DSC 423
Facebook Interaction Dataset Analysis

Team: Data Crew

Denvir Gama, Erik Pak, Jiten Mishra, Liang Tse Hsu, Martello Pollock

DePaul University
Data Analysis and Regression 423
Dr. Alvin Chin
March 11, 2023

Abstract

This study presents a research approach for performing the metrics of Lifetime post-consumer for a post published in a brand's Facebook page. The performance metrics extracted from a cosmetic company's page were modeled. The Mean Absolute Percentage Error (MAPE) for our final model is 6.000214%. The model was built on Lifetime post-consumer as our response variable. The dataset includes identification, content, categorization, and performance features that enable unique identification of each post, categorization of the post, and performance metrics such as post reach, impressions, and engagement. Other variables such as post type, category, post month, post weekday, and post hour are also included, providing a rich source of information for predicting social media performance metrics and evaluating their impact on brand building. The analysis of this model information may be used by managers to make decisions on whether to publish a post.

Introduction

Social media has become an integral part of businesses today. It's a valuable tool for marketing and brand building. To create an effective marketing strategy, understanding your customer base is essential, and social media metrics play a vital role in this. Metrics such as Impressions, Engagement, and Reach help businesses predict the impact of their social media content on brand building. This study aims to develop a predictive model that uses the characteristics of social media posts to determine their effects on Facebook. The study will focus on identifying the most significant factors that contribute to social media post success. These factors include post type, post length, timing, use of multimedia, and target audience demographics. By analyzing these factors, businesses can tailor their content to their target audience's interests and increase customer engagement, leading to more robust brand building and overall business success.

Utilizing a predictive model can help businesses make informed decisions about their social media content. It can help them optimize their social media strategies, improve customer engagement, and build stronger brands. By understanding the key factors contributing to social media success, businesses can create a targeted marketing strategy that resonates with their audience and drives brand growth.

In conclusion, this study aims to provide businesses with a framework for analyzing their social media success. By developing a predictive model, businesses can gain valuable insights into what factors drive engagement on social media. Utilizing this model can help businesses optimize their social media strategies, leading to more significant brand building and overall business success.

Problem Description

Understanding your customer base is essential for any business, especially on social media. Social media metrics, such as Impressions, Engagement, and Reach, play a vital role in creating an effective marketing strategy. These metrics help businesses predict the impact of their social media content on brand building. This study aims to develop a predictive model that uses the characteristics of social media posts to determine their effects on Facebook. By analyzing the factors contributing to social media success, businesses can tailor their content to their target audience's interests and increase customer engagement, leading to more robust brand building and overall business success.

The study will analyze various factors contributing to social media post success, such as post type, post length, timing, use of multimedia, and target audience demographics. By identifying the most significant factors driving engagement, businesses can develop a predictive model that helps them make informed decisions about their social media content. Utilizing this model can help businesses optimize their social media strategies, improve customer engagement, and build stronger brands. Understanding the key factors contributing to social media success can increase brand awareness, customer loyalty, and, ultimately, more significant business success.

Dataset

The dataset analyzed in this study was obtained from the Facebook page of an anonymous cosmetics brand, which tracks a comprehensive set of social media performance metrics relevant to the page. The dataset consists of 500 rows, covering posts published during 2014. The dataset includes seven features known before post publication and twelve features used to evaluate the post's impact. These metrics were carefully selected to provide a more comprehensive understanding of social media's impact on brand building.

The four main features of the dataset are identification, content, categorization, and performance. The identification features enable unique identification of each post, while the content feature includes the post's textual content. Categorization features categorize the post based on whether it's an action, product, or inspiration. Performance metrics such as lifetime post total reach, lifetime post total impressions, lifetime engaged users, and lifetime post consumers are also included in the dataset.

Other variables included in the dataset are page total likes, post type (link, photo, status, or video), category (used for manual content characterization), post month, post weekday, and post hour. Additionally, the paid variable distinguishes between paid and non-paid posts, and there are variables for comments, total likes, total shares, and total interactions (sum of comments, likes, and shares).

