[Fall 2021] Lab 2: Facebook User Engagement Analysis for Cosmetic Marketing

w203: Statistics for Data Science

Shanie Hsieh, Elias Saravia, Alejandro Pelcastre

December 5, 2021

Introduction

The Data and Research Design

The world is increasingly becoming more datafied every year. As a result, more and more companies are utilizing the power of data science to promote their business and create an online presence. With digitalization, our society slowly turns towards social media to form relationships, have greater reach, and even provide lifestyles to billions around the world. Facebook, being one of the largest platforms on the internet, is an attractive domain to garner space in the digital era. Several companies began creating a Facebook page to promote their products with advertisements on the site. We will investigate how a renowned cosmetics brand's advertising managed to perform by evaluating how users interacted with the company's post. To do this, we will investigate a dataset containing the time of day, the hour of the day, whether the post was paid for or not, and other features to analyze these factors that best boost interactions. Ultimately, we are focusing on the hour variable to answer our research questions:

> *Does the time of day of a cosmetics Facebook post have an influence on user engagements?*

We will be creating a regression analysis by using the large sample model to create our model with only the hour variable on the total interactions variable to analyze impact. We will address the linear model assumptions and convince the reader that our dataset satisfies the requirements. We will begin by preprocessing our data to make sure there are no complications with null values and erroneous data as well as one-hot encode any necessary categorical variables. Next, we plan to create several different models that contain other variables such as paid (if the post was paid to be promoted) and weekday (what day of the week the post was posted) in order to determine the significance of our main variable, hour.

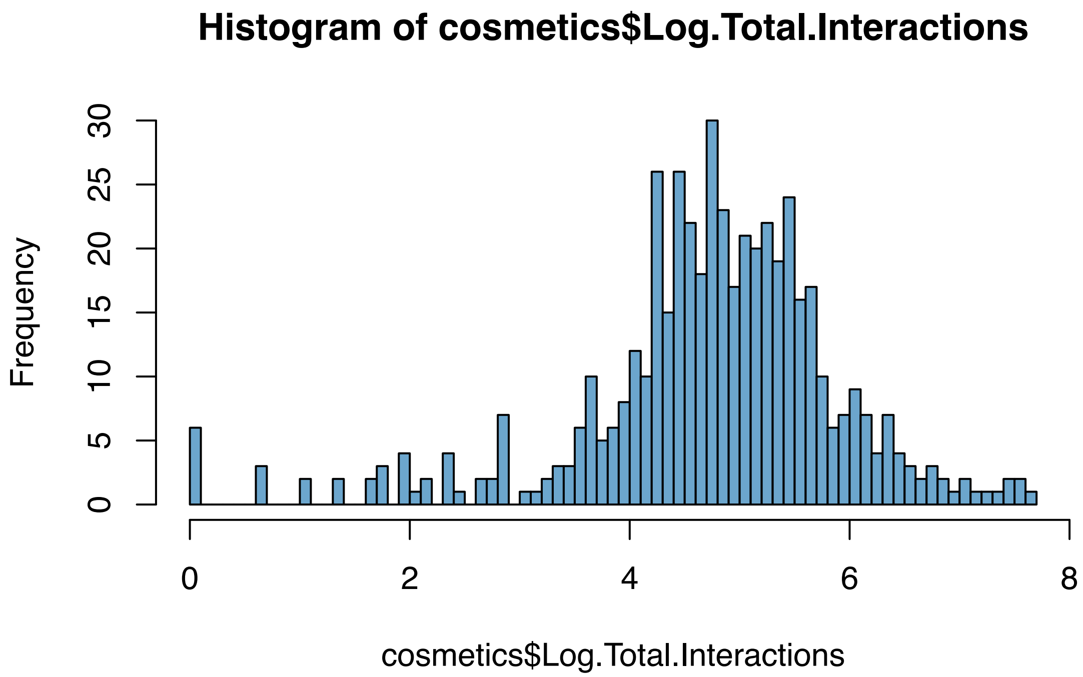
By answering this question, we could better give insight to this cosmetics brand to better reach their audience. Moreover, this research could extend to other companies. Though we are focusing solely on the impact of one variable, our research could be replicated with several different variables to predict their impact.

