email

Description automatically generated

1. **Graphical user interface, text

   Description automatically generatedRidge Regression Model**

**Graphical user interface, text, application, email

Description automatically generated**

1. **Lasso Regression Model**

**Graphical user interface, text, application, email

Description automatically generated**Graphical user interface, text, application, email

Description automatically generated

1) As in first case I applied simple linear regression on model and it’s r2\_score was 0.60 which is not considered as good result.

2) In second case I applied Polynomial feature technique on dataset before giving it to Ridge regression. I also used GridSearchCV method and pipeline technique to make process fast and easy to find out best best parameters for our model to have good prediction.

So by doing this, I found that parameters i.e polynomial feature and alpha values 3 and 14 are best for our model to have r2\_score of 0.89 .

3) Similarly for case three I used Lasso Regression, GridSearchCV and pipeline technique and founded that value of 2 for degree and 10 for alpha is best for our model to have good r2\_score of 0.89 .

**A paragraph explaining which of your regressions you recommend as a final model that best fits your needs in terms of accuracy and explainability.**

For choosing model that best fir our data, Ridge and Lasso both are better than simple linear regression and both have same r2\_score, So we can choose any one from both of them but if we want interpretability along with our main goal of prediction then Lasso will be choice.

**Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your linear regression model**

Some of key point for this data and regression model is that as we know that correlation in data is not good for our model and it have impact on model accuracy but in our case when we used technique of VIF to eliminate correlation from data, I found that it does not have good impact on accuracy of model because in this technique we drop one of column that are correlated with each other, So by doing that we will end up with having less number of feature for our model to train on which is again problem. i.e Problem of underfitting.

**Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model adding specific data features to achieve a better explanation or a better prediction**

So my next suggestion for analyzing this data will be to have more data because as we can see our data consist of only 205 row or observation that is not good enough to train model.