

**MKT 568 Final Project – Predicting Engagements on Instagram**

By: Aditya Nene, Maxime Anderson, and Preshit Gujar

Date: December 11, 2024

Professor: Prof. Farnoush Reshadi

TA: Sanjeeth Nagappa Chakrasali

**Executive Abstract**

One of the most important success factors for social media influencers is how much engagements their posts get, as this can indicate how interested other social media users are in them, and finding indicators that lead to engagement is a problem that many influencers on social media face. We were given a historical dataset of posts on Instagram, their interaction data, and various other variables about the posts such as the caption, date, post type, users in photo, link, count of hashtags, count of mentions, followers, and following. We cleaned the data, removing all null values and removing the USERS IN PHOTO column, as many entries were represented by a “- “, and could not be interpreted accurately by us. After this, we also conducted data transformation, creating new variables from our categorical variables that could not be used for modeling. Our models were ultimately not great for accurately predicting the amount of engagement that a post would receive, as our best model had an adjusted r^2 value of 0.382. But, we gained insight about what the variables were that lead to engagements on posts which were follower count, following count, if the post was in English, and timing of the post. Followers had the greatest impact on the models when it came to driving likes on posts which makes sense – the more people who follow an account will see their posts and be more prone to liking them. Lastly, we found that some aspects of posts such as the caption sentiment and the type of post (photo or video) did not have a significant impact on engagement and should therefore not be a focus of influencers. Overall, this report talks about the various analysis tasks that we conducted, and what our recommendations to influencers are based on this analysis.

**Introduction**

The problem that we are looking at in this project is to understand what types of posts will lead to more engagements for influencers. The goal of this project is to make inferences for influencers beyond their own posts about what types and kinds of posts will drive the most engagements. This is important for influencers, as engagements drive their revenue streams. The more engagements that influencers receive, the more they will be able to receive through paid promotions and other marketing deals with companies. Because of this, this analysis will be useful to both influencers and companies looking to work with influencers. Influencers will be able to use this analysis to see what types and kinds of content that they should post to generate the most engagements. Companies that look to collaborate with influencers can also use this analysis, as it will be able to provide them with insight into many different types of influencers for which type of posts will get the most exposure and engagements. Our plan for solving this problem and providing the necessary insights is to run four different analyses: Linear Regression for Likes and Comments, Linear Regression with feature engineered variables for Likes and Comments, Linear Regression with day of the week as a binary for if it is a weekend or not, and separate linear regressions for micro and macro influencers on Instagram. After we have ran all of these different models, we will look at the performance of the model as well as the significant of the independent variables in each of the models to determine what the best recommendations are for influencers looking to increase their likes and comments on their postings.

**Methodology**

The CRISP(Cross-Industry Standard Process) data analysis methodology was used to prepare the data for the multiple analysis that we would conduct on it. This methodology consists of 6 phases which are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The process begins with understanding the business context, the business objectives, and transforming the business problem into a data analysis solution. The Data Understanding Phase consisted of exploring the data to find anything unusual or other patterns in the data. Data Preparation involves cleaning the data and creating the necessary variables for analysis. The Modeling phase consisted of running multiple models to find the best way to predict the amount of likes and the number of comments that a posting would receive. The Evaluation phase assesses the performance of the models that we created and leads into the deployment phase in which we decide if the models that we created would be able to provide insight in a real-world scenario.

**Data**

The data that will be used in our analysis was scraped from Instagram by Corentin Dugué. Here is a table of the variables that were used in our analysis and their type:

|  |  |  |
| --- | --- | --- |
| Variable Name | Definition | Type |
| LIKES | the number of likes the post received at the time of data collection | Continuous |
| COMMENTS | the number of comments the post received at the time of data collection | Continuous |
| TEXT | the text used as the post caption | String |
| DATE | the date the post was created | DATE - Categorical |
| TYPE | what type of post it was (1 = PHOTO, 2 = VIDEO) | Categorical |
| USERS IN PHOTO | the number of users in the post | Continuous |
| LINK | the Link to the Instagram post | N/A |
| FOLLOWERS | The number of followers that an account has | Continuous |
| FOLLOWING | The number of accounts that the poster is following | Continuous |
| number\_of\_tags | The number of hashtags in a posting | Continuous |
| number\_of\_mentions | The number of accounts mentioned in a posting | Continuous |