The Data

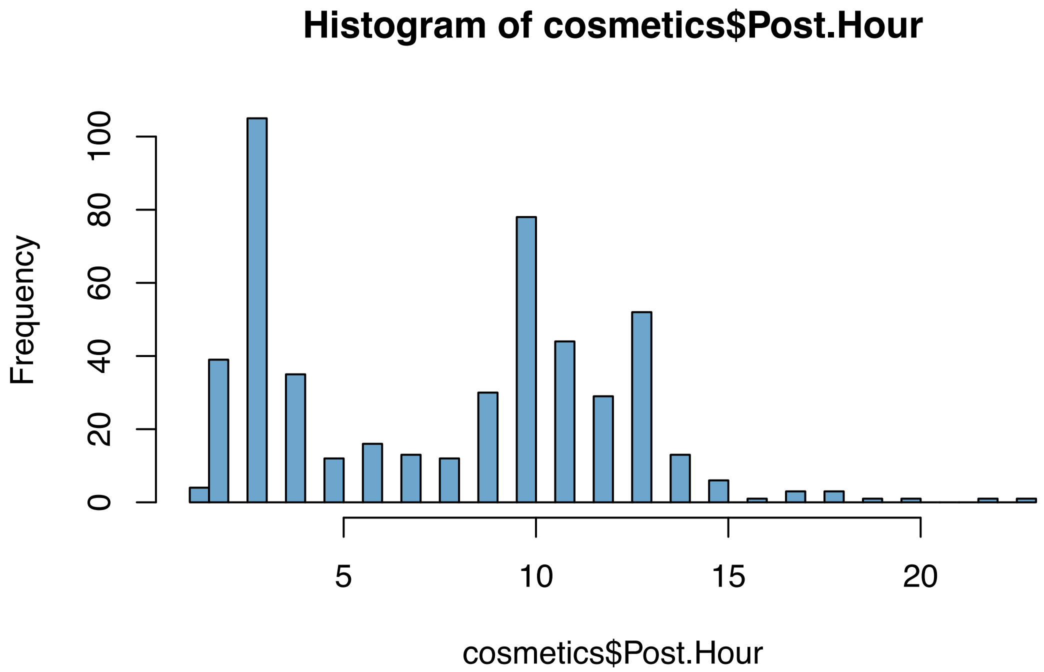
The dataset we will be exploring to answer our research question comes from the UCI Machine Learning Repository by a research paper: Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach (Sérgio Moro a,b,∗, Paulo Rita a, Bernardo Vala). The dataset explores posts that were published on Facebook by a cosmetics brand between January 1st, 2014 to December 31st, 2014. Originally the data had 790 observations, but due to some of the rows containing deanonymizing data the dataset had to be reduced to n=500 observations with 19 variables. Essentially, 7 variables are features prior to the post and the other 12 variables evaluate the post. The outcome variable we are using is Total Interactions and will be utilized as our success metric for our model. The list of all variables within the dataset are:

- Page total likes
- **Post information:** Type, Category, Paid
- **Post time:** Post Month, Post Weekday, Post Hour
- **Lifetime post effects** Lifetime Post Total Reach, Lifetime Post Total Impressions, Lifetime Engaged Users, Lifetime Post Consumers, Lifetime Post Consumptions, Lifetime Post Impressions by people who have liked your Page, Lifetime Post reach by people who like your Page, Lifetime People who have liked your Page and engaged with your post
- **Post engagement:** Comment, Like, Share, Total Interactions

