أكاديمية سدايا
SDAIA Academy

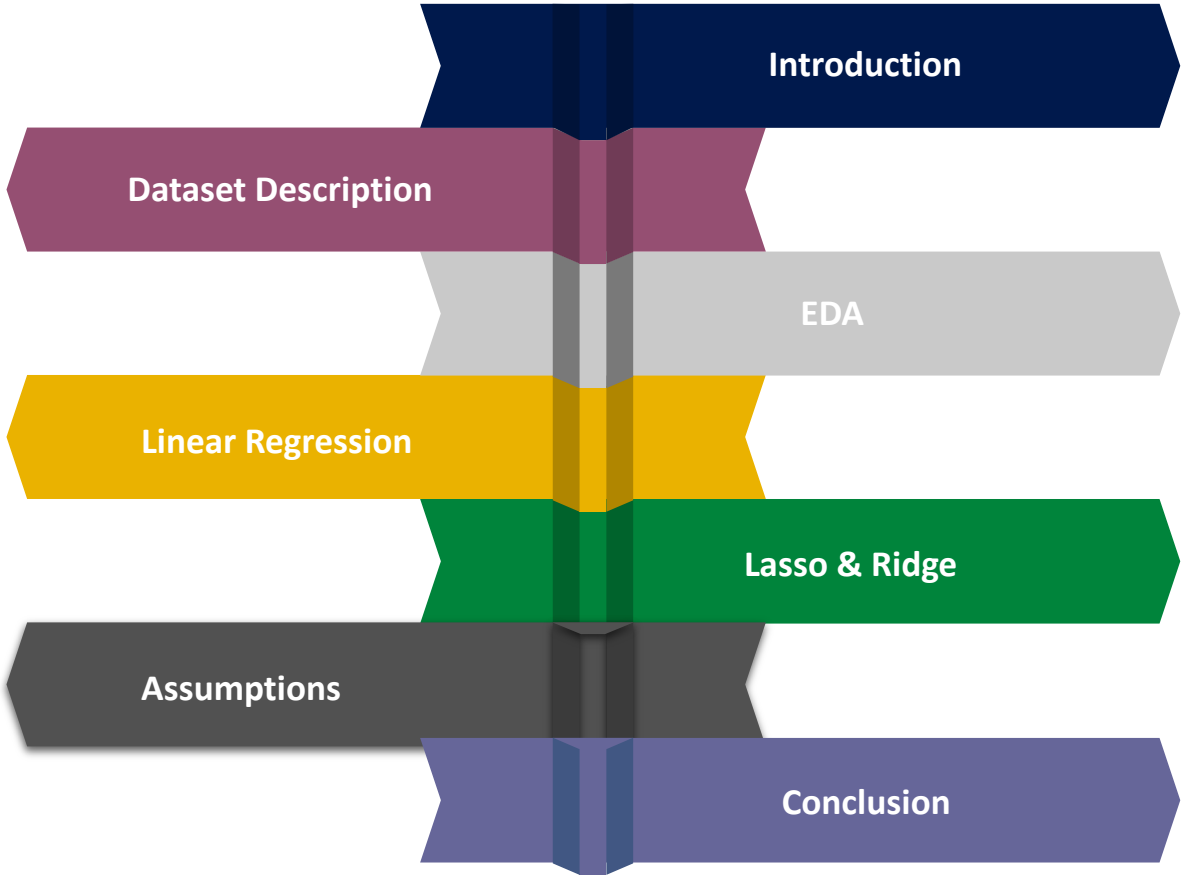
Predicting car price using linear regression

Presented by :

Arwa AlBassam & Nouf Alsaeed

9/26/21

OUTLINE


Introduction
Dataset Description
EDA
Linear Regression
Lasso & Ridge
Assumptions
Conclusion

Introduction

Lots of people renovate and sell their cars.

But they face the problem of determining the car's selling price based on the specifications of their cars. To help them determine the expected price for the car, we used the data from cars.com and we build the linear regression model to predict the price of the car.

Dataset Description

From Cars.com

400 Rows X 6 Columns

FEATURE	DESCRIPTION
DESCRIPTION	Column contains the car name and the model of the car
DEALER	Name of the automobile dealer.
MILES_DRIVEN	The kilometers driven for the used cars.
RATE	Rate of the dealer out of 5
NO_OF_REVIEWS	Number of reviews for each cumulative rate.
PRICE	Price of the car.

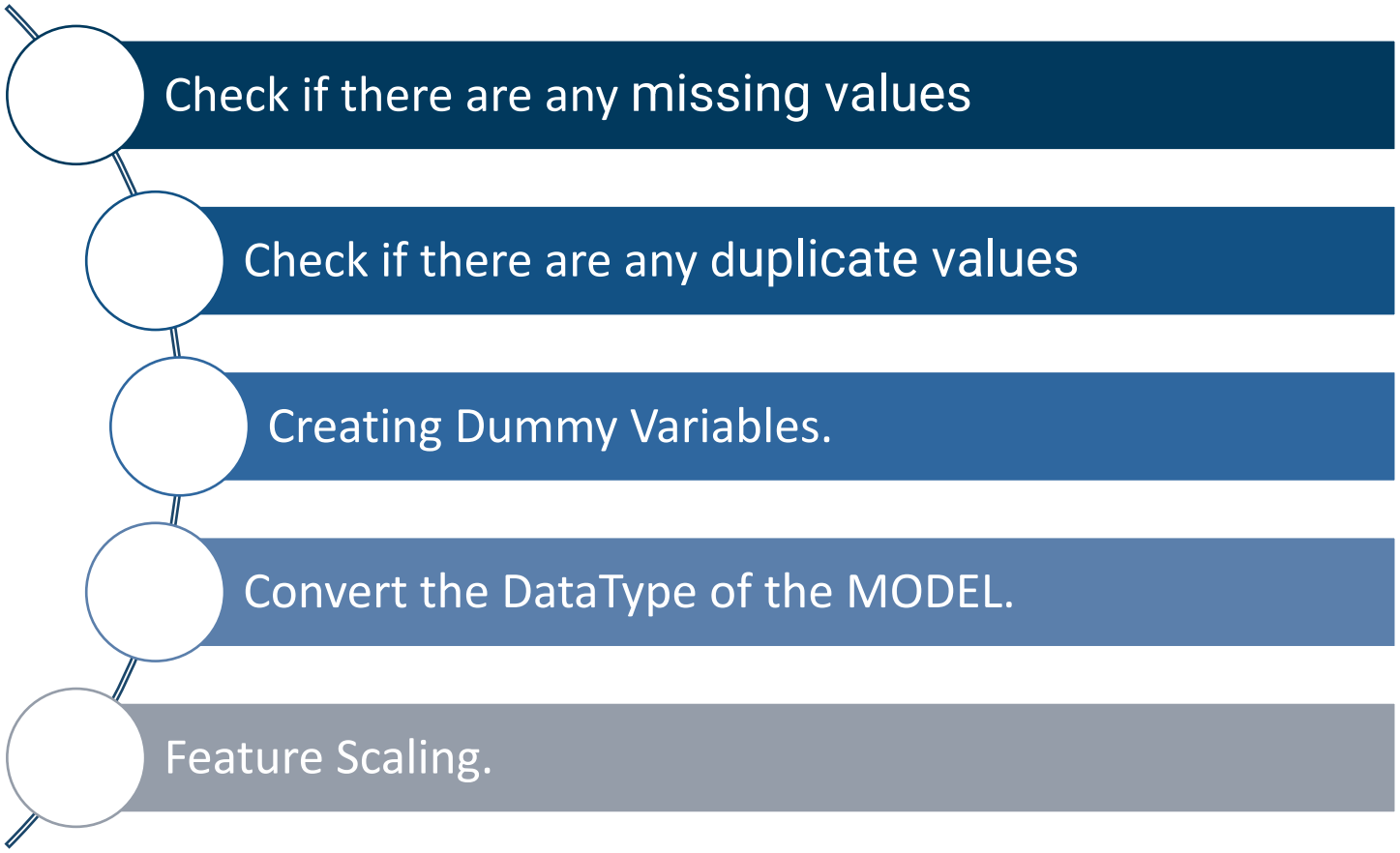
Dataset Description (cont.)

Features Engineering columns:

FEATURE	DESCRIPTION
MODEL	Model year of each car.
CAR_NAME	The name of the car.
CAR_NAME_Other	The name of the car that is repeated less than 25 times.