The Data set contains the following features:

- Identification—features that allow identifying each individual post.
- Content—the textual content of the post
- Categorization—features that characterize the post.
- Performance—metrics for measuring the impact of the post (or the impact of the page, in the case of “Lifetime Post Consumers”)

The Facebook metrics dataset has the below variables:

- Page total likes: No of people who liked the page
- Type: Type of content Link/Photo/Status/Video
- Category: Manual content characterization: action (special offers), product (direct advertisement,) and inspiration (non-explicit brand related content)
- Post Month: Month the post was published [1 – 12]
- Post Weekday: Weekday the post was published [1-7]
- Post Hour: Hour the post was published [0-23]
- Paid: Paid /nonpaid [1/0]
- 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.
- Comment: Total Comment on the post
- Like: Total like on the post
- Share: Total Share of the post
- Total Interactions: Sum of Comment + Like + Share

Overall, this dataset provides a rich source of information for predicting social media performance metrics and evaluating their impact on brand building.

Data Pre-Processing:

Before building a model, performing a data preprocessing is essential. For our purposes, we have focused exclusively on the Lifetime Post Consumers performance metric, which we considered as our response variable. To simplify our data set, we have eliminated other metrics, to this we have also removed Shares, Likes and Comments as these were already a part of Total Interaction. Furthermore, we have created dummy variables for the Type and Category variables, which we added to the data set by removing the original columns. We have also removed null values and zero values for the descriptors as we were using a Linear regression model which had log transformations. This approach ensures that our data set contains only the most critical and informative variables for our analysis.

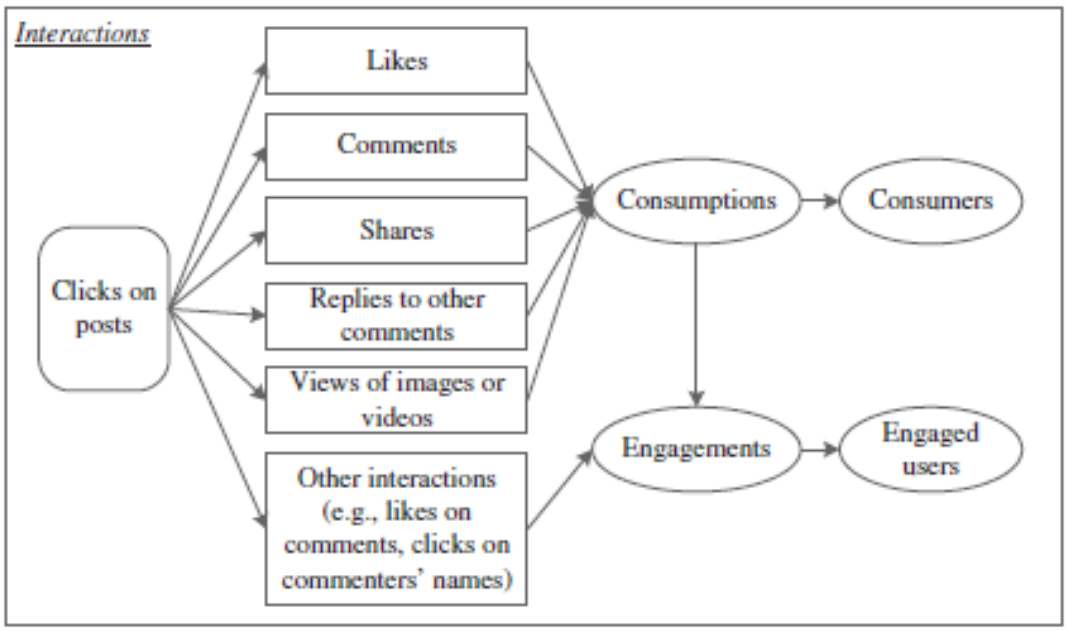


Figure 1: Interaction Criterion of a post

Procedure:

Exploratory Analysis of Variables

The initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

The Histogram of the target variable (Lifetime post-consumers) had a non-symmetrical distribution; therefore a log transformation was performed which turned out to have a symmetric distribution.