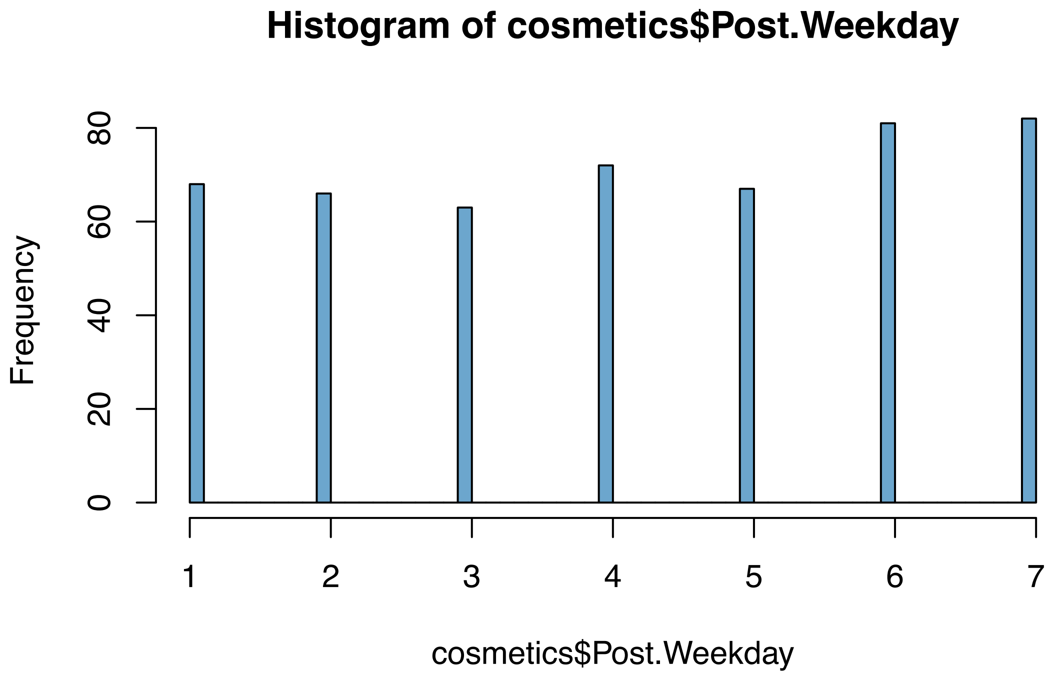
An initial data exploration of the data showed all of our features are integer types and there are no missing values. Limiting our dataset to show only the variables we cared about (total_interactions, post.hour, time_of_day post.weekday, paid, and type), we decided to take a look at the relative distribution of our data and note any outliers. Our outcome variable,total_interactions, had a right-skewed distribution with a high frequency of instances at lower interactions with the occasional post with high interactions. We also found one observation with an extremely high total_interactions number (a hit!) that could skew our model. Since there was only one outlier we decided to remove the datapoint leaving us with n = 499 observations. Furthermore, we were interested in looking at the time of day, therefore we created the following bins to represent "morning post," "afternoon post," and "evening post": hours 0 - 8, hours 9 - 16, and hours 17 - 23. We also performed a logarithmic transformation on the total_interactions variable (due to large values) which produced an amazingly normal distribution:



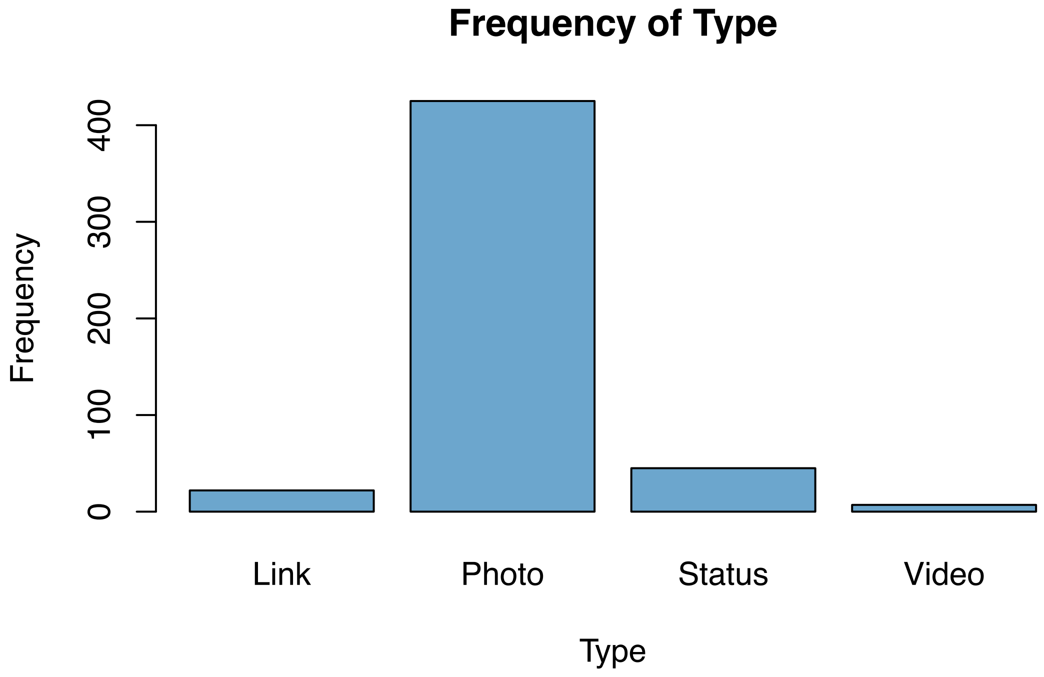
The focus of our model is on the Post.Hour variable. Taking a look at the distribution of this metric variable, we see the range from 0-23 with more posts early in the day. We decided we must apply a transformation that can make our variable more useful for analysis by making it categorical. We will explore this in our model building process.