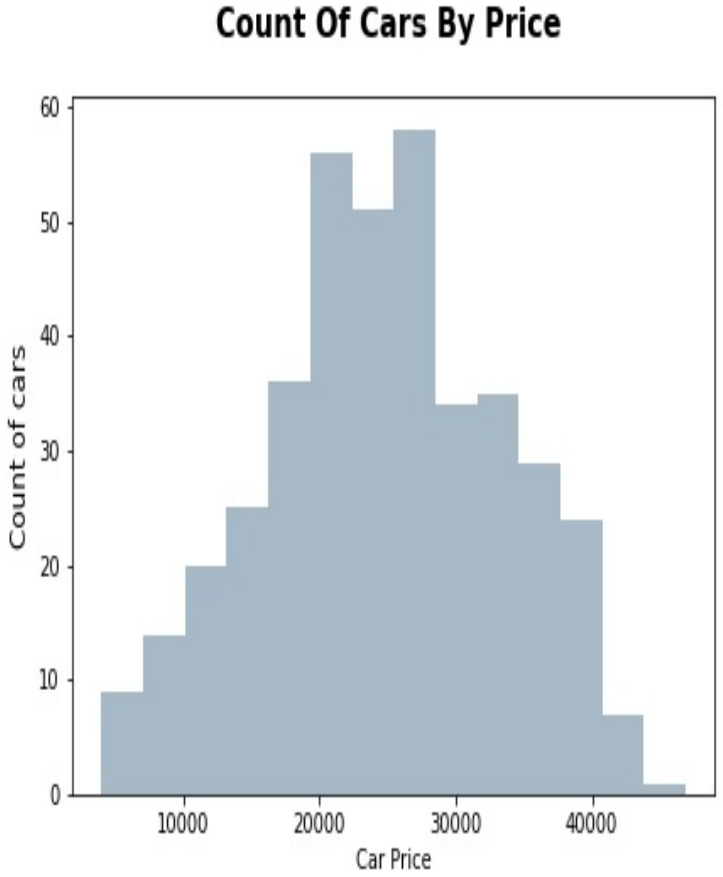
EDA

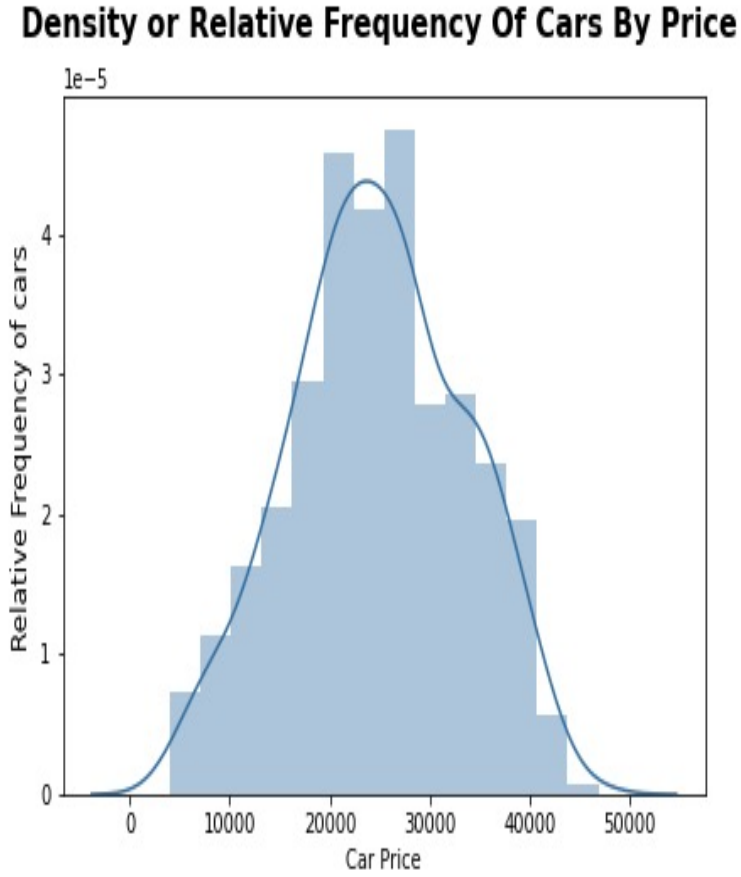
Preprocessing



EDA

Visualizations

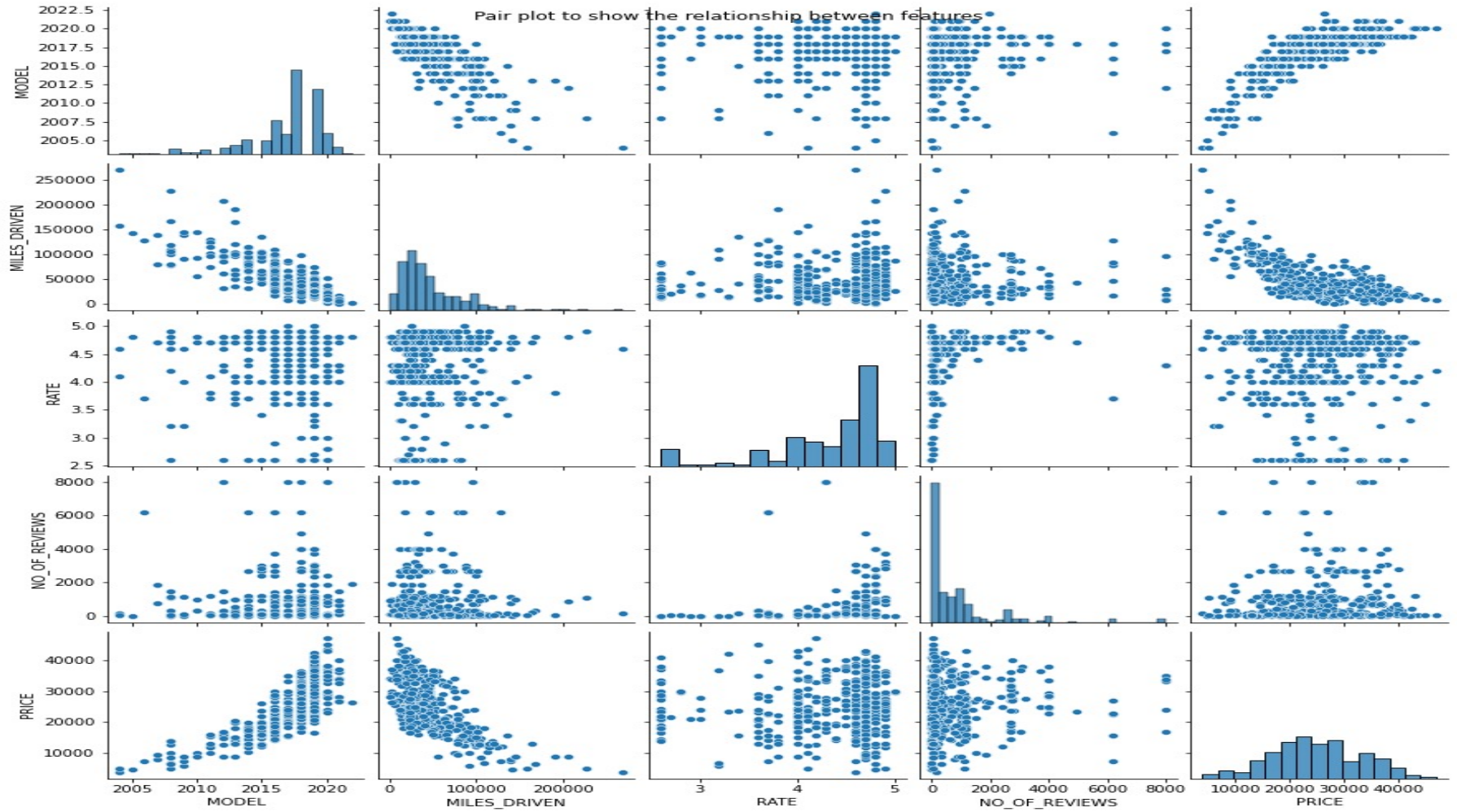




The first plot is a histogram plot to show the Count Of Cars By Price, and the second plot is a distplot that represent the Density or Relative Frequency Of Cars By Price.