There is a total of 19682 records in the data. For the numerical data, the statistical takeaways before data cleaning can be seen in the table below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Min** | **25%** | **50%** | **75%** | **Max** | **Std** |
| FOLLOWERS | 62564.13 | 17993 | 23299 | 36699 | 62791 | 1134619 | 104234.9 |
| FOLLOWING | 1489.77 | 0 | 174 | 506 | 1367 | 7586 | 2252.68 |
| LIKES | 2497.77 | 0 | 420 | 1073 | 2683 | 158338 | 5574.99 |
| COMMENTS | 39.83 | 0 | 1 | 5 | 17 | 26011 | 447.97 |
| DATE | 2017-04-21 | 2016-03-27 | 2017-04-17 | 2017-04-22 | 2017-04-27 | 2017-05-02 | NaN |
| USERS IN PHOTO | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| number\_of\_tags | 6.74 | 0 | 0 | 3 | 10 | 41 | 8.78 |
| number\_of\_mentions | 0.72 | 0 | 0 | 0 | 1 | 34 | 1.70 |

As you can see, there are problems with the USERS IN PHOTO column, that will be addressed during the data cleaning step.

Here are all our categorical data columns, and the unique values that can be seen in each of the columns:

|  |  |
| --- | --- |
| **Variable** | **Unique Values** |
| TYPE (1 PHOTO, 2 VIDEO) | 1,2 |

There were 6 null values in the Text column, 5819 null values in the list\_of\_tags column, and 12935 null values in the number\_of\_mentions column. All of these null values are normal as null in the text column represents an empty caption, and null in the list variables represents a post that had 0 tags or 0 mentions which is normal.

In this analysis, the independent variables that we look at are likes and comments, as we are interested in what leads to more engagements for social media influencers. The dependent variables are the rest of the variables which are FOLLOWERS, FOLLOWING, TEXT, DATE, TYPE(1 PHOTO,2 VIDEO), USERS IN PHOTO, LINK, list\_of\_tags, number\_of\_tags, list\_of\_mentions, and number\_of\_mentions, as well as the additional variables crated during the data cleaning and data transformation steps.

**Data Cleaning**

The main challenge with cleaning this data had to do with the USERS IN PHOTO variable, as many of the records were designated with a “-“. We ultimately decided t drop this column for two main reasons, that it may interfere with the analysis regarding the number\_of\_mentions column, and that over half of the rows had a “-“ as a value, which we did not know how to interpret properly.

To clean the data, we first read the data into python and inspected it to see if we noticed anything unusual. During this, we noticed that the USERS IN PHOTO column had “-“ as a value in many pf the columns, and decided to drop it as we did not know how to properly address this. We did not notice any other unusual data entries from the statistics, so we moved forward with cleaning. After dropping this column, we then dropped all of the columns that did not have a caption, as we felt this may affect the analysis. We then went to detect outliers and remove any rows that were outliers. We set our threshold for detecting outliers at a z score of 3, meaning any continuous variables with a z score higher than -3 or 3 were removed. We decided to remove outliers, as we did not want them to make our predictions from our analysis worse. After both steps, the data frame had 17032 rows. Some additional steps that we took on our own to clean the data were changing the TYPE column to be 0 = Photo and 1 = Video. This column was also renamed to be simpler. We also changed the type of the Text column to be string. Next duplicate rows were checked for and removed. A correlation table for the remaining variables can be seen here:

