Data analysis-APL

Lakshmiram S

October 2023

Introduction

We have data on various variables that possibly affect a student's chances of getting into a university. Our task is to come up with a model that would predict the probability of getting in(prob), as a dependence on the various parameters mentioned.

Approach and analysis

The various parameters mentioned are: cgpa(cg), gre score(gre), toefl score(toefl), sop, lor, university rating(uni-rating) and research.

- Before beginning to analyse the data, I normalise all the independent variables, so that the coefficients given by the model will then be the weightage of that variable in determining prob.
- First I try to understand which variables might actually affect prob. In this line of reasoning, i plot prob vs uni-rating and i get a counter-intuitive graph; there is a positive correlation between prob and uni-rating. From this, i understand that the uni-rating might be another dependent variable and so I decide to ignore it as one of the deciding "independent" variables that affect prob.
- Second, I plot prob vs cg and I observe a good correlation between the 2; i use scipy.optimize.curvefit to understand what polynomial best approximates the relation and the result is that $prob\,\alpha \approx cg^{1.58}$
- Now I just do Linear Least squares regression, treating all the variables as independent. I measure the Mean Square Error(mse) to understand how good my model is.
- I now use that fact that prob is proportional to $cg^{1.58}$ and use this in my least square fit and I see an immediate reduction in the mse value.

Observation and Inferences

the final expression for prob as a function of the given parameters is:

-193.56068653363013 gre + 0.3262198710463339 toefl + 0.045257585117966924 sop + 0.08374711916894897 lor + 1.0773586059377631 cg**1.585 + 0.04525004261586248 research.

We note that the coefficient of cgpa is the highest positive number and gre has a negative correlation.(I'm not really sure how to explain it, my guess is that it is a consequence of not including uni-rating as an independent variable, although I see no reason why I should do so, after my prior reasoning.)

Here are some plots to support my analysis

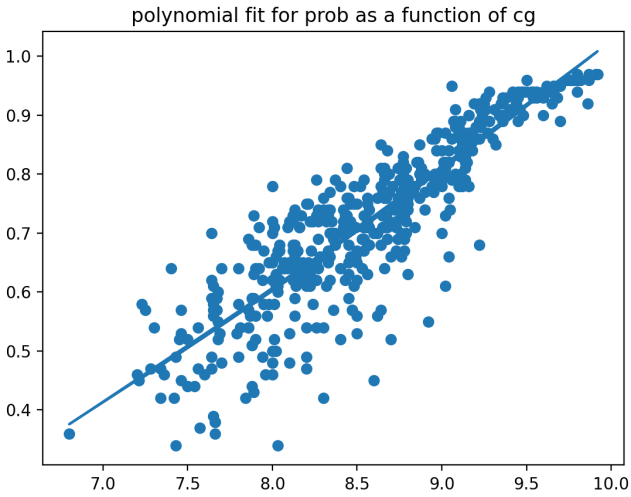

Figure 1: Your Image Caption

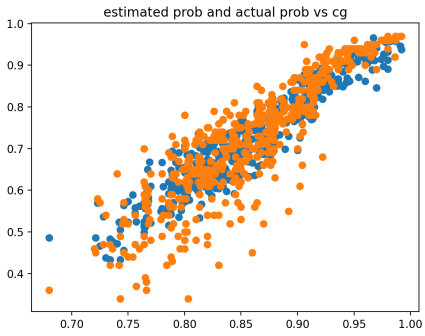

Figure 2: mse = 0.004762531915998518

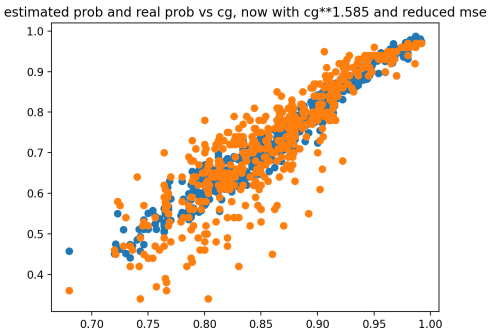

Figure 3: mse = 0.00407174118819384 44