EDA

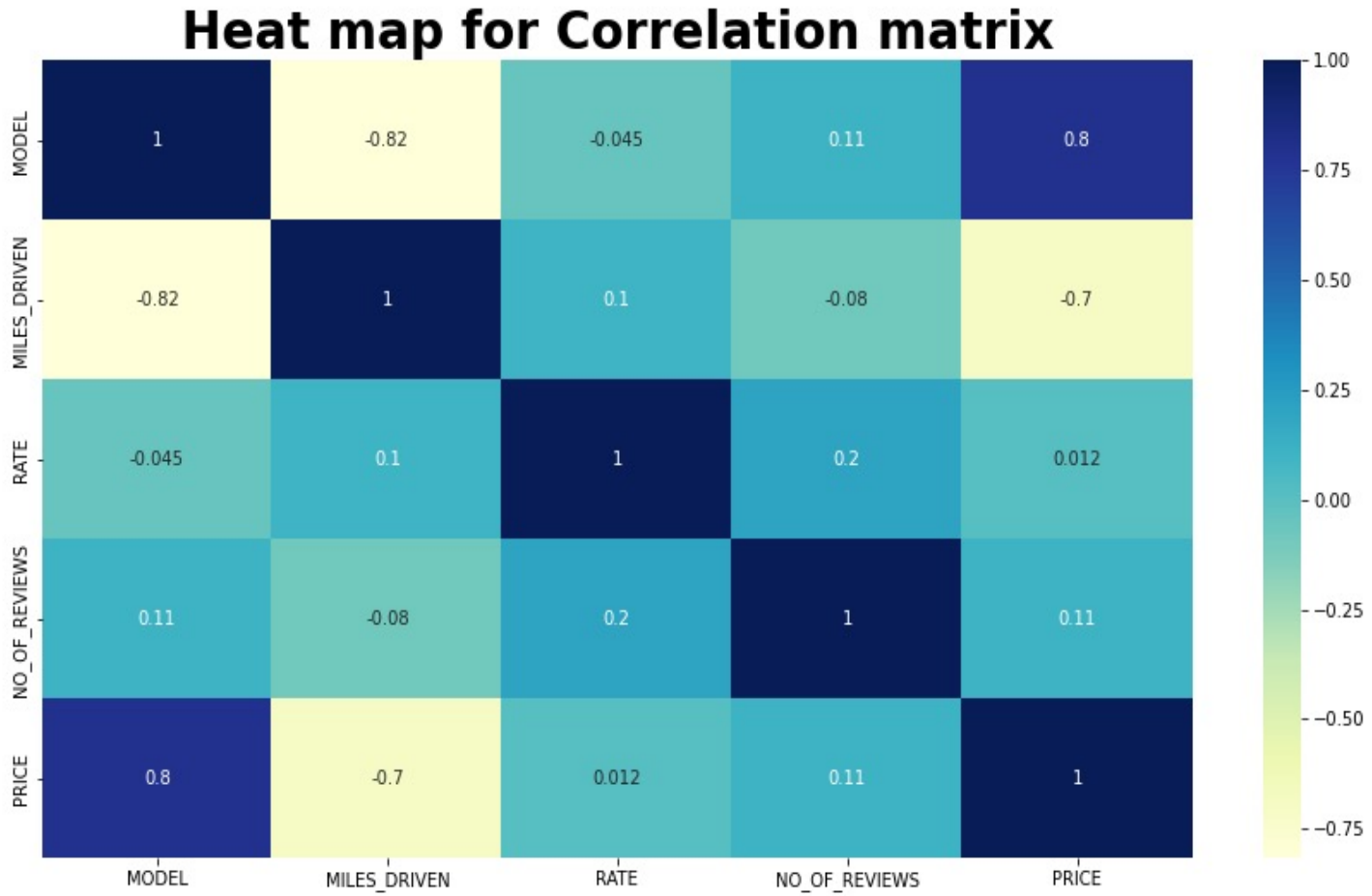
Visualizations



Using pairplot to shows the relationship for (n, 2) combination of features in a DataFrame.

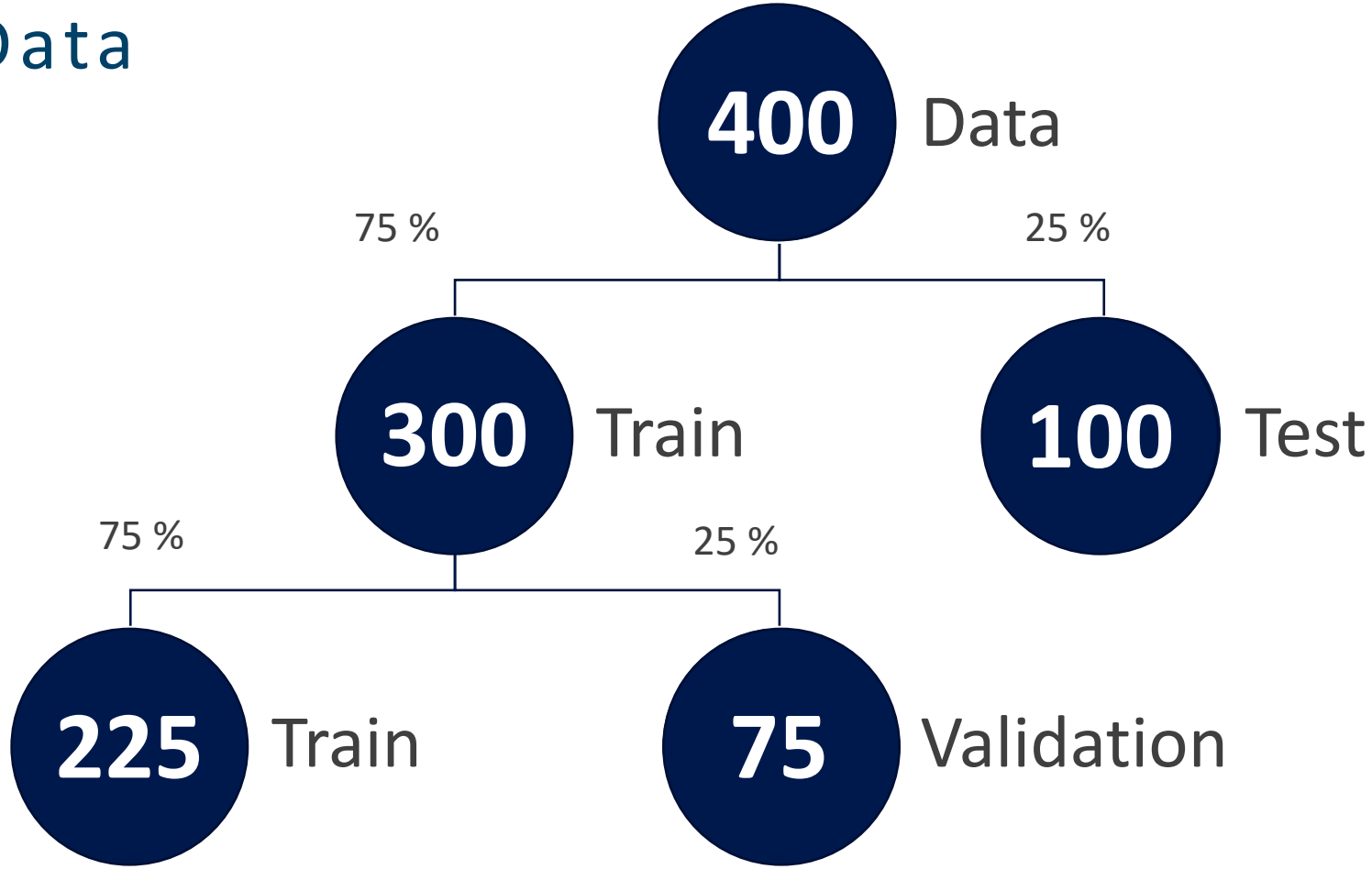
EDA

Visualizations (cont.)



