Dataset Facebook Regression Report

A: Summary:

The Facebook dataset is downloaded from UCI Machine Learning Repository. After restructuring the model and now we have 10 independent variables, 1) lifetime post total reach, the number of people who saw a page post, 2) lifetime post total impressions, the number of times a post from a page is displayed, 3) lifetime engaged users, the number of people who clicked anywhere in a post, 4) lifetime post consumptions, the number of clicks anywhere in a post, 5) lifetime post impressions by people who have liked a page, total number of impressions just from people who have liked a page, 6) lifetime post reach by people who like a page, the number of the people who saw a page post because they have liked that page, 7) lifetime people who have liked a page and engaged with a post, the number of the people who have liked a Page and clicked anywhere in a post, 8) comment, number of comments on the publication, 9) likes, number of “Likes” on the publication, 10) shares, number of times the publication was shared. Meanwhile, we have one dependent variable called popularity. In this data case, we have 496 lines dataset all together.

(1) The analysis of the specific variables.

First of all, I use average, variance, max, min, median, mode, graph and correlation to analyze the performance of all independent variables. For example, we have the graph 1 shows the relationship between the lifetime engaged users and popularity. And we can saliently find that the majority of the performance of the variables works fine comparatively.

Graph 1

After that, I use the module correlation get the whole relation map between each other. I want to figure out that which have strong relation with y, which means that the correlation coefficient should be larger than 0,4, and I also need to check that between which two independent variables there is a strong tie which shows that the correlation coefficient is large than 0,7. In principle, we want that should have strong connection with y instead of others. In table 1, which is the part of the correlation map, and I use ABC to denote every variable in order to make it simple to read. And we notice that B and E have weak tie with K, and A has strong connection with F, also for C and G. This information will help us understand the relations between all these variables and may later give us support to better improve the regression model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E |
| A | 1 |  |  |  |  |
| B | 0,69453357 | 1 |  |  |  |
| C | 0,56927037 | 0,36741607 | 1 |  |  |
| D | 0,32254343 | 0,22479873 | 0,67577678 | 1 |  |
| E | 0,32157995 | 0,85071457 | 0,2594794 | 0,16387803 | 1 |
| F | 0,74225416 | 0,65156556 | 0,61040038 | 0,35313787 | 0,58416663 |
| G | 0,39883277 | 0,32260169 | 0,83858703 | 0,57585661 | 0,30906546 |
| H | 0,42630064 | 0,31590116 | 0,50508482 | 0,26002194 | 0,24193879 |
| I | 0,54435541 | 0,34435486 | 0,56868991 | 0,23429682 | 0,25256881 |
| J | 0,45631191 | 0,28682901 | 0,53126058 | 0,2013004 | 0,18936071 |
| K | 0,47632134 | 0,31397727 | 0,96809309 | 0,70574453 | 0,22202372 |

Table 1

(2) Multiple regression.

I get the coefficient, standard error of the variables with Function LINEST. For table 2, at first glance, we can find that is 0,9987 which is very close to 1, it means that we already have the better quality of the regression model. However, is there multicollinearity in my case? On the one hand, we have found the existence of two more variables are highly correlated, on the other hand, we also find that the coefficient of the b6, as shown in Table 3, b6 has strong connection with y, but still b6 is not significant, so I assume that there is multicollinearity in my data case. As to the coefficient of determination and the adjusted coefficient of determination, , while . Using can avoid being trapped into specifying an overly complex model.

|  |  |  |  |
| --- | --- | --- | --- |
| b10 | b9 | b8 | b7 |
| 0,898159635 | -0,934378369 | 0,51038672 | 0,023610115 |
| 0,094601084 | 0,012836599 | 0,14582313 | 0,005681506 |
| 0,998701121 | 32,19991163 | #N/A | #N/A |
| 37291,38763 | 485 | #N/A | #N/A |
| 386649901,3 | 502864,64 | #N/A | #N/A |

Table 2

|  |  |  |
| --- | --- | --- |
| tb6 | 0,6682869 | < Tcr |

Table 3

|  |  |
| --- | --- |
| RSS | 386649901,3 |
| ESS | 502864,64 |
| TSS | 387152765,9 |
| R^2 | 0,998701121 |
| R^2adj | 0,998671662 |

Table 4

(3) Model Comparison.

