

### Homework 3

Please consider the “dataset\_Facebook.csv” data attached. It includes 500 observations and 19 variables. More details about the data are provided in the attached paper. Mainly, it has the following set of outcome/dependent variables:

Table 1:

List of output features to be modeled

Feature	Description <sup>a</sup>
Lifetime post total reach	The number of people who saw a page post (unique users).
Lifetime post total impressions	Impressions are the number of times a post from a page is displayed, whether the post is clicked or not. People may see multiple impressions of the same post. For example, someone might see a Page update in News Feed once, and then a second time if a friend shares it.
Lifetime engaged users	The number of people who clicked anywhere in a post (unique users).
Lifetime post consumers	The number of people who clicked anywhere in a post.
Lifetime post consumptions	The number of clicks anywhere in a post.
Lifetime post impressions by people who have liked a page	Total number of impressions just from people who have liked a page.
Lifetime post reach by people who like a page	The number of people who saw a page post because they have liked that page (unique users).
Lifetime people who have liked a page and engaged with a post	The number of people who have liked a Page and clicked anywhere in a post (Unique users).
Comments	Number of comments on the publication.
Likes	Number of “Likes” on the publication.
Shares	Number of times the publication was shared.
Total interactions	The sum of “likes,” “comments,” and “shares” of the post.

<sup>a</sup> Descriptions extracted from:

- <http://www.agorapulse.com/blog/facebook-reach-metrics-ultimate-guide>
- <https://www.facebook.com/help/274400362581037>

And the following Explanatory/ predictor variables:

Table 2:

**TABLE 2**  
List of input features used for modeling

Feature	Description
Category	Manual content characterization: action (special offers and contests), product (direct advertisement, explicit brand content), and inspiration (non-explicit brand related content).
Page total likes	Number of people who have liked the company's page.
Type	Type of content (Link, Photo, Status, Video).
Post month	Month the post was published (January, February, March, ..., December).
Post hour	Hour the post was published (0, 1, 2, 3, 4, ..., 23).
Post weekday	Weekday the post was published (Sunday, Monday, ..., Saturday).
Paid	If the company paid to Facebook for advertising (yes, no).

Now, please do the following.

1. Please consider the “Likes” as the outcome variables and all the variables in Table 2 as the explanatory variables.
  - a. Fit a multiple regression model and explain the results including coefficient significance and R-squared values. Refit the model using “log(Page total likes)” instead of “Page total likes” and “log(Likes)” instead of “Likes”. What did you observe?
  - b. For the best model in question (a), are there any variables that are not significant and can we drop them from the model? Do you observe any significant changes or the results remain similar even after they are dropped from the model?
  - c. Please carefully check whether the model assumptions are met or not? Please suggest remedies if the assumptions are not met.
  - d. Are there any variables that have nonlinear association with the outcome variable? How do you deal with them?
2. Please repeat question 1 for the outcome variables “Comments” and “Shares”? Which model you find interesting and why?