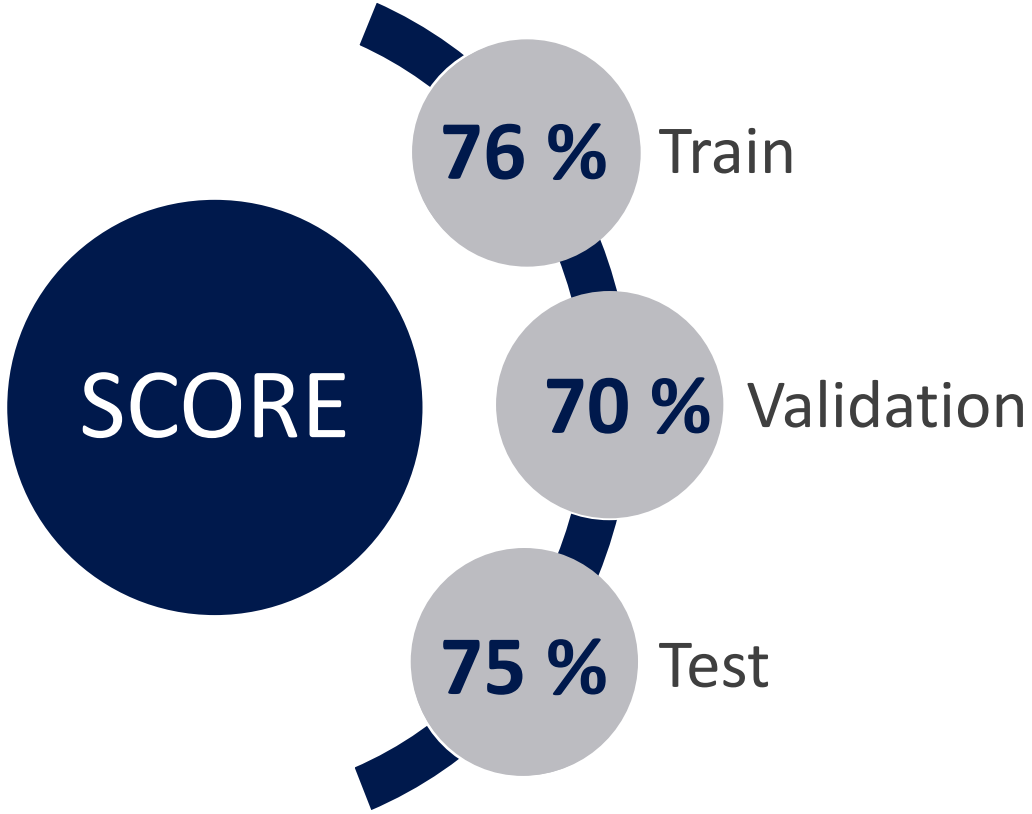
Linear Regression

Split Data



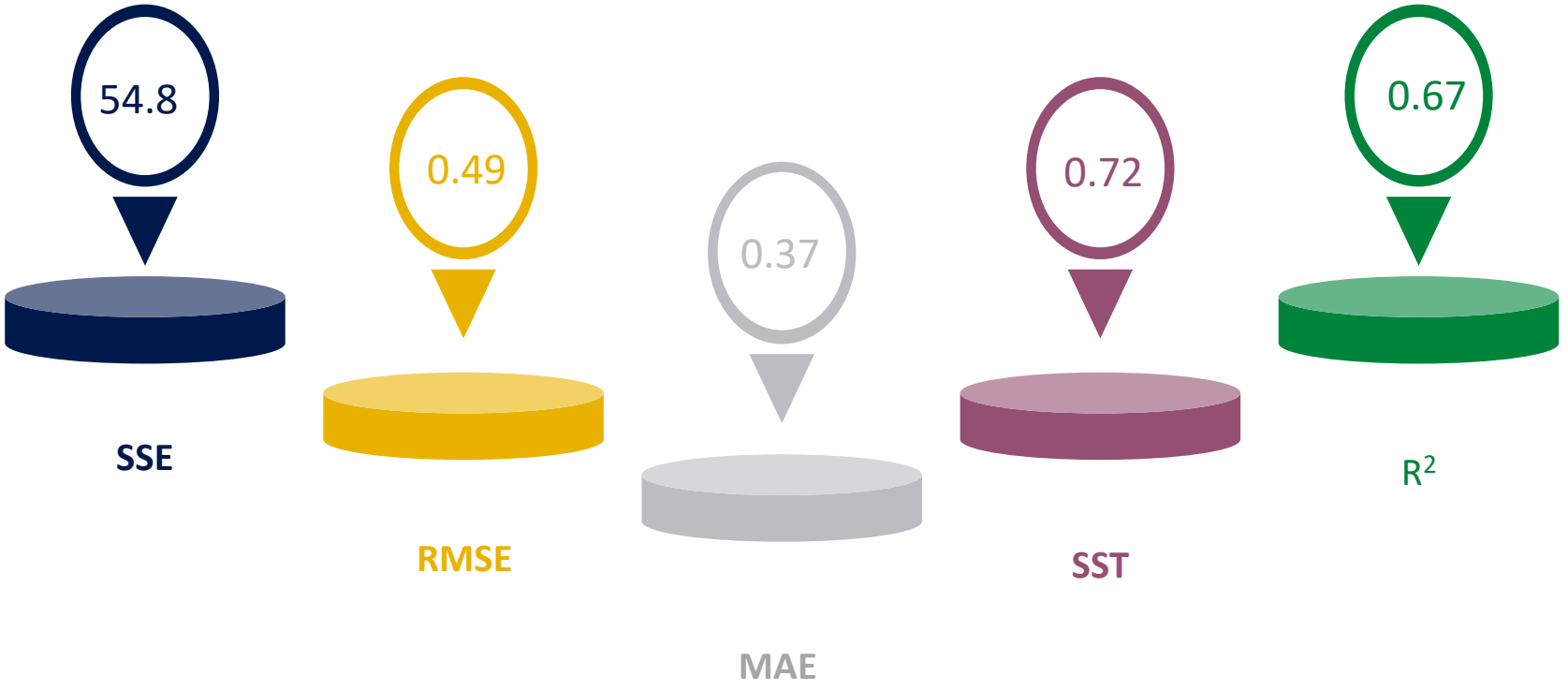
Linear Regression (Cont.)

- Build Model
- Evaluation:



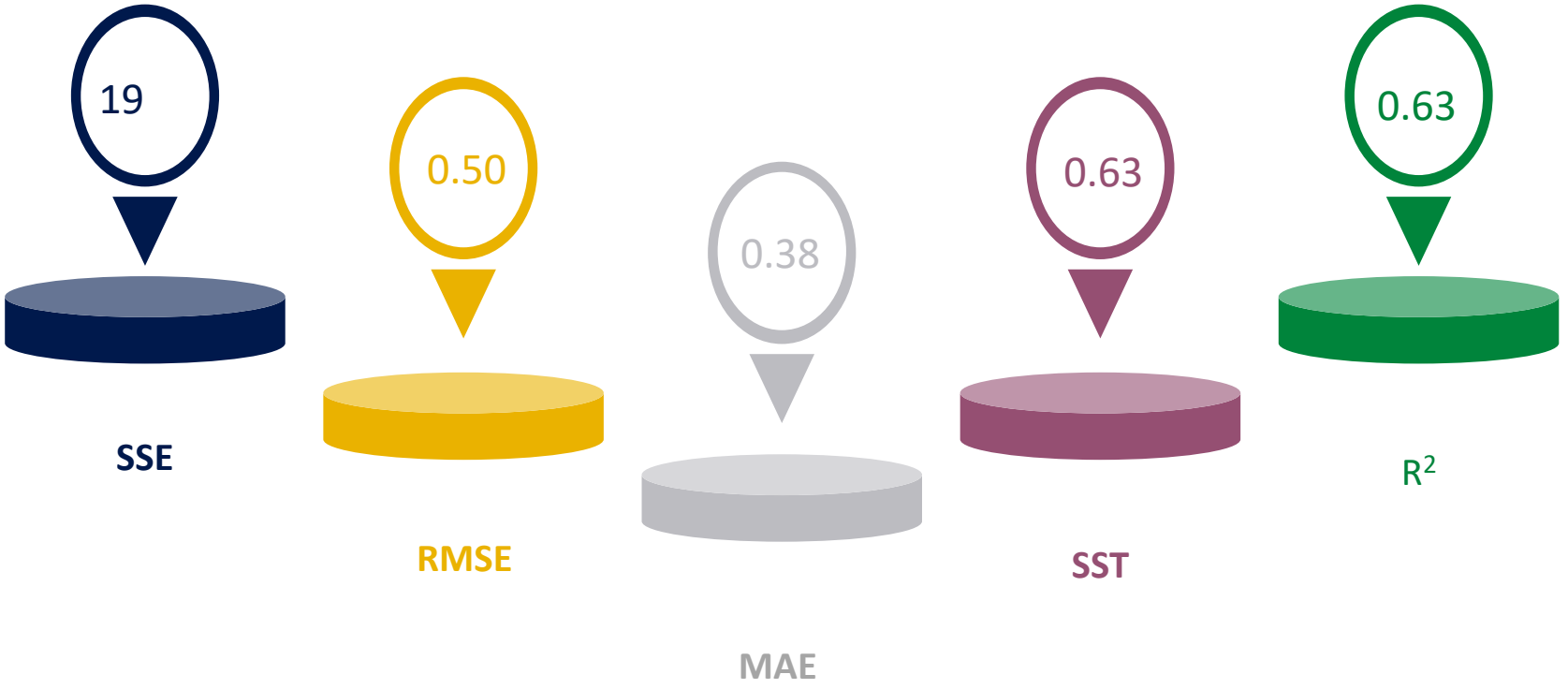
Linear Regression (Cont.)

- Evaluation (Train):



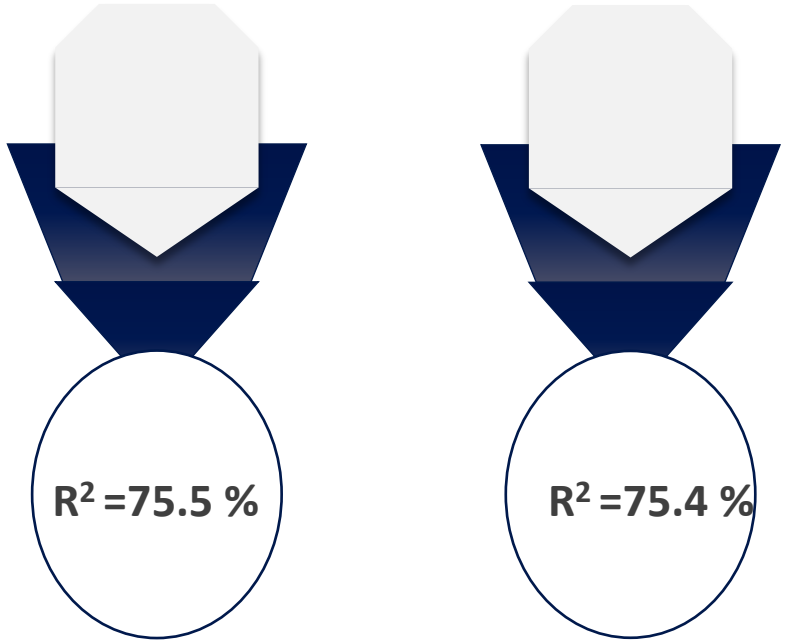
Linear Regression (Cont.)