From the correlation map and other information, we need to make comparison and decide which one is the best model for our case. In my case, I have divided them into 18 types of model. However, we still find that the original model which contains 10 independent variables is the best one with the biggest , and the least .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | F | Se | R^2adj | R^2 |
| A | 144,9710434 | 778,395998 | 0,22531701 | 0,22688202 |
| C | 7372,772212 | 221,8415524 | 0,93707712 | 0,93720423 |
| D | 490,2114501 | 627,1868969 | 0,4970593 | 0,49807534 |
| F | 166,2912864 | 765,7260113 | 0,25033086 | 0,25184534 |
| G | 961,424415 | 515,7584537 | 0,65989302 | 0,66058011 |
| AC | 4274,654009 | 206,9195545 | 0,94525734 | 0,94547852 |
| AD | 322,9185493 | 583,0561493 | 0,56534609 | 0,56710227 |
| AG | 543,8536646 | 494,8977162 | 0,68684905 | 0,68811431 |
| DC | 4010,206034 | 213,2501551 | 0,94185645 | 0,94209137 |
| DG | 720,6268931 | 447,388128 | 0,74408736 | 0,74512135 |
| FC | 4669,001487 | 198,4459634 | 0,94964908 | 0,94985252 |
| FD | 328,0687726 | 580,4371165 | 0,56924217 | 0,5709826 |
| FG | 481,901692 | 515,5142947 | 0,66021496 | 0,66158783 |
| CDA | 3063,538011 | 199,9608034 | 0,94887744 | 0,94918728 |
| CDF | 3348,342288 | 191,6823564 | 0,9530228 | 0,95330751 |
| DGA | 720,6268931 | 447,388128 | 0,74356721 | 0,74512135 |
| DGF | 480,7382607 | 447,3926388 | 0,7440822 | 0,74563321 |
| A-J | 37291,38763 | 32,19991163 | 0,99867434 | 0,99870112 |

Table 5

(4) New point forecast.

By using Function MMULT, MINVERSE, TRANSPOSE, with = , here A is the row vector of new data set while adding 1 on the first place and the first column of X is also should be 1 column. Then we can calculate the value of and also the confidence interval of new point as shown in Table 6 and Table 7.

|  |  |
| --- | --- |
|  |  |
| 10% | 1,45709584 |
| 20% | 1,49041536 |

Table 6

|  |  |
| --- | --- |
| Interval | |
| Lower | Upper |
| 881,30137 | 887,027369 |
| 966,271297 | 972,128234 |

Table 7

(5) Several tests.

1) Durbin-Waston test.

Durbin-Waston statistic is . For the value of dL, dU, we can find the corresponding number in the Durbin-Waston table according to K and N. In my case, dL=1,817, dU=1,89. And we can notice d is located between [0, dL], it means that there is statistical evidence that the error terms are positively autocorrelated.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Durbin Waston |  |  |  |  |
|  | K=10, N=496 |  |  |  |
| dL | 1,817 |  |  |  |
| dU | 1,89 |  |  |  |
| 4-dU | 2,11 |  |  |  |
| 4-dL | 2,183 |  |  |  |
| d | 1,806893836 | d>dL | positively autocorrelated | |

Table 8

2) Jarque-Bera test.

With the usage of , , . And in my case, JB is larger than Chi-square distribution with degree of freedom is equal to 2 which represents that the residuals are not normally distributed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Jarque-Bera |  |  |  |  |
| S | 2,645829522 | CHIINV | 5,99146455 |  |
| K | 33,33009098 |  |  |  |
| JB | 19590,26553 | do not have the normal distribution | | |

Table 9

3) Goldfeld-Quandt test.

In this test, we need to divide the data sample into three parts, and we only take the first and last part into consideration using regression. And , , comparing and , if , we would say that there is heterogeneity in this variable. Back to my case, it means that the variable I has heterogeneity.

|  |  |
| --- | --- |
| Low |  |
| 4,71758175 | 274,378561 |
| 1,25202181 | 55,3783546 |
| 0,08012301 | 348,7253 |
| 14,1976048 | 163 |
| 1726561,27 | 19822321,5 |

|  |  |
| --- | --- |
| High |  |
| 1,59673156 | 0 |
| 0,18163507 | #N/A |
| 0,32029042 | 1470,79651 |
| 77,2795198 | 164 |
| 167174332 | 354771751 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fg | 17,8975883 |  |  | |  |
| Fcr | 1,29488632 | <Fg | heterogeneity |

Table 10

4) The value of mean approximation error is below 8%, so the accuracy of model is good.

|  |  |
| --- | --- |
|  | 3,00% |

Table 11

(6) Modification.

As for the original model, after observing the performance of several points, I delete the irrelevant ones, and I get a new model with close and less . Table 9 is the part of the output table.

|  |  |  |  |
| --- | --- | --- | --- |
| 1,29305688 | -0,9502127 | 0,91567042 | -0,0078454 |
| 0,0929861 | 0,01133709 | 0,14263933 | 0,00803153 |
| 0,99853497 | 27,9948722 | #N/A | #N/A |
| 32783,8999 | 481 | #N/A | #N/A |
| 256931642 | 376965,889 | #N/A | #N/A |

Table 12

B: Question:

1) I don’t understand why there are highly correlated variables in my case, but still the original mode is the best?

2) why does work? I get the large than 1.