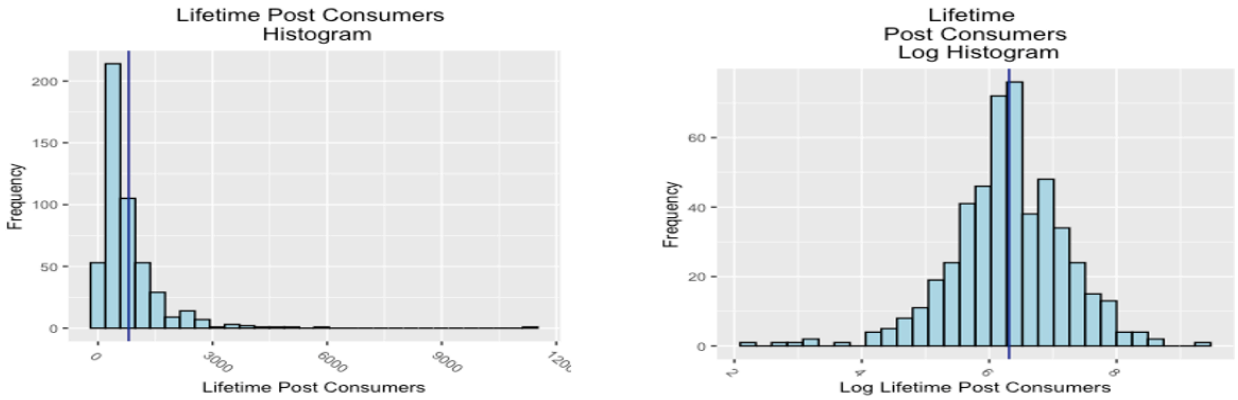
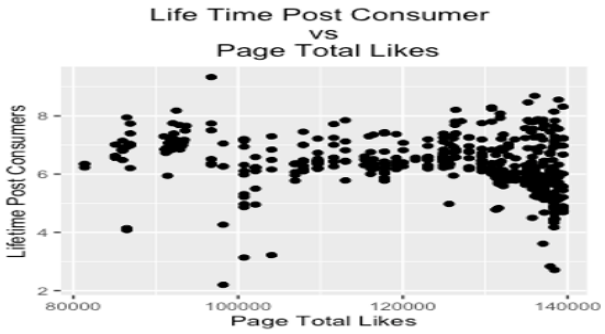


Figure 2: Histogram for LPConsumer and Log(LPConsumers)

To visualize the strength of the relationship between the response variable and individual descriptors, a Scatter plot analysis was done on continuous descriptors (LPConsumer vs page total like & Total interaction), and Box Plot analysis was done on Numeric variables. From the scatterplot there seems to be a weak linear relationship with a non-constant variance. From the visualization of scatterplot on total interaction, a log transformation was performed which gave better linearity. From the boxplots, it is evident that post hour, post weekday, post month and paid has a negligible effect on LPConsumer.
It's also evident that Total Interaction has the highest influence on the response variable.



Correlation Values

	PTLike	LPConsumer	TotalInterac
PTLike	1.00000000	-0.1496334	0.04831435
LPConsumer	-0.14963338	1.0000000	0.34941125
TotalInterac	0.04831435	0.3494112	1.00000000

Figure 3: Scatterplot and Correlation Matrix

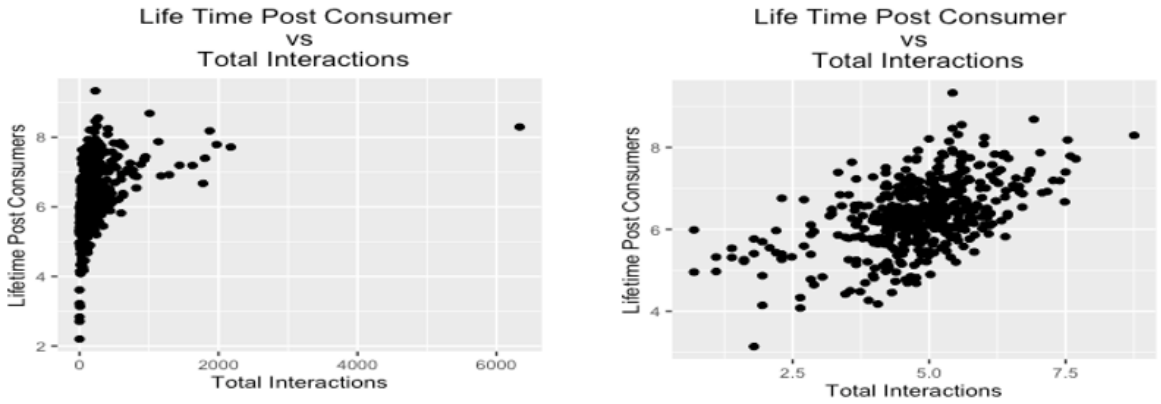
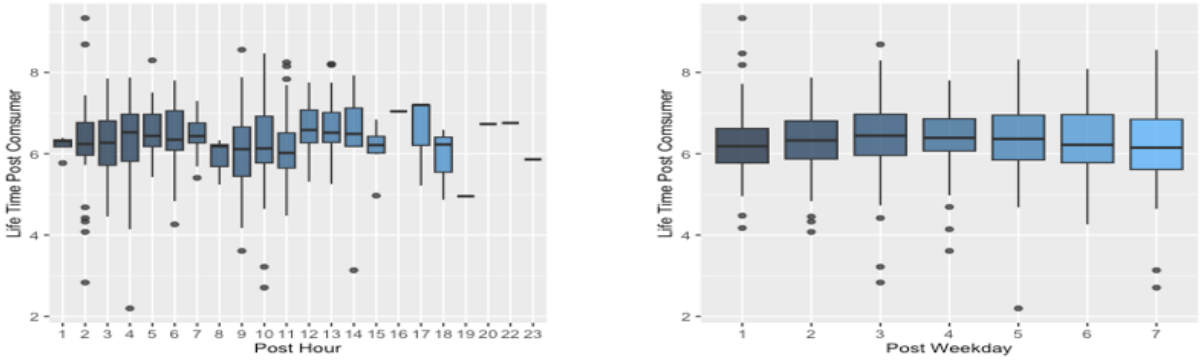