- Evaluation (Validation):



Linear Regression (Cont.)

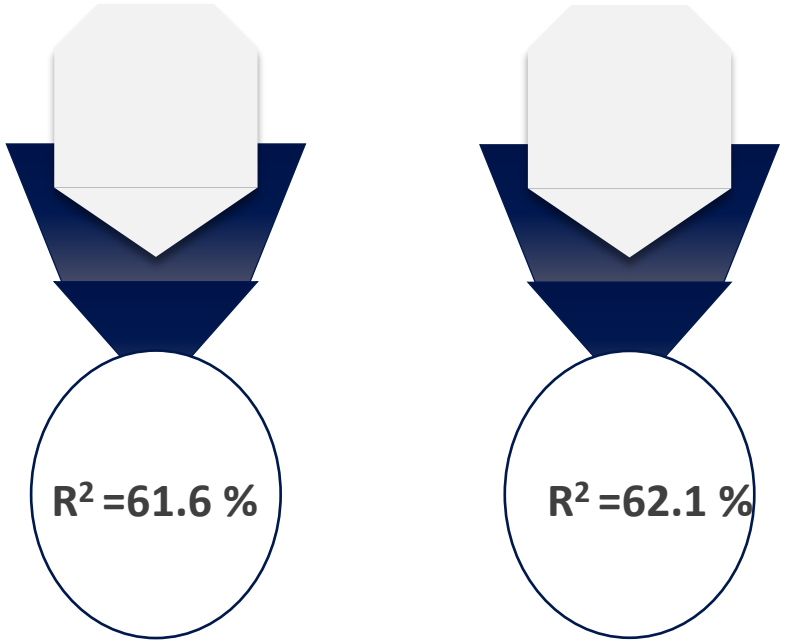
OLS for Train set :



R-squared Adj R-squared

Linear Regression (Cont.)

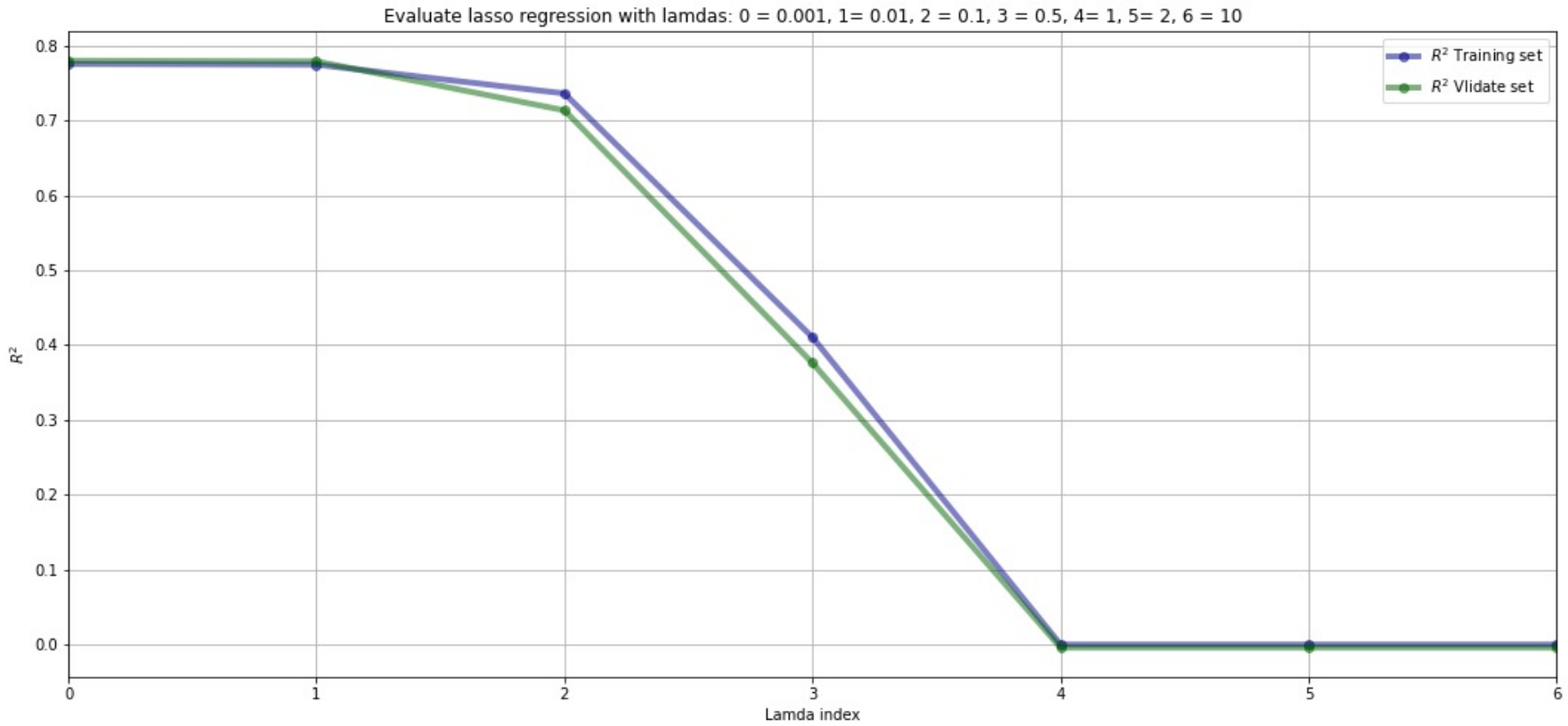
OLS for Validation set :



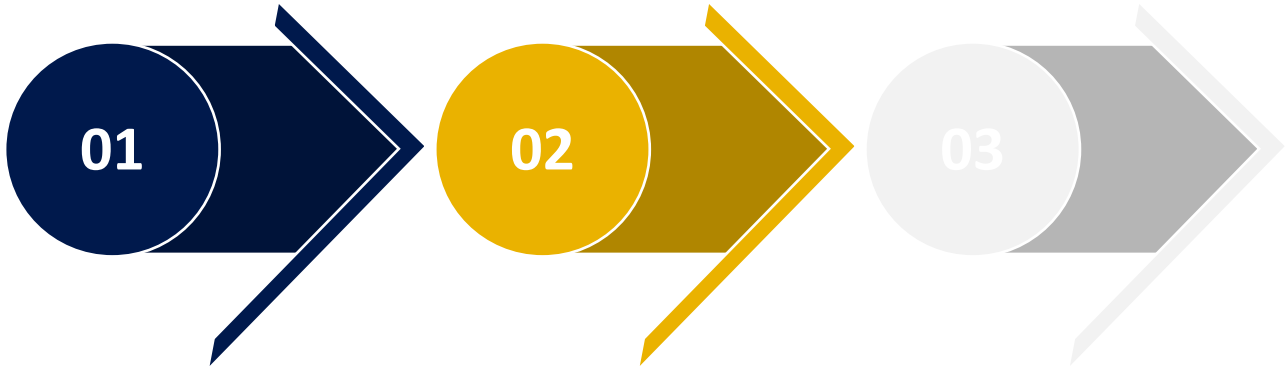
R-squared Adj R-squared

Regularized Linear Regression (LASSO)

Lasso with different Lambdas



Regularized Linear Regression (LASSO) (cont.)