A diagram of a number of followers

Description automatically generated

**Data Transformation**

For each of the questions, new variables were created for our analysis to be in line with the question being asked. For questions 1, it was necessary to transform the Datetime column into times of the day which were morning, afternoon, evening, and night, as well as days of the week, and the month of posting. All these variables that were created were categorical variables, and therefore had to be dummy coded before we ran our analysis. For month, MONTH\_April was dropped meaning that when all other month values are 0, the month of posting was April. For the day of the week, DAY\_Friday was dropped meaning that when all other columns are 0, the day of the week is Friday. For the time of day, time\_of\_day\_afternoon was dropped meaning that when all other time of day columns are 0, the time of day of posting was the afternoon. This makes the baseline for the analysis a Friday in April in which the post was made in the afternoon, which is represented by the coefficient in the analysis. All variables were made into integers, so that they could be interpreted by the model.

For question 2, it was necessary to create three new variables for the analysis which were

For question 3, the same variables were used that were used in question 1, but the DAY columns were altered. These columns were altered into a single binary that would designamte if the day was a weekend day or not. This meant that when DAY\_Saturday or DAY\_Sunday equals 1, then Is\_weekend also equals 1. This variable was then made to be an integer so that it could be used in the regression models

**Analysis Results**

**Question 1**

For question 1, two linear regression models were run to predict likes and comments that posts would receive based on many variables. Overall, these models did not perform well as the model looking to predict likes had an r^2 value of 0.378 and an adjusted r^2 value of 0.377, and the model looking to predict comments had a r^2 value of 0.027 and an adjusted r^2 value of 0.026. Both models would not be good for predictions, as the r^2 and adjusted r^2 values are too low for the models to give us any insights.

The variables that had a significant influence on the amount of likes a post received were Followers, Following, Type of post(Photo or Video), number of tags, MONTH\_May, DAY\_Wednesday, and text\_length, as all of these variables had a p-value less than 0.05. It is also important to note that the const was statistically significant, meaning that the combination of MONTH\_April, DAY\_Friday, and time\_of\_day\_Afternoon were all significant to the model. The variables that had a significant influence on the amount of comments a post received were Followers, Following, Type of post(Photo or Video), number of tags, text\_length, and MONTH\_May as all of these values had a p-value less than 0.05. It is also important to note that the const was statistically significant, meaning that the combination of MONTH\_April, DAY\_Friday, and time\_of\_day\_Afternoon were all significant to predicting the number of comments a post received.

For the model predicting likes, this table displays the coefficients of the variables that were statistically significant, and how they can be interpreted to predict the number of likes that a posting will receive.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Coefficient** | **Interpretation** |
| FOLLOWERS | 0.0328 | For every 100 followers someone has, they can expect 3.28 more likes per posting. |
| FOLLOWING | -0.168 | For every 100 people that someone follows, they can expect 16.8 less likes |
| TYPE OF POST | 1.44\*10^(-11) | One may see an extremely small decrease in likes for videos that are posted |
| Number\_of\_tags | -8.3745 | For every person tagged in a post, one can expect 8.375 less likes. |
| Text\_length | -0.2899 | For every 100 words in a caption of a posting, one can expect about 29 less likes. |
| MONTH\_May | -394.532 | If a post is in May, it can expect 394.53 less likes than if it was posted another time of year. |
| DAY\_Wednesday | 127.652 | If a posting was made on a Wednesday, it can expect to see 127.65 more likes. |
| Intercept | 691.560 | The baseline posting(A Friday afternoon in April) can expect to start at 691.56 likes. |

When looking at the standardized model for predicting likes, it can be seen that The number of followers has the largest impact on the model with a standardized beta coefficient value of 0.565. The next two variables with the largest impact are MONTH\_December with a standardized beta coefficient value of -0.321 and MONTH\_February with a standardized beta coefficient value of -0.300.