Figure 4: Scatterplot of Total Interaction



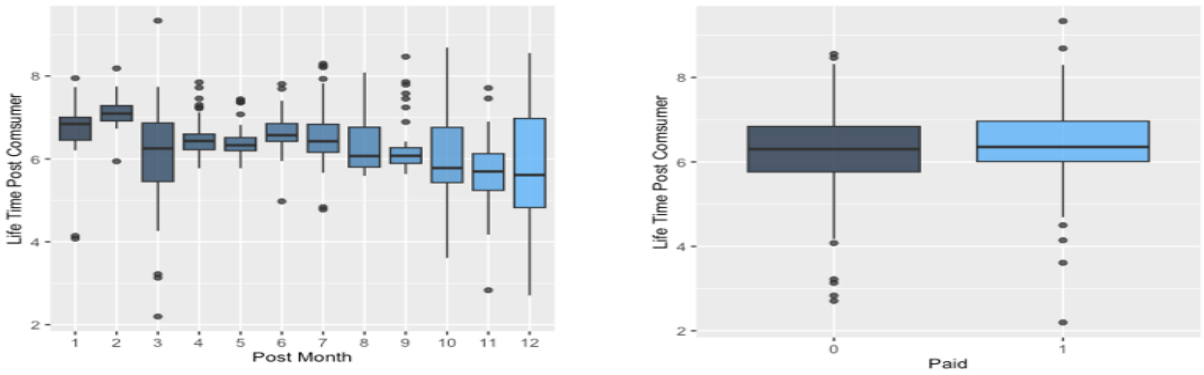


Figure 5: Boxplots of Post Hour, Post weekday, Post month and Paid

Model Selection

To create comparison models, we had 3 different approaches.

- No Interaction Model: This includes all descriptors without any interaction.
- Partial Interaction Model: This includes all descriptors with paid as an interaction with other variables.
- Full Interaction Model: This includes all the descriptors interacting with each other.

No Interaction Model:

On the full interaction model the multiple R^2 is 61.86% having some non-significant variable. Hence a backward variable selection process was performed, where all the variables were significant with multiple R^2 as 61.46%.
However, the multiple R^2 is less than the previous model but the F statistic score is high making the model simpler. Hence this was the first model.

```
Call:
lm(formula = log(LPConsumer) ~ PTLike + PosHr + Paid + log(TotalInterac) +
    typeP + typeS + typeV + category1 + category2, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21410 -0.26494 -0.00665  0.26933  1.96768

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.591e+00  2.407e-01  23.223  < 2e-16 ***
PTLike            -2.104e-05  1.533e-06 -13.725  < 2e-16 ***
PosHr              3.325e-03  5.573e-03   0.597    0.551
Paid               9.047e-02  5.308e-02   1.704    0.089 .
log(TotalInterac)  4.119e-01  2.321e-02  17.748  < 2e-16 ***
typeP              1.035e+00  1.186e-01   8.729  < 2e-16 ***
typeS              2.202e+00  1.483e-01  14.849  < 2e-16 ***
typeV              1.579e+00  2.310e-01   6.836 2.47e-11 ***
category1          4.481e-01  6.010e-02   7.457 4.18e-13 ***
category2          9.520e-02  6.761e-02   1.408    0.160
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5185 on 480 degrees of freedom
Multiple R-squared:  0.6186, Adjusted R-squared:  0.6114
F-statistic: 86.48 on 9 and 480 DF,  p-value: < 2.2e-16
```