Identify best lambda

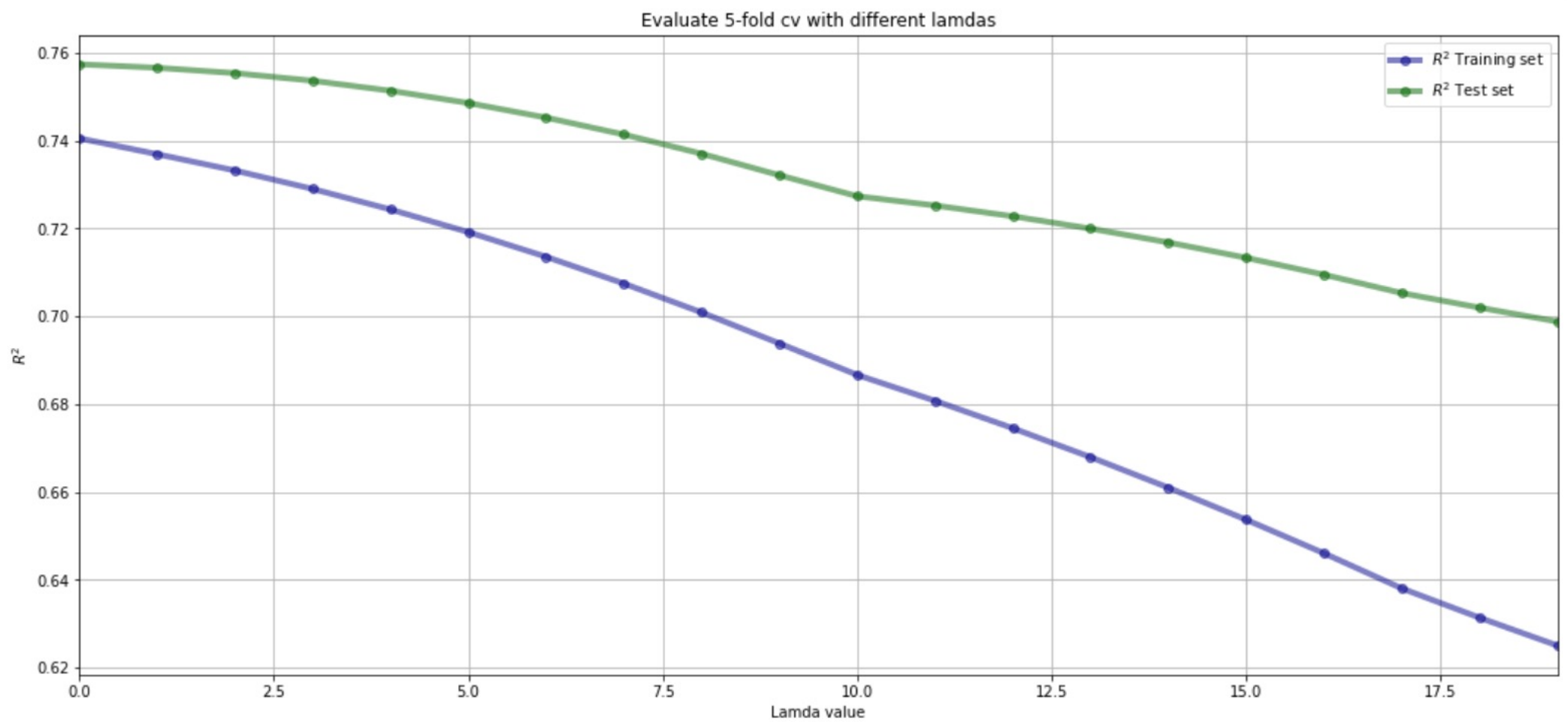
R^2 = 59%

Fit Model

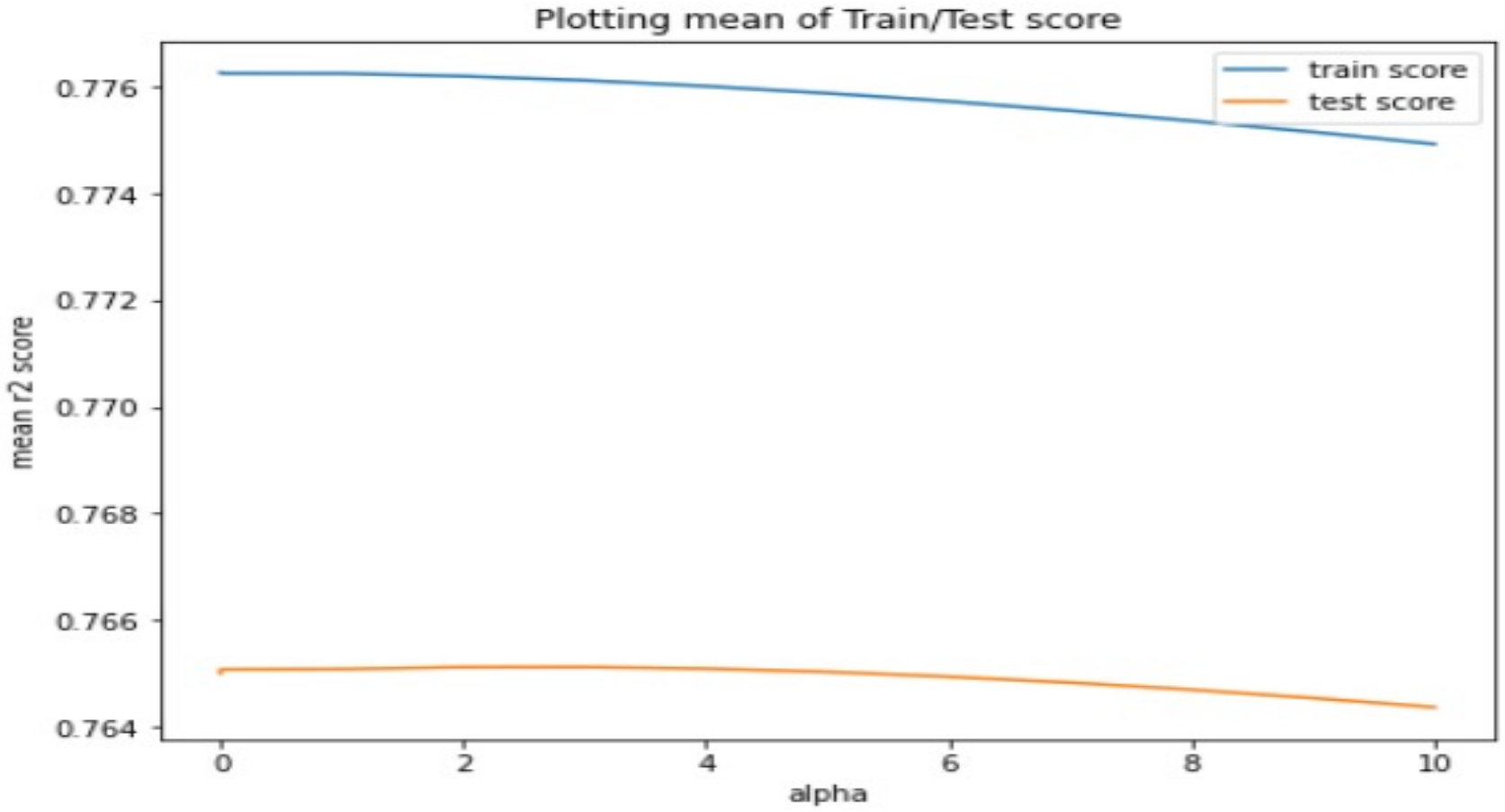
Cross validation using 5 Folds

R^2 = 76%

Regularized Linear Regression (LASSO) (cont.)

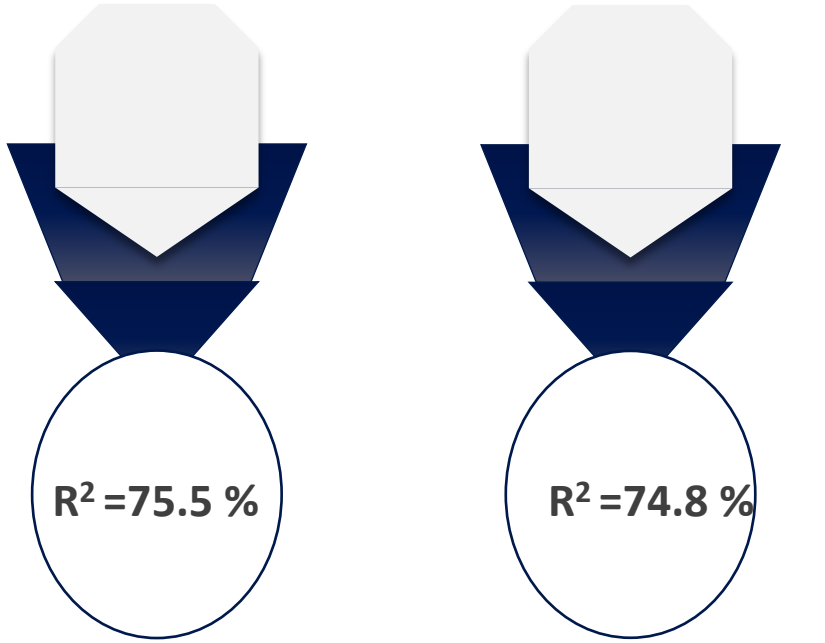


Ridge Regression



Ridge Regression (Cont.)