For the model predicting comments, this table displays the coefficients of the variables that were statistically significant, and how they can be interpreted to predict the number of comments that a posting will receive.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Coefficient** | **Interpretation** |
| FOLLOWERS | 0.0002 | For every 10000 followers someone has, they can expect 2 more comments per posting. |
| FOLLOWING | -0.0018 | For every 1000 people that someone follows, they can expect 1.8 less comments |
| TYPE OF POST | 2.93\*10^(-13) | One may see an extremely small decrease in comments for videos that are posted |
| Number\_of\_tags | -0.4253 | For every person tagged in a post, one can expect 0.4253 less comments. |
| Text\_length | 0.026 | For every 100 words in a caption, one can expect to see 2.6 more comments |
| MONTH\_May | -4.5266 | A posting made in May can expect to see -4.5266 less comments |
| Intercept | 12.40 | The baseline posting(A Friday afternoon in April) can expect to start at 12.409 comments. |

When looking at the standardized model for predicting comments, it can be seen that MONTH\_January had the largest impact on the model with a standardized beta coefficient value of -0.1784. The next two variables with the largest impact are MONTH\_March with a standardized beta coefficient value of -0.149 and FOLLOWERS with a standardized beta coefficient value of 0.1213.

**Question 2**

For question 2, we created three new variables which were is\_english, sentiments\_score, and time\_since\_posting. We belived that each of these variables would influence the number of engagements that posts were receiving in various ways. We believed that a variable that determines if the caption for the post was in English or not could help the model in a positive way, as posts in English may be able to be seen and understood by a larger audience. We believed that a sentiment\_score could influence the number of engagements that a post gets, as people may be more likely to engage with posts that are positive. We believed that the time\_since\_posting could help determine how many engagements a post has gotten, as we believed the longer a post has been up, the more opportunity it has had to be engaged with. We added these variables to the same linear regression analyses that we had run in question 1, to predict how many likes and comments a posting will receive.

Adding these variables to the model did improve the performance of the model, on a very minor level. After checking for multicollinearity and finding none, these variables were added as additional independent variables. For the model predicting the amount of likes that a posting would have, the adjusted r^2 value was raised to 0.382 from 0.377 in the previous model. Fo the model predicting the number of comments that a posting had, the adjusted r^2 value was raised to 0.031 from 0.026 in the previous model.

For the model predicting likes, the variables that had a significant influence on the amount of likes a posting would receive were FOLLOWERS, FOLLOWING, is\_english, text\_length, MONTH\_MAY, DAY\_Wednesday, and all of the time\_of\_day categorical variables. These variables were deemed to be significant as they all had p-values less than 0.050. For this model this table displays the coefficients of the variables that were statistically significant, and how they can be interpreted to predict the number of likes that a posting will receive.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Coefficient** | **Interpretation** |
| FOLLOWERS | 0.0328 | For every 100 followers someone has, they can expect 3.28 more likes per posting. |
| FOLLOWING | -0.160 | For every 100 people that someone follows, they can expect 16 less likes |
| is\_english | -300.183 | If the caption of a post is in English, 300.183 less likes can be expected |
| Text\_length | -0.3128 | For every 100 words in a caption of a posting, one can expect about 31.28 less likes. |
| MONTH\_May | -401.0642 | If a post is in May, it can expect 401.06 less likes than if it was posted another time of year. |
| DAY\_Monday | 125.306 | If a post is made on a Monday, it can expect to see 125.31 more likes |
| DAY\_Wednesday | 120.870 | If a posting was made on a Wednesday, it can expect to see 120.87 more likes. |
| time\_of\_day\_Evening | 93.802 | A post made in the evening compared to the afternoon can expect to see 93.80 more likes |
| Time\_of\_day\_Morning | -150.456 | A post made in the morning compared to the afternoon can expect to see 150.46 less likes |
| Time\_of\_day\_Night | 254.056 | A post made in the night compared to the afternoon can expect to see 254.06 more likes |