We also decided to take a look at the distributions of various other variables that we are including in our supplementary models. Looking at Post.Weekday, which is also a metric variable, we see a relatively even distribution across all 7 days. We will apply a similar transformation as Post.Hour to Post.Weekday to make the variable categorical.



Our next variable looks at Type which is also a metric variable with strings as inputs. There are 4 types of posts, link, photo, status, and video. We decided to look at the barplot to visualize the frequency of each type.



Lastly, we decide to take a look at Paid, which is a binary variable with 0 being unpaid posts and 1 being paid posts. We found that paid posts are about half the amount of unpaid posts.

The Model Building Process

For our investigations, we used the Ordinary Least Squares (OLS) model to interpret the relationship between controllable variables (what time the post was made, and whether it was a paid/promoted) and total user interactions on the company's posts. Since our dataset contains n = 499 observations, we decided to use the Large-Sample model given that we exceed the general $n > 100$ requirement. The Large-Sample model is appealing over the Classical Linear Model (CLM) because there are less assumptions that need to be met than in the CLM in order for our results to have guarantees. We will briefly go over these assumptions and whether or not they have been satisfied:

1. First, the Large-Sample Model requires that our data is independent and identically distributed (I.I.D.). We would expect there to be some correlation between posts - for example, given a post about a promotion it is possible that another post could remind users that the promotion is ending. We can expect that after time passes, posts at one time will be relatively independent of those posted at much different times. This assumption leads us to conclude that we can treat the data as independent, though we should keep in mind that it is only an approximation. Another crude assumption is believing they are all identically distributed. Prior business insights and experience may have inclined the cosmetics company to post at certain times of the day. Another consideration is that working individuals make the posts so it could be that company schedule influences when in time the posts are generated. After exploring the data and seeing a uniform distribution in terms of posts in a given weekday, we can sensibly support the I.I.D. assumption.

2. The second and last requirement is that a unique Best Linear Predictor (BLP) for our model exists. In order to access if the BLP exist, we have to check that the covariance of our features (cov[Xi, Xj]) and the covariance of our dependant and independent variables (cov[Xi, Y]) are finite (no heavy tails), and that there is no perfect collinearity (i.e. E[XTX] is invertible). This led us to print the following covariance matrix of our features:

```
# covariance matrix here (make sure there are no large values)
cov(cosmetics[, (colnames(cosmetics) %in% c("Log.Total.Interactions", "Post.Hour", "Post.Weekday", "Pai
```

```
##                         Post.Weekday  Post.Hour          Paid
## Post.Weekday            4.1376289868  0.4006852 -0.0001212092
## Post.Hour               0.4006852359 19.1152901 -0.1323361858
## Paid                   -0.0001212092 -0.1323362  0.2007224067
## Log.Total.Interactions -0.3391253312 -0.5510144  0.0595764297
##                        Log.Total.Interactions
## Post.Weekday                      -0.33912533
## Post.Hour                         -0.55101436
## Paid                               0.05957643
## Log.Total.Interactions             1.48921957
```

We see that all pairs of features [Xi , Xj] have non-zero covariance indicating there is no perfect collinearity. We also see our covariance matrix shows reasonably finite numbers where we can conclude that the covariance of our dependent and independent variables are finite. By proving both parts of the statement, we can claim that a unique BLP exists for our data.

Model Building Process: The Model

After satisfying the assumptions laid out for our large-sample model, we moved forward with constructing a model based on hour, time of day, and a variety of other covariances.

