Instagram: Predicting Post Engagement

Merritt Cozby Anna Giesler
William Jacobe Andrew Lisi George Rao

1 Introduction

Why this is interesting. Instagram is the most popular social media platform of our generation, and the most prevalent metric of success on the platform is quantity of likes and comments. This led us to wonder: How can users increase engagement?

Summary of findings. We found that the most prominent predictor of engagement was followers. The poster's verified status, number of posts, and whether they were a personal or non-personal account also seemed to play a role.

2 Methods

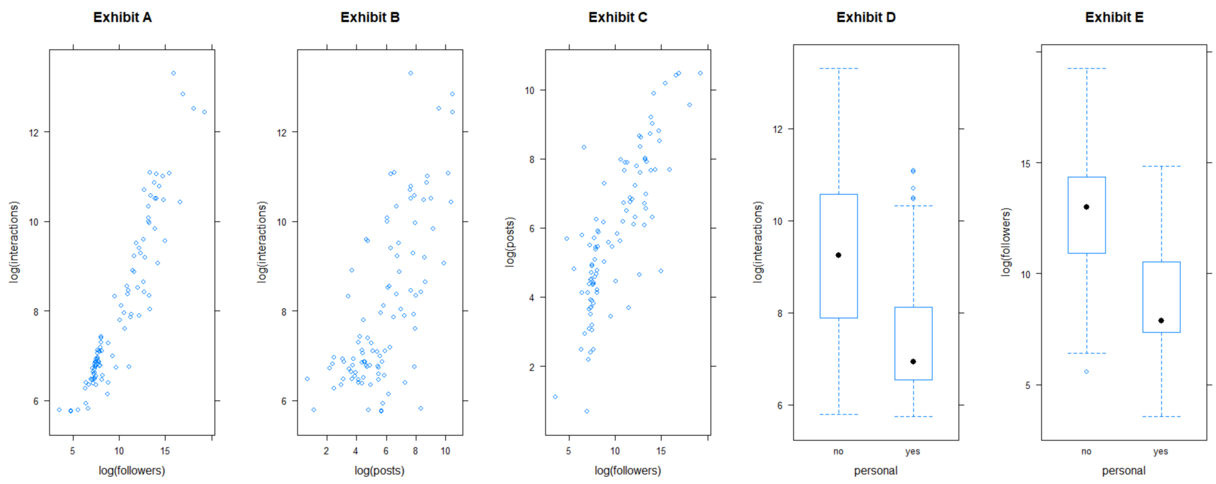
Collection method. We collected a sample of 100 posts, with each member collecting data on the first 20 posts in their feed, excluding those younger than a day or older than a week. We reasoned that, even though this method suffered from convenience sampling bias, there was no better way to collect a close-to-random sample.

Variables collected. In addition to engagement metrics, we collected a variety of other variables. The values of followers, posts, and following were taken at time of collection, rather than time of posting, but by limiting posts to those younger than a week, it's highly likely that these numbers were still similar and relevant to engagement.

Variables created ex-post. We created three notable variables: interactions (likes + comments + views), likesPerFollower, and peopled (true if people depicted > 0).

Data exploration. We found a few notable correlations involving interactions:

