upGrad

#LifeKoKaroLift

Linear Regression: Assignment
PG Diploma in Data Science

upGrad

Course : Data Science

Lecture On : Assignment

Instructor : Sumit Shukla

upGrad

What we will cover in this session?

1 Quick Revision on Linear Regression

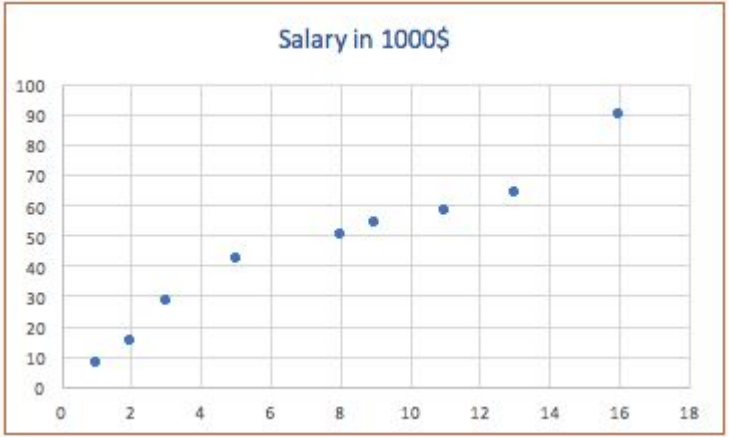
2 Assignment walkthrough

3 QnA

upGrad

Quick Revision

Years of Experience	Salary in 1000$
2	15
3	28
5	42
13	64
8	50
16	90
11	58
1	8
9	54



Quick Revision

Multiple Linear Regression

salary = b1*experience+b2*industry exposure+b3*excel knowledge

But why adding many variables is not good?

Quick Revision

What is Multicollinearity?

- How to detect
- How to handle it

How to check Model performance?

- Adjusted R-Square
- AIC/BIC

How to go about selecting features for a good model?

- RFE
- Manual
- Mixed

Poll Question

How to reduce the number of fields in a categorical variable when it's so many?

- Don't do anything, create dummies, we can remove categories that are insignificant.
- Remove those that are present less in number.
- Combine some field that have same or closer objective

Which model is a good model?

- A model with 10 variables
- A model with 30 variables

Assignment Problem Statement

Geely Auto have contracted an automobile consulting company to help them understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. The company wants to know:

- Which variables are significant in predicting the price of a car
- How well those variables describe the price of a car

What you need to do?

- Create a linear model that describe the effect of various features on price.
- The model should be interpretable so that the management can understand it.

Assignment Steps

Data Preparation

- Extract car-company name from carname column.
- Check the number of car companies present in the data. Did you saw anything interesting?
- Convert the columns to proper data type.
- Create dummies/ perform label encoding.

Model Building

- Divide the data to train and test.
- Perform scaling.
- Divide the data into X and y.
- Perform Linear Regression.
- Use mixed approach if you want.

Assignment Steps

Model Evaluation

- Check the various assumptions.
- Check the Adjusted R-Square for both test and train data.
- Report the final model.

Poll Question

My model R-Square is 89% for train but 50% for test, why so? :(

- You have unnecessary variables in your model.
- Your model is complex.
- You have less variables in your model.

Assignment-Subjective

Steps to answer subjective part

- Answer all the questions.
- You can write the answer using any software but submit the file in PDF format
- You can use images and plots to support your answer.
- Make sure the question is answered with sufficient number of word: No limit
- Please don’t copy for any online available literature.

Assignment-Endnote

What to keep in mind

- Add comments after every cell of code. So that we can understand your approach and method.
- Describe the results.
- For subjective answers, use DOC and type on it, if you wish to add images you can. But convert it to PDF before submitting.
- Create only one Jupyter notebook.
- Submit one zip file with the code and the PDF.
- Use StackOverflow for dealing with syntax errors. Rather than being stuck at one place or waiting for someone to resolve your doubts, take action and use the resources available on the internet to save time.
- Post on the discussion forums for resolving any doubts you have
- Finally, write code manually instead of copy-pasting from the in-content notebooks provided. Builds a habit of writing code. It's okay to look and write, but don't just copy-paste under any circumstance. Because of just copy-pasting, a lot of our students have faced difficulties in the past when they had to write some code on their interview.

upGrad

#LifeKoKaroLift

Thank You!

References:
towardsdatascience