When conducting our exploratory data analysis we discovered that the distribution of our total interactions variable was highly right-skewed. We wanted to focus on whether there was a relationship between total interactions and the hour of day, so we decided to rescale the output variable by taking a log transformation and creating a new feature called Log.Total.Interactions

Model 0: Since we want to use hour as our dependent variable in a Large Sample Model we needed to first take into account how the model accesses relative distance between each hour. Without any sort of transformations, our model would not replicate the practical nature of hours in the day. To fix this, we decided to use an indicator function and one-hot encode the hours so that the model emphasizes which hours of the day lead to higher interactions. Therefore, we first constructed a model using an indicator value for Post.Hour and the log transform of the total interactions as our outcome variable. This model had some significance and had an adjusted r-squared value of 0.0008675 and F statistic of 5.358.

$$log(TotalInteractions) = \beta_o I(Hour)$$

Model 1: Next, we constructed a new model taking into account the time of the day as an indicator since there were too many variables to account for if we utilized hour as an indicator. We took this as an opportunity to bin hours based on the time of day into one of three categories: morning, afternoon, and evening. The goal was to see whether a specific category signaled higher user interactions. We incorporated this along with the log transform of the total interactions as our outcome variable. This model had no significance and had an adjusted r-squared value of 1.322e-05 and F statistic of 1.006. Indicators for Afternoon and Evening were dropped from the model.

$$log(TotalInteractions) = \beta_0 I(TimeofDay)$$

Lastly, we created 3 other models that included covariates / other variables in our data. This included paid, post type, and weekday. The following models are outlined below along with their statistical information:

Model 2: This model included an indicator for time of day and indicator for post weekday. The model had a significant p-value for post weekday and had an increase in our adjusted r-squared (0.0193). The F-statistic was 5.802.

$$log(TotalInteractions) = \beta_0(TimeofDay) + \beta_1(PostWeekday)$$

Model 3: This model included an indicator for time of day, indicator for post weekday, and whether or not the post was paid. The model had a significant p-value for post weekday and paid and had an increase in our adjusted r-squared (0.0276). The F-statistic was 5.608.

$$log(TotalInteractions) = \beta_0(TimeofDay) + \beta_1(PostWeekday) + \beta_2(Paid)$$

Model 4: This model included an indicator for time of day, indicator for post weekday, whether or not the post was paid, and an indicator for the type of post (i.e. status, photo, and video). The model had a significant p-value for post weekday, paid, and post types and had an increase in our adjusted r-squared (0.05529). The F-statistic was 5.75. Indicator for link type was dropped from the model.

$$log(TotalInteractions) = \beta_0(TimeofDay) + \beta_1(PostWeekday) + \beta_2(Paid) + \beta_3(Type)$$

Results

We start with the results of Model 0. We see that our adjusted R-squared value is extremely low (0.0008675) signaling that the model either is insufficient or that the hour of the day has no impact on the number of interactions gained on a post. This conclusion is further supported by Model 1 with now simplifying the problem from a specific hour to more generally one of three times of the day - morning, afternoon, and evening. The adjusted R-squared value for this model is even lower: 1.322e-05 and with an F statistic of 1.006 meaning that the model was not improved. Moving on to Model 2 where we expand our search to include the weekday that the post was made, the results show a minor improvement of the model: adjusted R-squared of 0.0193 and F-statistic of 5.802. These values, however, are still too small to explain any relationship between our dependent and independent variables. As for Model 3, the adjusted R-squared value and F-statistic are 0.0276 and 5.608 respectively. We would expect an increase in the adjusted R-squared value since Facebook has claimed that paid promotions increase user engagement. The value is still small, however, again reinforcing that the time relationship between time of post and total engagement is not significant. Finally, our Model 4 includes the type of post (photo, status, and video) and again increases the adjusted r-squared value to 0.05529 with an F-statistic of 5.75. While the adjusted R-value is still low, we see that the improvement may signal a greater relationship between the type of post and interactions as well as if the post was paid for than our original independent variable of hour.

```
stargazer(model0, model1, model2, model3, model4, title = "Results", align=TRUE, type="text", font.size=
```

```
##
## Results
## =====================================================================================================
##                                                                    Dependent variable:
##                        -------------------------------------------------------------------------------
##                                                                    Log.Total.Interactions
##                                (1)                  (2)                  (3)                    (4)
## ----------------------------------------------------------------------------------------------------
## I(Post.Hour)                -0.029**
##                               (0.012)
##
## I(Time_of_Day)Morning                               0.109                0.089                  0.076
##                                                    (0.109)              (0.108)                (0.108)
##
## I(Post.Weekday)                                                        -0.087***              -0.087**
##                                                                         (0.027)                (0.027)
##
## I(Type)Photo
##
##
## I(Type)Status
##
##
## I(Type)Video
##
##
## Paid                                                                                           0.271**
##                                                                                                (0.120)
##
## Constant                    4.949***             4.687***               5.057***               4.988**
##                               (0.112)              (0.075)              (0.136)                (0.139)
##
## ----------------------------------------------------------------------------------------------------
## Observations                   499                  489                  489                    488
## R2                            0.011                0.002                0.023                  0.034
## Adjusted R2                   0.009               0.00001               0.019                  0.028
## Residual Std. Error     1.214 (df = 497)     1.200 (df = 487)      1.189 (df = 486)       1.185 (df =
## F Statistic           5.358** (df = 1; 497) 1.006 (df = 1; 487) 5.802*** (df = 2; 486) 5.608*** (df =
## ====================================================================================================
## Note:
```