For the model predicting comments, the variables that had a significant influence on the amount of comments a posting would receive were FOLLOWERS, FOLLOWING, number\_of\_tags, is\_english, text\_length, MONTH\_May, and time\_of\_day\_Night. These variables were deemd to be significant to the predictive model, as they all had p-values less than 0.050. For this model this table displays the coefficients of the variables that were statistically significant, and how they can be interpreted to predict the number of likes that a posting will receive.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Coefficient** | **Interpretation** |
| FOLLOWERS | 0.0002 | For every 10000 followers someone has, they can expect 2 more comments per posting. |
| FOLLOWING | -0.0015 | For every 1000 people that someone follows, they can expect 1.5 less comments |
| is\_english | -8.6925 | If the caption of a post is in English, 8.69 less comments can be expected |
| Text\_length | 0.0251 | For every 100 words in a caption of a posting, one can expect about 2.5 more comments. |
| MONTH\_May | -4.6211 | If a post is in May, it can expect –4.62 less comments than if it was posted another time of year. |
| Time\_of\_day\_Night | 4.7834 | A post made in the night compared to the afternoon can expect to see 4.78 more comments. |

Overall, the variables that we decided to add to the model did improve the performance of the linear regression; however, the model is still not very accurate and should not yet be deployed to give suggestions to social media accounts on how to drive more engagements.

**Question 3**

For this question, another multiple linear regression was run, but this time, instead of every day of the week being a binary, the new binary was determined by if the data was a weekend day or not. For both the model looking to predict likes and the model looking to predict comments, the results were slightly better than the original models, as the model predicting likes had a r^2 value of 0.379 and an adjusted r^2 value of 0.379 and the model predicting comments had a r^2 value of 0.028 and an adjusted r^2 of 0.027. The variable “Is\_weekend” was not significant in prediction of both likes and comments, meaning that there is no evidence that posting on the weekend will lead to more engagements. For predicting likes, the variables that had a statistically significant impact on the amount of likes a posting would receive were FOLLOWERS, FOLLOWING, TYPE, number\_of\_tags, text\_length, MONTH\_May, and all of the time\_of\_day binary variables as well as the intercept. For predicting comments, the variables that had a statistically significant impact on the amount of comments a posting would receive were FOLLOWERS, FOLLOWING, TYPE, number\_of\_tags, text\_length, and time\_of\_day\_Night. Overall, replacing the dummy coded day variables with the Is\_weekend variable did improve performance of both of the models, but only very minorly.

**Question 4**

For this question, 4 models were run in total. Two models were run to predict likes and comments for micro influencers, and two models were run to predict likes and comments for macro influencers.