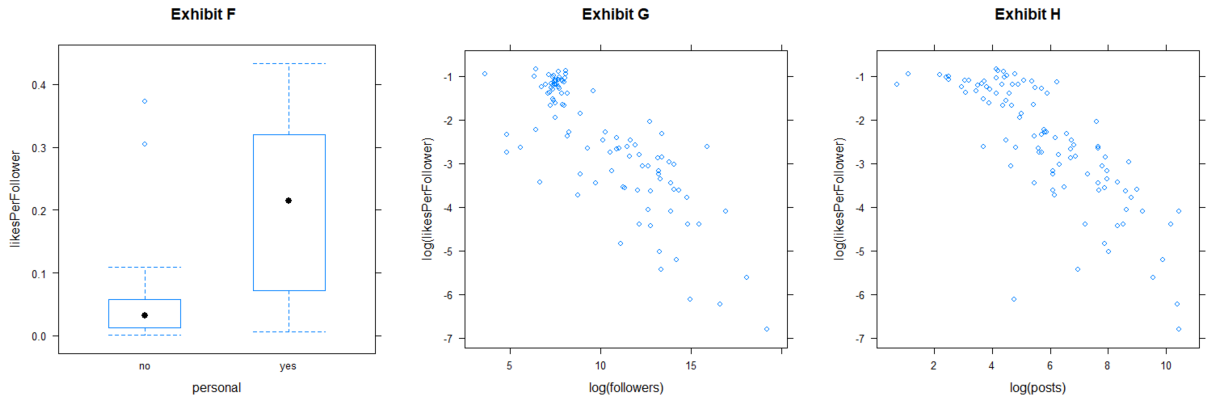
> Naturally, a strong correlation between interactions and followers. Out of all transformations, log–log seemed to result in the most linear relationship (Exhibit A). From this, we knew we would have to log-transform both terms in our model.
>
> An association between interactions and posts (Exhibit B); however, it was likely confounded by followers, which also correlated strongly with posts (Exhibit C).
>
> Similarly, fewer interactions if an account was personal (Exhibit D), which could be a result of confounding—after all, personal accounts also tended to have fewer followers, according to Exhibit E.



The question of whether these relationships were real or only due to confounding informed our decision to build a multiple regression model to adjust for confounders.

Additionally, we found associations involving likesPerFollower. Personal accounts had much higher likes per follower on average (Exhibit F). Likes per follower was negatively correlated with followers, after applying a log transformation to both variables (Exhibit G). Also, log(likesPerFollower) was negatively correlated with log(posts) (Exhibit H).

Based on this, we knew that we would have to use a log–log transformation when regressing likesPerFollower against either followers or posts.



Description of models. **First, regressed interactions and followers, taking log–log to ensure a linear fit:** `log(interactions) ~ log(followers)`

Second, we made a multiple regression model to determine which other variables explained interactions, adding each variable starting from the simple regression to see if it improved our model. Ultimately, we found that only “verified” added noticeable predictive power for interactions: `log(interactions) ~ log(followers) + verified`

Finally, we changed the response to likes per follower, which helped us find two more explanatory variables that seemed to improve the model: log(posts) and personal:

```
log(likesPerFollower) ~ log(followers) + log(posts) + personal + verified
```

3 Results

Predicting interactions. Follower count plays a big role in predicting interactions, with a simple log–log regression producing an r-squared of 88% (±4%). That implies that only 12% of the variation in interactions remains unexplained once followers are considered.

Correspondingly, adding additional variables to the model proved unfruitful for the most part—after coupling each variable with log(followers) and calculating a bootstrapped 95% confidence interval for its coefficient, only one, verified, seemed to have a value significantly different from zero, resulting in our multiple regression model:

```
log(interactions) ~ log(followers) + verified
```

```
            name      lower      upper level   estimate
  log.followers.  0.5231568  0.6198798  0.95  0.5694637
     verifiedyes -0.9632363 -0.1371156  0.95 -0.5785402
```

According to this output, each 1% increase in followers was associated with a 0.57% (±0.05%) increase in interactions, verified status unchanged, which was intuitive.

However, less intuitive was the coefficient for verifiedyes: this output implied that, followers unchanged, getting verified was linked with a 44% decrease in interactions, with bounds of 13% and 62%.[1] But why might being verified hurt engagement?

Well, at its core, the goal of interaction is to send a signal to the account, with benefits equal to the value of the signal multiplied by the probability that it's noticed. For branded or popular accounts, the quantities of both factors seem smaller—those accounts might not value the support as much, and are less likely to notice.

Most brands are verified, and since verified status is prominently displayed in the feed (in contrast to follower count, which is hidden), users may be likely to use the blue checkmark as a proxy for popularity in their mental calculus.

On the other hand, this may have just been a fluke. Testing at a 95% confidence level over so many variables (20+), it's likely that one or two came out looking important by chance alone, and this result's counterintuitive nature made us even less confident.