We wanted to see if our longer models performed better than our shorter models. To do this, we used the var.test(model1, model4) function to give us the following results: F = 1.0563, p-value = 0.5469, and the ratio of variances 1.056344 with the alternative hypothesis being that the true ratio of variances is not equal to 1.

```r
# F-Test here
var.test(model1, model4)
```

```
##
##  F test to compare two variances
##
## data:  model1 and model4
## F = 1.0563, num df = 487, denom df = 481, p-value = 0.5469
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.883708 1.262578
## sample estimates:
```

```
## ratio of variances
##           1.056344
```

Overall we did not see any significant relationship between the hour of the day a post was made and the amount of user interaction it gained. We did see some increase in the model's performance once we added other variables such as if the post was paid, what type of post, and weekday. This could be used in future research to guide institutions or any general organization to look more into the type of post and weekday to optimize user interactions. Based on our findings, we would expect the cosmetics company to freely post any time of the day and refocus their resources on investigating other factors to promote engagement.

Model Limitations

Statistical Limitations

There were a couple of statistical limitations that we faced that may have affected our models. First, in our data exploration of our main independent variable, Post.Hour, we found that there were very sparse data points between 17 (5pm) to 0 (12am) which could have greatly affected the predictions of our model. Had there been more equal data producing and collecting across the hours, we could more accurately determine whether hour is actually important to higher interactions. Another limitation we faced was that the data was solely collected from 2014. Perhaps there was something that socially impacted people that prevented them from accessing or viewing these posts. We would be missing significant data points if there was an impediment in that year. We also identified that many of the variables, though integers, were metric variables. Applying statistical models onto metric variables without transformation creates results that may be hard to interpret and analyze. We applied transformations that we hoped could overcome this limitation, however, these transformations may not have been enough to accurately determine our predictions in our models.

Structural Limitations

Omitted Variables and bias: At the time of each post, the total number of followers that our cosmetics page has could be different. Having more followers would mean more users would get the posts in their timeline, which increases the opportunities for people to interact with the post. This means that the number of followers in an unaccounted variable (an omitted variable) can help explain the total user interactions.

Given the omitted variable bias equation:

> *Omitted variable bias = (How much do omitted variables affect the outcome?) * (How related are measured and omitted variables?)*

We suspect that given more followers, more people get to see the post and thus cannot lower the chance of interactions - it can only increase. So the number of followers should reasonably affect the omitted bias (away from zero). However, we should not expect that the time of day a company makes a post drastically changes the number of followers (you would lose, gain, tons of followers every hour). With this in mind, we would expect the second variable on the right-hand side of the equation to be tiny. Therefore, given a small number multiplied by some other number we expect the overall omitted variable bias to go towards zero.

Conclusion

Originally, we hypothesized that the time of day of a post made on Facebook had a causal relationship with the total number of engagements. However, through our analysis we concluded that there is no significance in the time of day the post is made. This was based on information statistics on low adjusted r-squared values, low F-statistics values, and low values for F-test to compare two model variances. However, other variables in our model showed higher significance (ranked from highest to lowest significance): post status type, post weekday, post video type, post photo, and paid. Therefore, as a cosmetics brand, the time of day the post is made should not have an impact on performance of engagements. Instead there should be a higher focus on creating more visual posts (i.e. videos and photos).