Kaggle_Dataset Analysis

Aim of the Problem:

The provided code aims to analyze a dataset related to graduate admission, specifically focusing on factors such as GRE scores, TOEFL scores, university ratings, statements of purpose (SoP), letters of recommendation (LoR), CGPA, and research experience. The objective is to predict the chance of admission using regression modeling techniques.

Approach of the Problem:

The code begins by loading the dataset from the 'Admission_Predict_Ver1.1.csv' file. The data is read using the csv.reader function, and the header row is skipped.

The code takes in the parameters GRE, TOEFL, UniRat, SOP, LOR, CGPA, Research and ChanceAdmit given in the .csv file as the parameters and creates an array using them, and appends all of the values present in the .csv file.

Then a matrix M is constructed, which has 8 parameters taken as the basis on which linear regression is to be performed.

By initially taking the parameters as a linear dependence and plotting the various plots of (GRE, ChanceAdmit) and the other parameters, we see that the dependence of ChanceAdmit is hugely higher on CGPA, TOEFL and GRE scores.

Then the np.linalg.lstsq takes in the parameter matrix M and the ChanceAdmit as inputs to perform regression on.

Post that, we plot the values of the predicted versus the actual ChanceAdmit using the plt.scatter function.

The 8 parameters obtained by multiple hit and trials were GRE, TOEFL, and CGPA raised to the zeroeth, first and second power (as CGPA was highly correlated to the ChanceAdmit),the LOR, the SOP raised to the second power, the product of research and LOR, and the product of TOEFL and UniRat.

The parameters inputted into the M matrix need to be carefully taken into account as they influence the Predicted ChanceAdmit.

The code performs data analysis by creating a design matrix M which combines various features. The linear regression model is fitted using the np.linalg.lstsq function, and the regression coefficients are calculated.

Regression Model:

The linear regression model is expressed as:

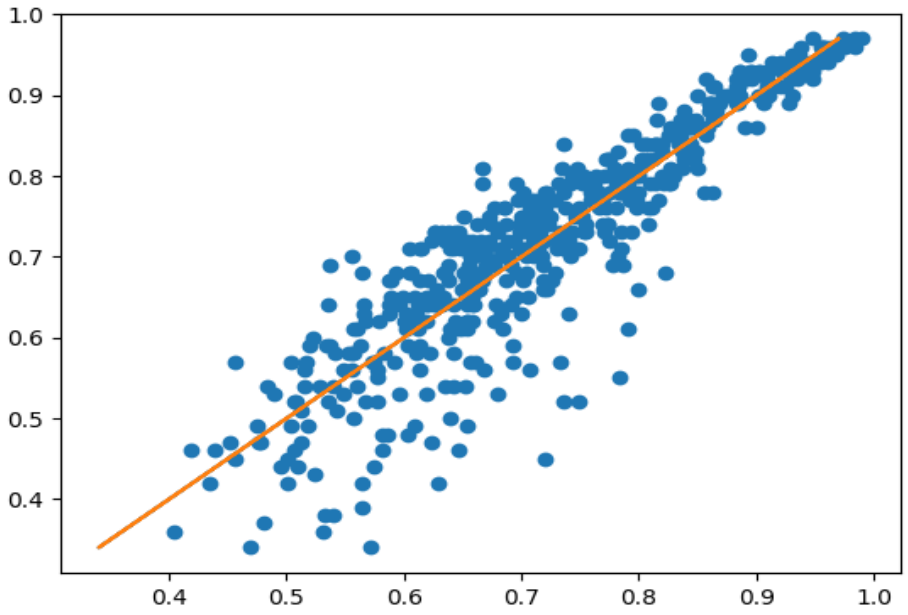
M = np.column_stack([GRE**1, TOEFL**1, CGPA**0, CGPA**2, CGPA**1, LOR**1, SOP**2, Research*LOR, TOEFL*UniRat**1])

p, _, _, _ = np.linalg.lstsq(M, ChanceAdmit, rcond = None)

The code then attempts to create a scatter plot and a line plot to visualize the model's predictions. It also attempts to find the correlation coefficient (in the comments).

Results:

The actual admission values are stored in Actual Admit, and the predicted . The code calculates the correlation coefficient between the actual and predicted admission values, in the commented out section, and the coefficient is 0.82, which is close to 1, showing a high value of correlation.
The obtained plot:



The obtained equation is 0.0018852432690599841*GRE**1 + 0.002745108076999792*TOEFL**1 + -1.8753159821390752*CGPA**0 + -0.008687338713741083*CGPA**2 + 0.2640138532893347*CGPA**1 + 0.012803025740327035*SOP**1 + 0.0004201704925007233 5*Research*LOR + 0.007054014368358375*TOEFL*UniRat**1

Discussions:

I had discussed the values of the parameters to be taken in for the M matrix with ee22b122 and I had also had a few discussions with a few other classmates regarding the approach to be followed for the particular problem, including the curve_fit approach.