Figure 6: Full Non-Interaction Model

Backward variable selection process

```
Call:
lm(formula = log(LPConsumer) ~ PTLike + Paid + log(TotalInterac) +
    typeP + typeS + typeV + category1, data = mydata)

Coefficients:
      (Intercept)             PTLike               Paid  log(TotalInterac)
        5.626e+00         -2.079e-05          8.762e-02           4.094e-01
            typeP              typeS              typeV          category1
        1.043e+00          2.253e+00          1.588e+00           4.143e-01
```

Final Non-Interaction Model:

```
Call:
lm(formula = log(LPConsumer) ~ PTLike + log(TotalInterac) + typeP +
    typeS + typeV + category1, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-2.32981 -0.26168  0.00091  0.27016  2.00912

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.620e+00  2.330e-01  24.124  < 2e-16 ***
PTLike             -2.077e-05  1.490e-06 -13.935  < 2e-16 ***
log(TotalInterac)   4.149e-01  2.291e-02  18.115  < 2e-16 ***
typeP               1.042e+00  1.172e-01   8.894  < 2e-16 ***
typeS               2.247e+00  1.443e-01  15.568  < 2e-16 ***
typeV               1.605e+00  2.299e-01   6.979 9.85e-12 ***
category1           4.202e-01  5.307e-02   7.918 1.67e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5196 on 483 degrees of freedom
Multiple R-squared:  0.6146, Adjusted R-squared:  0.6098
F-statistic: 128.4 on 6 and 483 DF,  p-value: < 2.2e-16
```