[1] Estimates/bounds of coefficients in log–level relationships were interpreted by subtracting 1 from e^coef.

Predicting likes per follower. Since followers alone seemed to explain so much of the variation in interactions, perhaps predicting likesPerFollower, a measure of follower-adjusted engagement, might give us more insight into the other predictors of engagement. After testing a few sets of explanatory variables, we kept log(followers) and verified while adding log(posts) and personal:

```
log(likesPerFollower) ~ log(followers) + log(posts) + personal + verified
```

```
             name       lower       upper level   estimate
   log.followers. -0.22034560 -0.02979712  0.95 -0.1252303
       log.posts. -0.36805988 -0.08381593  0.95 -0.2277234
      verifiedyes -1.21574340 -0.38200232  0.95 -0.7793145
      personalyes  0.07759307  0.84135607  0.95  0.4717313
        r.squared  0.69451645  0.84812703  0.95  0.7695191
```

Other explanatory variables constant, these relationships seemed to hold, on average:

For each 1% increase in followers, likesPerFollower decreased 0.125% (±0.1%). This is sensible because as accounts gain traction over time, their follower base might include lower proportions of core, truly loyal followers.

For each 1% increase in post count, likesPerFollower decreased 0.23% (±0.14%)—number of posts might be a proxy for post frequency, and the more often an account posts, the less notable each post might appear to a follower.

Having verified status was associated with a 54% decrease in likesPerFollower, with bounds of 32% and 70%. The explanation is similar to the one given previously. However, in this case, verified does not capture the lower perceived value of liking a post by a brand, since personal already captures brand effects.

Having a personal account was linked with a 60% increase in likesPerFollower (bounds of 8% and 132%)—personal accounts may have stronger individual relationships with their followers, making them more likely to interact.

Dropping otherwise-intuitive explanatory variables. To decrease complexity, we decided to drop a few variables from our models, even though they could have seemed like intuitive predictors of engagement to a normal Instagram user.

Initially, we hypothesized that age would help predict engagement since older posts might be seen by more users, but we failed to observe a relationship. This could be because posts tend to reach a maximum level of interactions within the first day, and we restricted our data to exactly that time frame.

We also guessed that tags or mentions might affect engagement—either positively by attracting more attention, or negatively by distracting followers to the accounts tagged or mentioned, away from the post itself. However, we observed no such relationship after accounting for followers. It's possible that the two effects we mentioned cancelled out, or that there were no effects at all.

Finally, we presumed that having people depicted (peopled) might raise engagement (possibly due to a preference for people over objects), but no relationship was found.

4 Discussion

Evaluating our hypotheses. Except for the relationship with followers, most of our hypotheses about engagement did not find good evidence. Moreover, we continue to wonder why changing the explanatory variable to likesPerFollower caused additional explanatory variables to become relevant.

Limitations of analysis. We found three main limitations of our analysis. First, we never qualified the subject of each post in our data beyond counting the number of people, so our model lacks the influence of categories like food, landscapes, or selfies—it's certainly possible that post subject has an influence on engagement not already captured by the other variables we collected.

Second, our data was taken from the feeds of our team members, rather than randomly sampled. Although it's unclear what random sampling would even mean in this context, we can say that, if we are to be rigorous, the implications of our analysis must be interpreted in the context of a special population of posts, rather than posts in general. This population might be described as "posts made by accounts followed by Cozby et al., weighted by popularity among the researchers."

Finally, we encountered problems with specific variables. For example, Instagram only displays the number of views on single-frame video posts, so much of the data on views was unavailable. Moreover, the "self" and "personal" variables were too similar in their construction to analyze simultaneously in a multiple regression model. Also, we only collected 100 observations, which could have made big models too unreliable.

Key takeaways. Our key takeaway from this project is that the best (and nearly only) way to increase engagement is to increase followers, and that, contrary to popular belief, getting verified might actually decrease engagement. Users focused on increasing likes per follower might also want to post less frequently, use a personal account, and try to maintain a loyal following.