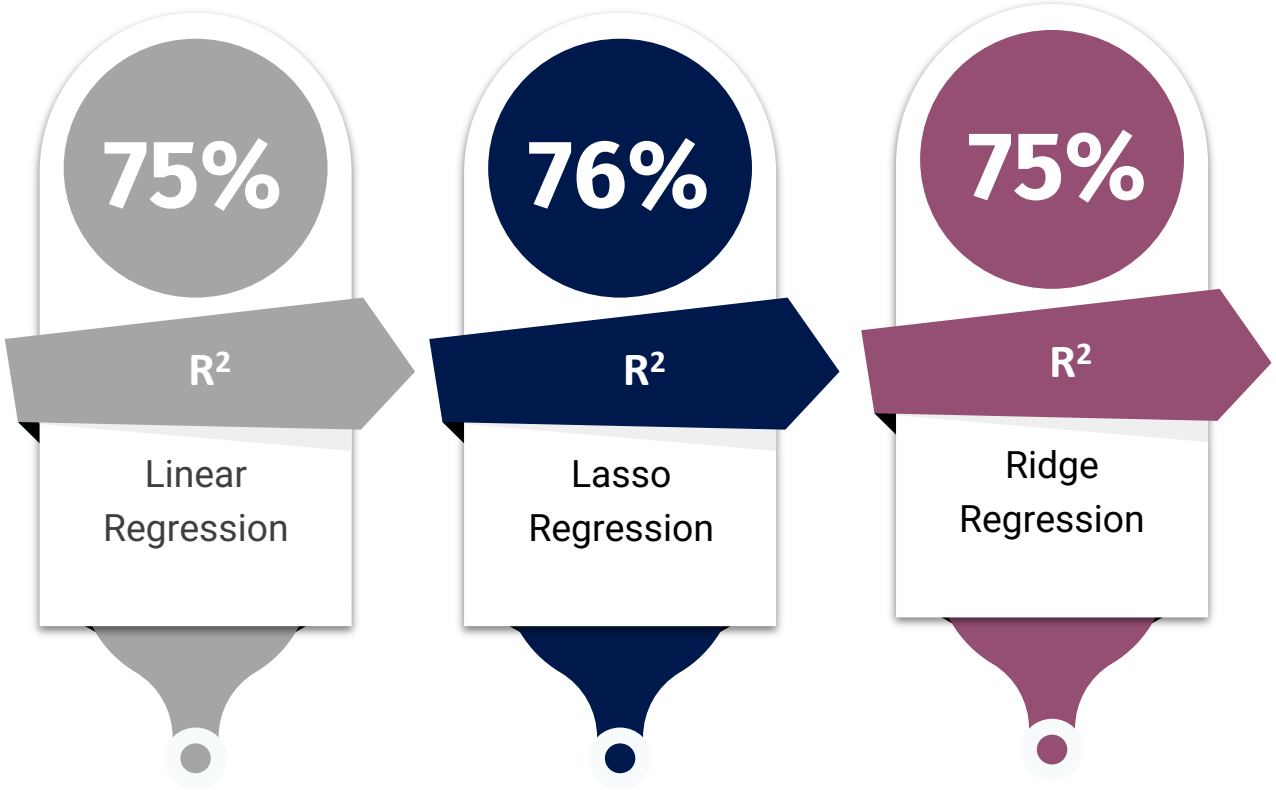
Model Evaluation Ridge Regression:



Model performance on the Train set

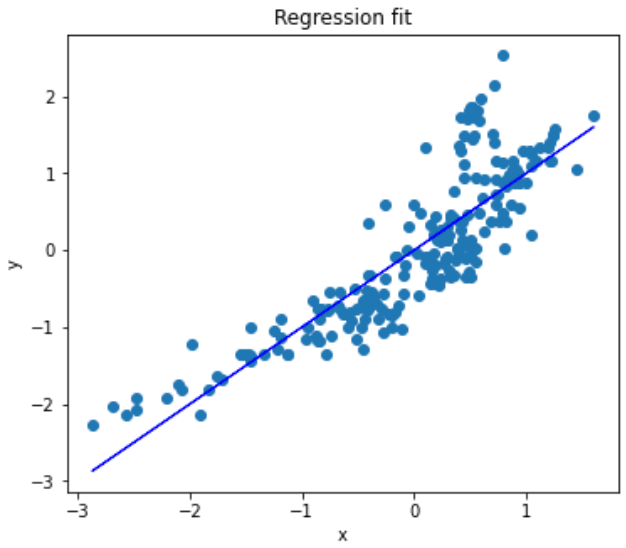
Model performance on the Test set

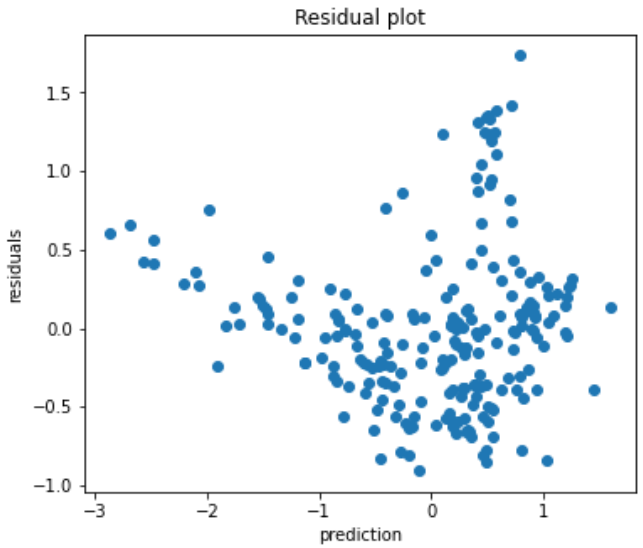
Evaluating the Three models using Test set across these three models

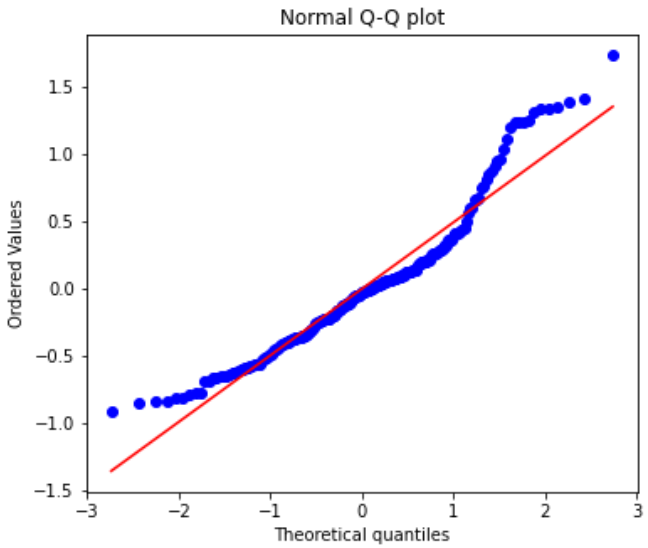


Assumptions

Assumption 1: **Regression is linear in parameters and correctly specified.**



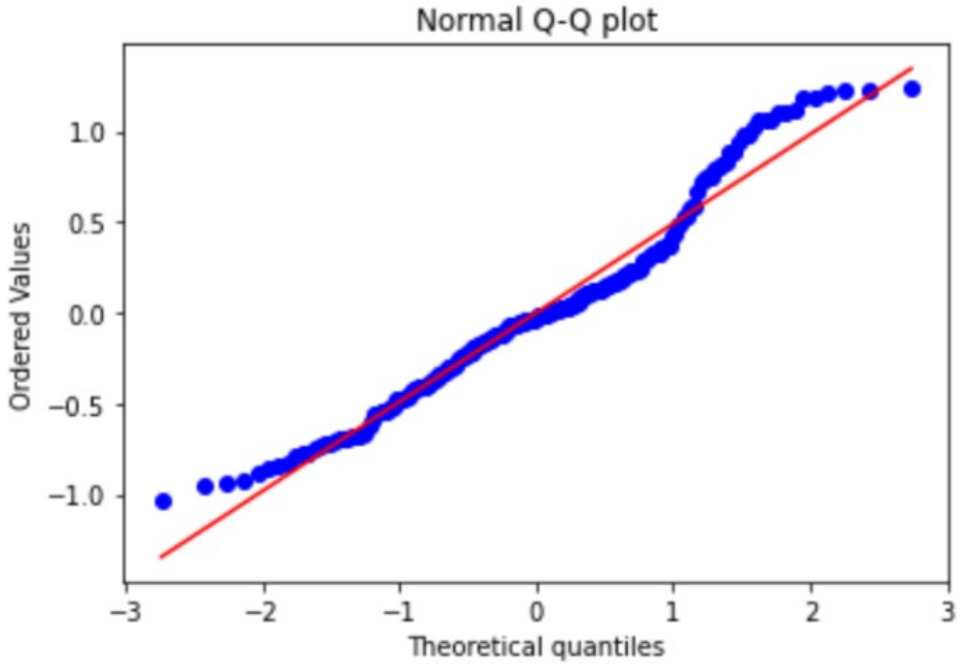




There is a pattern in figure 2, which means the model needs to be improved.