These models did not perform well compared to the initial models that we had ran. The results of the variables effect on the models can be seen in the tables below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Coefficient –Likes-Micro Influencer** | **P-value–Likes-Micro Influencer** | **Coefficient –Likes-Macro Influencer** | **P-Value–Likes-Macro Influencer** |
| const | 697.607 | 0 | 1290.446 | 0 |
| FOLLOWING | -0.1138 | 0 | -0.338 | 0 |
| FOLLOWERS | 0.0257 | 0 | 0.031 | 0 |
| MONTH\_December | -625.464 | 0.524 | 2.586\*10^-11 | 0.443 |
| MONTH\_February | -259.723 | 0.574 | -1491.839 | 0.439 |
| MONTH\_January | -729.699 | 0.457 | 1437.076 | 0.597 |
| MONTH\_March | 201.431 | 0.745 | -1.872\*10^-12 | 0.720 |
| MONTH\_May | -296.511 | 0 | -502.075 | 0.005 |
| DAY\_Monday | 83.757 | 0.099 | 237.698 | 0.110 |
| DAY\_Saturday | 16.179 | 0.730 | 34.630 | 0.799 |
| DAY\_Sunday | 3.564 | 0.938 | 10.425 | 0.937 |
| DAY\_Thursday | 1.594 | 0.972 | 175.480 | 0.185 |
| DAY\_Tuesday | 35.642 | 0.475 | 85.127 | 0.559 |
| DAY\_Wednesday | 18.556 | 0.714 | 268.974 | 0.066 |
| Time\_of\_day\_Morning | -58.525 | 0.162 | -428.949 | 0.001 |
| Time\_of\_day\_Evening | 77.553 | 0.019 | 114.109 | 0.246 |
| Time\_of\_day\_Night | 116.785 | 0 | 321.175 | 0.001 |
| Number\_of\_tags | 19.548 | 0 | -49.733 | 0 |
| Number\_of\_mentions | -33.951 | 0.020 | -10.408 | 0.819 |
| TYPE(0=PHOTO) | 3.065\*10^-13 | 0.332 | 0 | nan |
| Text\_length | -0.4567 | 0 | 0.3171 | 0.349 |
| Is\_english | -252.3114 | 0 | -379.213 | 0.000 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Coefficient –Comments-Micro Influencer** | **P-value–Comments-Micro Influencer** | **Coefficient –Comments-Macro Influencer** | **P-Value–Comments-Macro Influencer** |
| const | 12.4115 | 0 | 18.8573 | 0 |
| FOLLOWING | -0.0012 | 0 | -0.0022 | 0 |
| FOLLOWERS | 0.0001 | 0.004 | 0.0001 | 0 |
| MONTH\_December | -3.0756 | 0.931 | 2.552\*10^-13 | 0.654 |
| MONTH\_February | -4.7855 | 0.774 | -5.8003 | 0.912 |
| MONTH\_January | -9.2271 | 0.794 | 4.0031 | 0.957 |
| MONTH\_March | -7.0537 | 0.753 | -4.117\*10^-15 | 0.977 |
| MONTH\_May | -3.6488 | 0.087 | -5.284 | 0.275 |
| DAY\_Monday | -0.8521 | 0.642 | 4.134 | 0.308 |
| DAY\_Saturday | -0.8031 | 0.635 | 1.6172 | 0.661 |
| DAY\_Sunday | -1.6231 | 0.328 | 2.8262 | 0.433 |
| DAY\_Thursday | -2.6184 | 0.115 | -0.8234 | 0.819 |
| DAY\_Tuesday | -2.0566 | 0.254 | 10.0443 | 0.011 |
| DAY\_Wednesday | -3.4550 | 0.059 | 5.6742 | 0.154 |
| Time\_of\_day\_Morning | -0.7750 | 0.608 | -2.8572 | 0.414 |
| Time\_of\_day\_Evening | -0.1490 | 0.900 | 2.5450 | 0.342 |
| Time\_of\_day\_Night | -0.0125 | 0.992 | 11.7412 | 0.000 |
| Number\_of\_tags | 0.1044 | 0.130 | -1.0669 | 0.000 |
| Number\_of\_mentions | 0.4983 | 0.346 | -0.5475 | 0.658 |
| TYPE(0=PHOTO) | 5.23\*10^-16 | 0.963 | 0 | nan |
| Text\_length | 0.0156 | 0 | 0.0674 | 0.000 |
| Is\_english | -5.8142 | 0 | -14.8801 | 0.000 |

When looking at the effect that each of the independent variables had on predicting the amount of likes in both models, the intercept, FOLLOWERS, FOLLOWING, MONTH\_May, and time\_of\_day\_Night, number\_of\_tags, and is\_english were seen to be significant in predicting likes for both macro and micro influencers. Time\_of\_day\_evening , text\_length, and number\_of\_mentions were significant in predicting likes for exclusively micro influencers while time\_of\_day\_Morning was significant in predicting likes exclusively for macro influencers.

When it came to the coeffects for both models looking to predict likes, all the coefficients were different from one another, meaning that the number of likes that the model would predict a posting to have would be different for a micro-influencer vs a macro-influencer. For example, the text\_length coefficient was –0.4567 for micro influencers and 0.3171 for macro influencers, meaning that a longer caption would lead to a lower predicted likes for micro influencers but a higher predicted likes for macro influencers.

When looking at the effect that each of the independent variables had on predicting the number of comments in both models, the intercept, FOLLOWERS, FOLLOWING, text\_length, and is\_english were seen to be significant in predicting likes for both macro and micro influencers. Day\_Tuesday, time\_of\_day\_Night, and number\_of\_tags were all significant in predicting likes exclusively for macro influencers.

