SIT384 Cyber security analytics

Credit Task 5.2C: Linear Regression

Task description:

You are given one dataset "admission_predict.csv".

Its description is as follows:

Context

This dataset is created for prediction of Graduate Admissions from an Indian perspective.

Content

The dataset contains several parameters which are considered important during the application for Masters Programs. The parameters included are: 1. GRE Scores (out of 340) 2. TOEFL Scores (out of 120) 3. University Rating (out of 5) 4. Statement of Purpose and Letter of Recommendation Strength (out of 5) 5. Undergraduate GPA (out of 10) 6. Research Experience (either 0 or 1) 7. Chance of Admit (ranging from 0 to 1)

Acknowledgements

This dataset is inspired by the UCLA Graduate Dataset. The test scores and GPA are in the older format. The dataset is owned by Mohan S Acharya.

Inspiration

This dataset was built with the purpose of helping students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their chances for a particular university.

Citation

Please cite the following if you are interested in using the dataset : Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019

Sample data:

Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
1	337	118	4	4.5	4.5	9.65	1	0.92
2	324	107	4	4	4.5	8.87	1	0.76
3	316	104	3	3	3.5	8	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.8
5	314	103	2	2	3	8.21	0	0.65

(The above data is for demonstration purposes only. Please download the full version of admission_predict.csv.)

You are asked to:

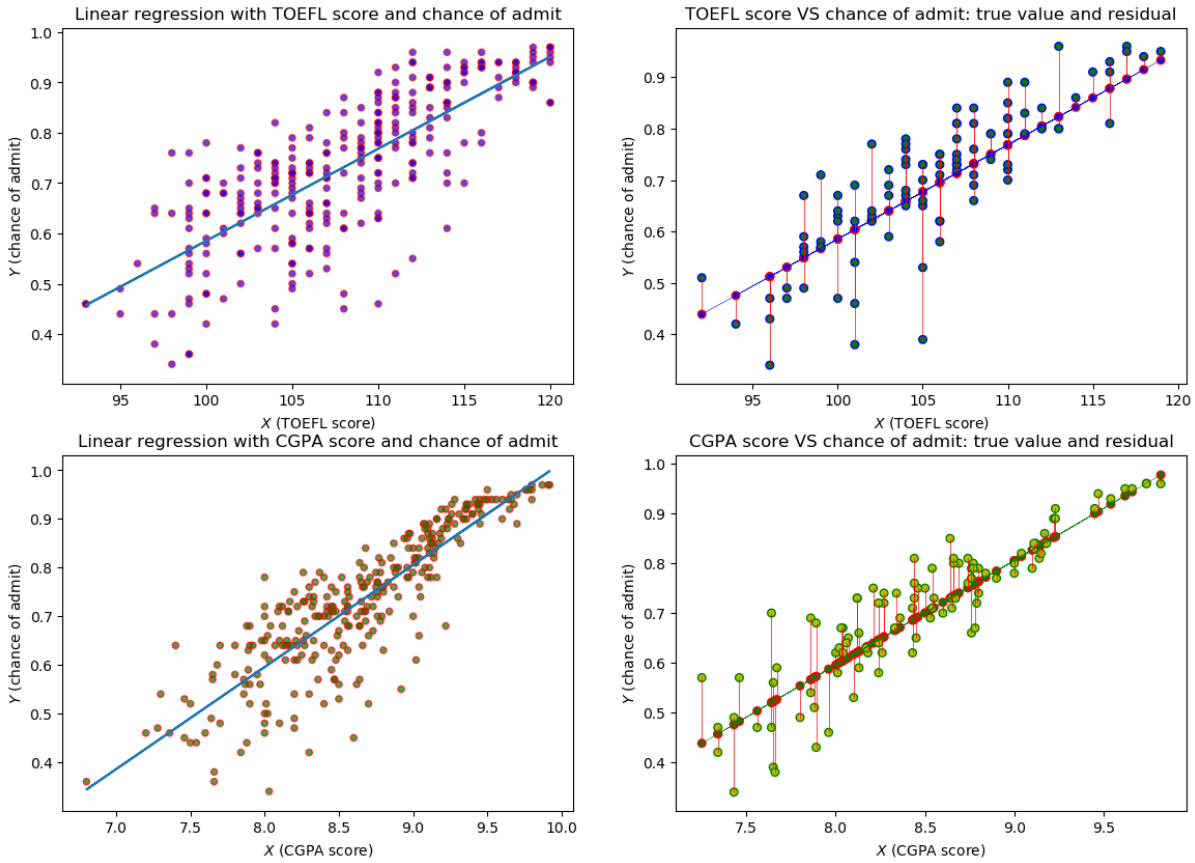
- split the datasets: first 300 records for **training** and last 100 for **testing**,

- build a linear regression model for “TOEFL score” and “chance of admit” from training data,
- build a linear regression model for “CGPA score” and “chance of admit” from training data,
- plot the regression line and the predicted points along the prediction line for the two models,
- plot the true values from testing set and the residual line for the two models.

Use the following settings:

- fig, ax = plt.subplots(nrows=2, ncols=2, figsize=(14, 10), dpi=100)
- X axis is TOEFL score or CGPA score
- Y axis is chance of admit
- Plot colors of your choice

Sample output as shown in the following figure is for demonstration purposes only.



Submission:

Submit the following files to OnTrack:

1. Your program source code (e.g. task5_2.py)
2. A screen shot of your program running

Check the following things before submitting:

1. Add proper comments to your code