1 Preliminary Analysis

I started off by plotting each of the parameters against chance of admit to get an idea of the data representation. Most of them had a simple relation , approximately directly proportional . To confirm this I tried curve fitting each of the parameters with the Chance of Admit as both linear approximations and quadratic approximations and then compared the error between the 2 fits. The result was as follows :

linear error	quadratic error	best fit
0.0068257023198923916	0.006789580643073724	quadratic
0.0074031107165580985	0.007399316352392437	quadratic
0.010411904712629224	0.010365868353573187	quadratic
0.010575720344772688	0.010500328282423935	quadratic
0.011600523035201177	0.011594261370230594	quadratic
0.004400570156519554	0.004390968335692146	quadratic
0.013956795649350645	0.013956795649350652	linear

From this table it is evident that the most of the independent variables (Scores , CGPA , etc)have a quadratic relation with the dependent variable (Chance of Admit)

But the msq error difference between linear and quadratic fit is negligible , so we can conclude that a linear model will work very well for this dataset.It is not nexessary to check the data for exponential , logarithmic , sinusoidal etc patterns , because if the data did follow any of these patterns , the quadratic fit would have a significantly lower mean square error than the linear msq error. This is clearly not the case in the given dataset.

I proceed to make a linear model of the following form :

$$f(t_1, t_2, t_3, t_4, t_5, t_6, t_7, a, b, c, d, e, f, g, h) = a * t_1 + b * t_2 + c * t_3 + d * t_4 + e * t_5 + f * t_6 + g * t_7 + h$$

I then use numpy.linalg.lstsq to solve for the 8 variables a,b,c,d,e,f,g,h .

- a = 0.00185851 , GRE Score
- b = 0.00277797 , TOEFL Score
- c = 0.00594137 , University Rating
- d = 0.00158614 , SOP
- e = 0.01685874 , LOR
- f = 0.11838505 , CGPA
- g = 0.02430748 , Research
- h = -1.27572508 , Constant factor

I have now plotted serial number vs Chance of Admit(Actual) and serial number vs Chance of Admit (Predicted) in 1

The mean square error between this lineaer data fit and the actual Chance of admit is 0.0035407508622541037

There are several possible reasons for why we don't get a highly accurate model for the Chance of Admit , A few are listed below

- Unavailable data , there may be some more factors affecting the Chance of Admit which were not recorded but they would still be affecting the outcome.
- Factors such as Research are taken in binary form , but in reality , quality of research is important and it should have a range of values , but again this is hard to quantify

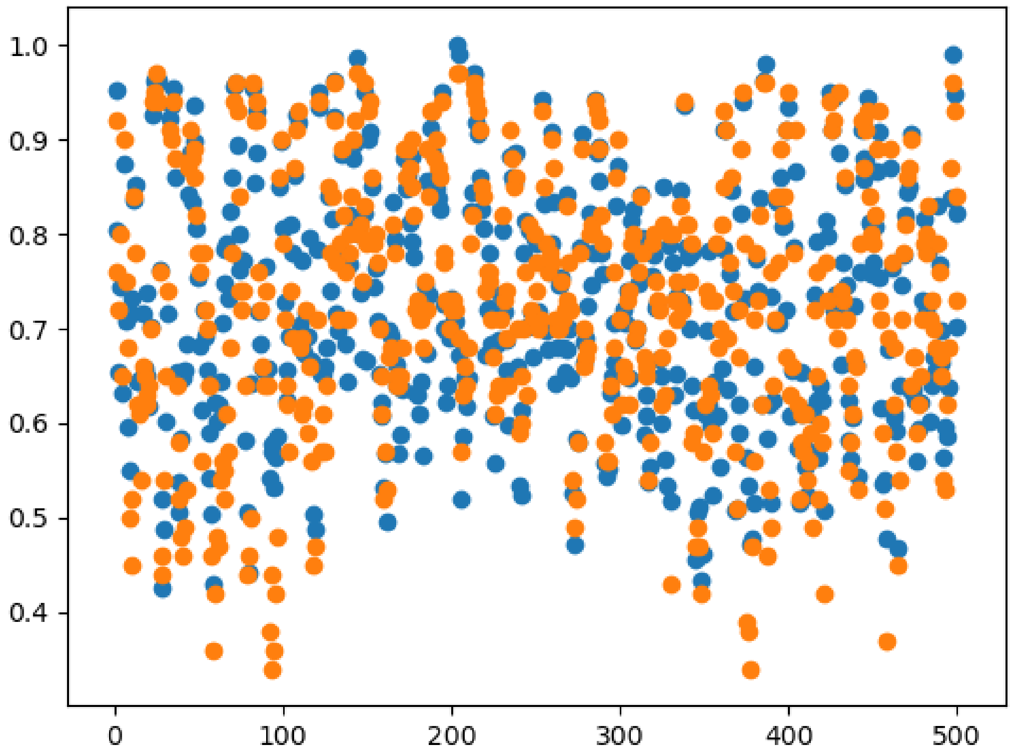

Figure 1: Results for the linear fit

1.1 Where to direct efforts?

At first look of the data , we would think that CGPA is the most important factor , followed by TOEFL score and then GRE score , because they have the highest coefficients in the linear fit. But this is a wrong analysis because the values taken by CGPA are in the range of 0 to 10 , but the GRE Score is on a scale of 0 to 340 . Therefore we have to normalise the coefficients by multiplying them by the range of each of the parameters . In this case the range is 0 to Max value. The normalised coefficients are :

- a = 0.6318922049034703 , GRE Score
- b = 0.3333566869703507 , TOEFL Score
- c = 0.029706840200885015 , University Rating
- d = 0.007930687278831549 , SOP
- e = 0.08429371176209421, LOR
- f = 1.1743797303007597 , CGPA
- g = 0.024307478582166024, Research

From this we can see that the highest weightage is going to CGPA , followed by GRE Score and then TOEFL Score. If we had not normalised the values , our analysis would have been wrong .

To get admitted in a University of rating 5 , a seperate analysis needs to be done taking in the data of all applicants who applied to tier 5 universities. That has been omitted here due to page limits , but the results are the same , Primary focus should be on *CGPA* followed by *GRE* and then *TOEFL.*