When it came to the coeffects for both models looking to predict comments, all the coefficients were different from one another, meaning that the number of comments that the model would predict a posting to have would be different for a micro-influencer vs a macro-influencer. For example, the number\_of\_mentions coefficient was 0.4983 for micro-influencers and –0.5475 for macro-influencers. This means that more mentions would lead to more predicted comments for micro-influencers but less for macro-influencers.

Overall, based on this analysis, we cannot make different predictions about increasing engagements for micro-influencers and macro-influencers separately as when the data was separated, both models that we ran performed worse than the original model with the complete dataset. This means that the full dataset is needed to predict likes and comments most accurately.

**Conclusions and Recommendations**

This project consisted of 6 main phases: Data Cleaning, Data Transformation, and Questions 1-4. In the data cleaning phase, we checked all of the columns to make sure there were no null values, removed a column in which we did not know how to properly interpret, and made sure all outliers were removed prior to analysis. In the data transformation phase, all categorical variables were dummy coded so that they could be used in our linear regression analysis. In question 1, two models were ran (, both standardized and unstandardized) to predict the amount of likes and comments that influencers received on Instagram. In question 2, we added three variables which were is\_english, Text\_Sentiment, and time\_since\_posting, to ultimately improve the models that we had created in question 1. In question 3, we looked to see if a binary variable for if a day was a weekend or not would be able to help improve our original model, which it did not. In question 4, we looked to separate the data into two different data sets for micro and macro-influencers and created new models to see if engagements for these subgroups could be better predicted, which they could not be.

Throughout the analysis process, although the models were not as accurate as we would have liked, we did come away with some key takeaways for influencers looking to improve the number of engagements they are seeing on their posts. We believe that data that was not seen in this data table likely is also having an impact on the amount of engagement that posts are receiving. This data is likely surrounding the post itself and the content that is included in it in which we do not know from the tabular data frame. Some of the other key takeaways from this analysis being done was that Followers, Following, and some aspect of our timing metrics(Variables or the intercept) was seen to be significant in all our models. Also, the variable that we added that is a binary for if the post is in English or not was also significant in all of our models. Overall, when looking at the standardized model in question 1, we found that Followers had the greatest impact on increasing both comments and likes, while MONTH\_December had the greatest impact on decreasing likes and MONTH\_January had the largest impact on decreasing the number of comments. Based on our best model in question 2, the best type of post for receiving more likes would be a post from an account with many followers and not following many people, with no tags, a short caption, very little or no mentions, that is not in English, was posted recently, in the month of March, on a Wednesday, at Night. Some variables that did not have a large effect on this prediction were the type of post(Photo or Video) and the Text Sentiment. Based on our best model in question 2, the best type of post for receiving more comments would be a post from an account with many followers and not following many people, with no tags, a longer caption, has a lot of mentions, that is not in English, was posted recently, in the month of December, on a Tuesday, at Night. Some variables that did not have a large effect on this prediction were the type of post (Photo or Video) and the Text Sentiment.

Our recommendations for social media influencers looking to increase their engagements would be the following:

* Focus on increasing the number of followers, and limit the amount of people you are following to increase engagements on your posts
* Post during the night as this maximizes both predicted likes and comments
* Do not worry about whether you should post a video or photo, as it does not have a large impact on engagement
* Make your posts able to be translated, so that people of all languages will be able to engage with it
* If you are looking to increase likes, keep captions short, hashtags low, and mentions low. If you are looking to increase comments, use longer captions, more hashtags, and less mentions

**References**

Dugué, C. (2017, May 14). *Predicting the number of likes on Instagram*. Medium. <https://towardsdatascience.com/predict-the-number-of-likes-on-instagram-a7ec5c020203>

*Introduction — statsmodels*. (n.d.). Www.statsmodels.org. <https://www.statsmodels.org/stable/index.html>