Assumptions (Cont.)

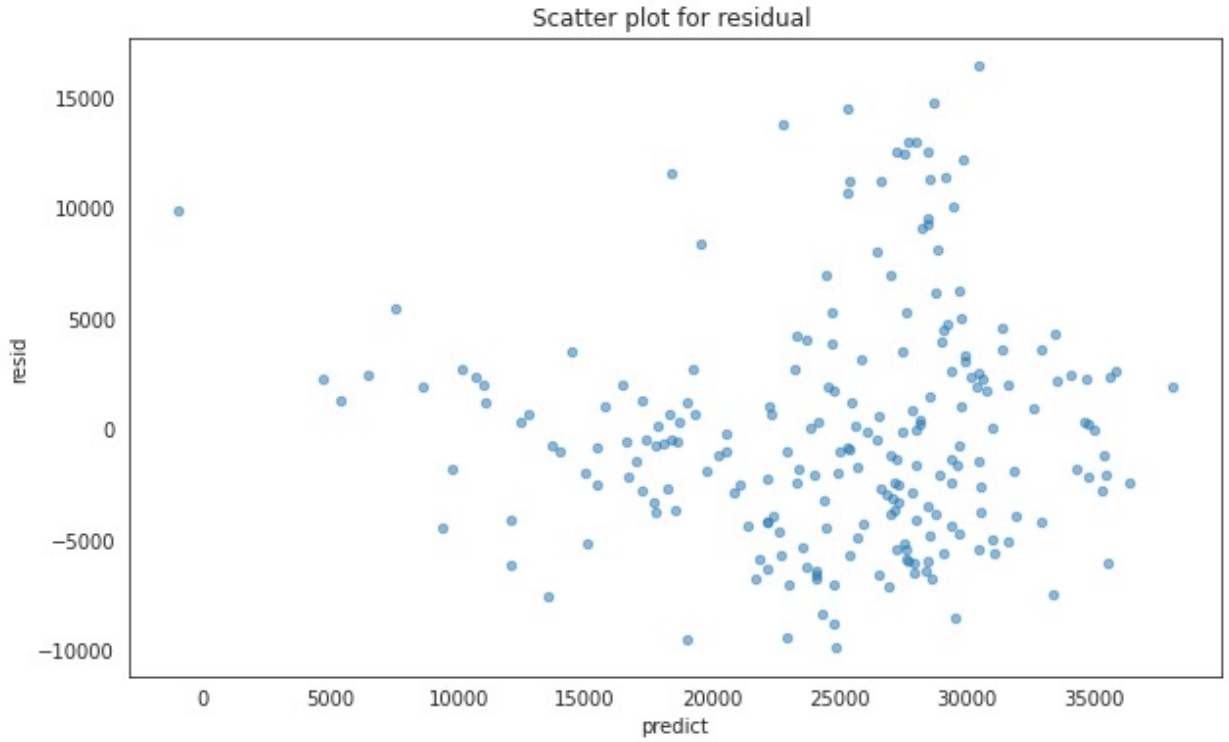
Assumption 2: **Residuals ($e_i = Y_i - \hat{Y}_i$) should be normally distributed with zero mean.** (light tailed)



The Q-Q plot is light tailed, which mean it is not normal

Assumptions (Cont.)

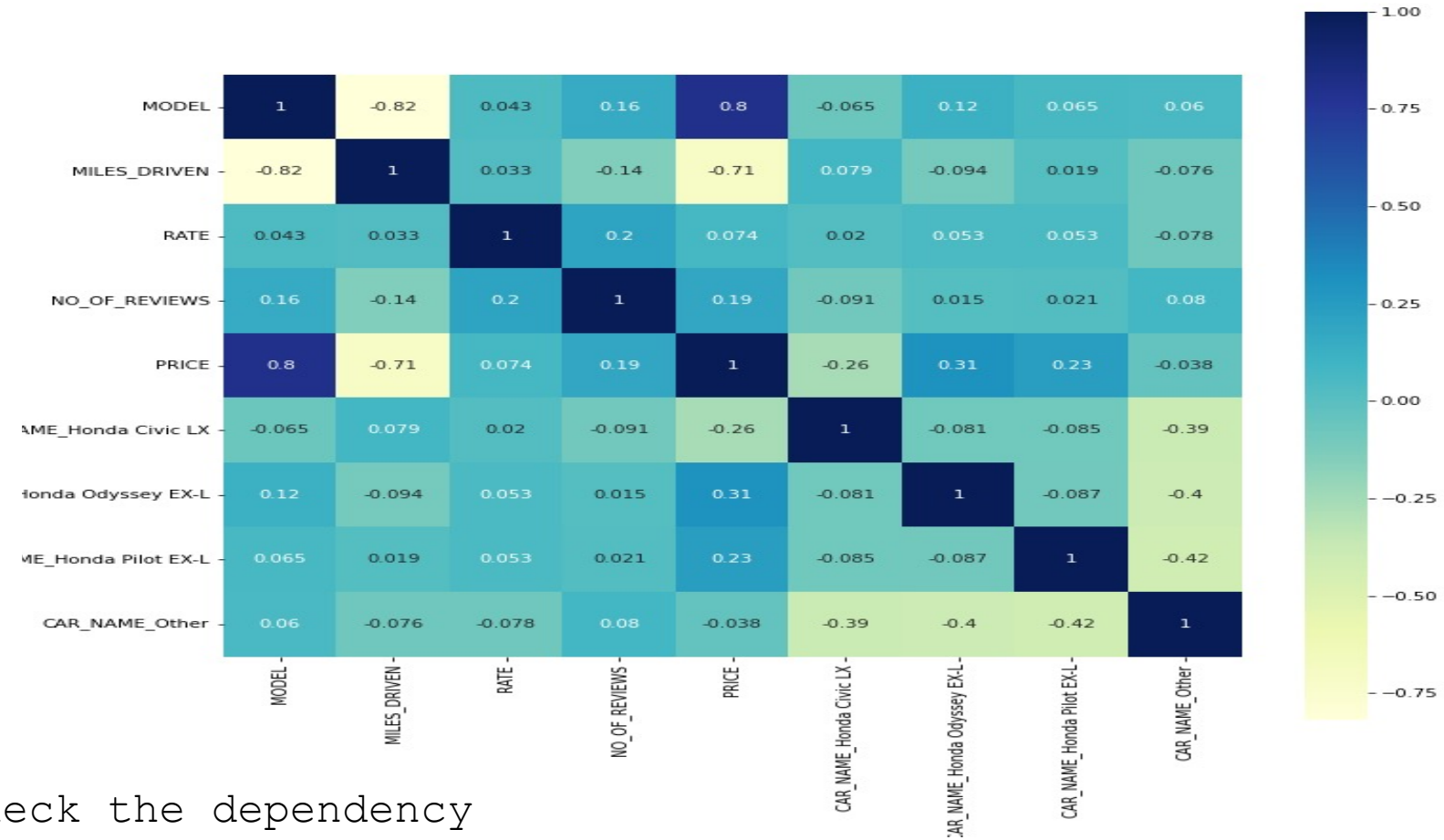
Assumption 3: **Error terms must have constant variance.**



The figure above does not have a constant variance

Assumptions (Cont.)

Assumption 5: **No perfect multi-collinearity**

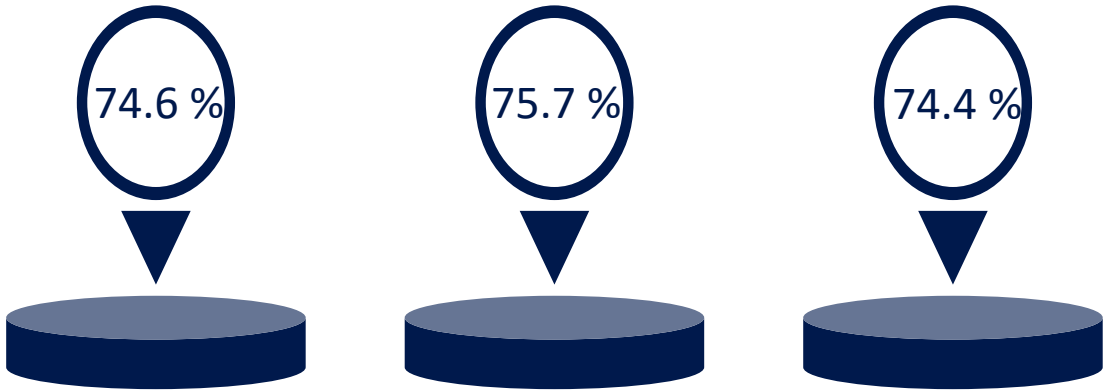


To check the dependency

→ Draw the heatmap and ensure the all feature dependent to each other.

Conclusion

Results of the R^2 (validation) for the three models



Linear Regression **Lasso Regression** **Ridge Regression**

Thank you !