Figure 7: Final Model 1

From the Analysis of Variance and Variance Inflation Factor (VIF), all the variables are significant and there is no multicollinearity.

```
Analysis of Variance Table

Response: log(LPConsumer)
                  Df  Sum Sq Mean Sq F value    Pr(>F)
PTLike             1  25.023  25.023  93.028 < 2.2e-16 ***
Paid               1   4.580   4.580  17.026 4.343e-05 ***
log(TotalInterac)  1  96.231  96.231 357.756 < 2.2e-16 ***
typeP              1  12.324  12.324  45.816 3.784e-11 ***
typeS              1  39.378  39.378 146.394 < 2.2e-16 ***
typeV              1  14.718  14.718  54.718 6.223e-13 ***
category1          1  16.376  16.376  60.880 3.794e-14 ***
Residuals        482 129.650   0.269
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variance Inflation Factor

PTLike	log(TotalInterac)	typeP	typeS
1.053647	1.150277	3.194659	3.154352

typeV	category1
1.351221	1.245845

Figure 8: ANOVA and VIF (Model 1)

Residual Plot Analysis:

The residual analysis was performed which shows randomness, normality and having the QQ plots showing linearity with some noise to the extreme ends which could be possible outliers.

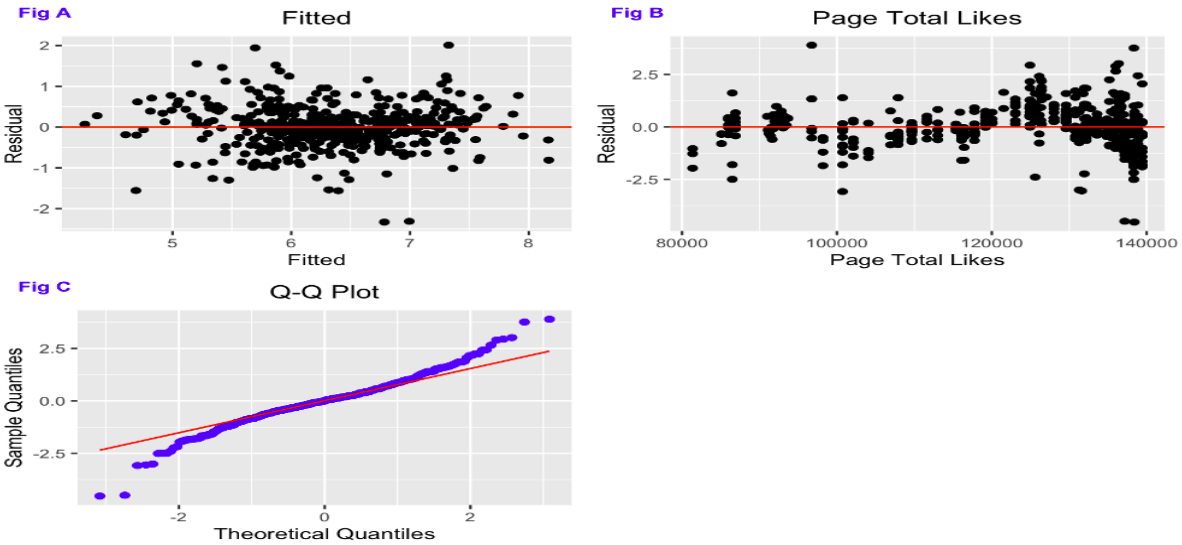


Figure 9: Residual Plots

Analyzing the hat value and cook's distance we see there are some outliers and influential points which is evident from other visual analysis.

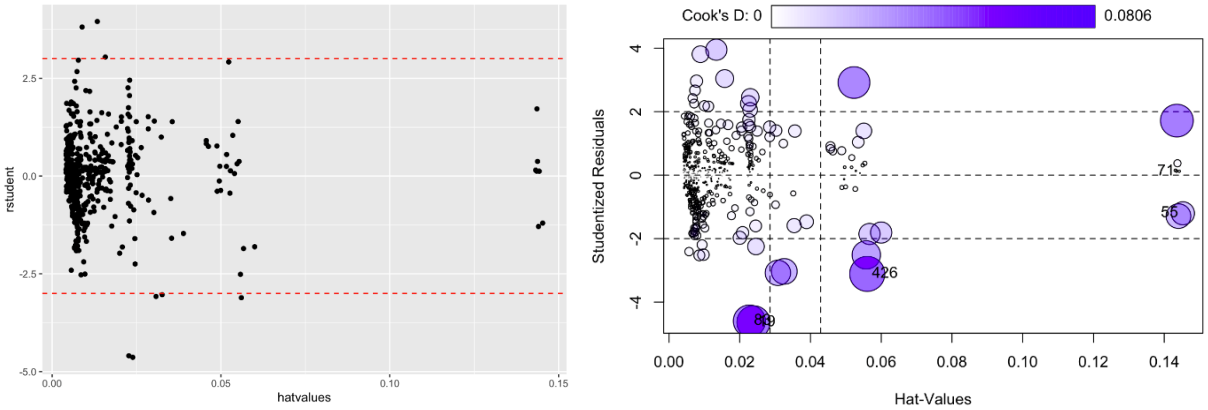


Figure 10: Influential Plot

Partial Interaction:

Even though paid was not a significant indicator according to our visualization, but paid posts play a crucial role in brand building on a social media platform, Hence its interaction was evaluated with the rest of the descriptors, and it was evident from our model that paid did not have effect on the other descriptors, and therefore we ended up with the same model as the previous one.

```
Call:
lm(formula = log(LPConsumer) ~ Paid * (PTLike + typeP + typeS +
    log(TotalInterac) + typeV + category1), data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-2.33456 -0.26515 -0.00041  0.26074  1.97198

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.553e+00  2.714e-01  20.458  < 2e-16 ***
Paid                     3.288e-01  5.412e-01   0.608    0.544
PTLike                  -2.055e-05  1.729e-06 -11.887  < 2e-16 ***
typeP                    1.005e+00  1.376e-01   7.305 1.18e-12 ***
typeS                    2.248e+00  1.686e-01  13.339  < 2e-16 ***
log(TotalInterac)        4.243e-01  2.703e-02  15.699  < 2e-16 ***
typeV                    1.453e+00  3.322e-01   4.373 1.51e-05 ***
category1                4.324e-01  6.394e-02   6.764 3.96e-11 ***
Paid:PTLike             -7.248e-07  3.472e-06  -0.209    0.835
Paid:typeP               1.640e-01  2.676e-01   0.613    0.540
Paid:typeS               2.629e-02  3.342e-01   0.079    0.937
Paid:log(TotalInterac) -5.682e-02  5.369e-02  -1.058    0.290
Paid:typeV               3.465e-01  4.795e-01   0.723    0.470
Paid:category1          -4.493e-02  1.165e-01  -0.386    0.700
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5207 on 476 degrees of freedom
Multiple R-squared:  0.6184, Adjusted R-squared:  0.608
F-statistic: 59.34 on 13 and 476 DF,  p-value: < 2.2e-16
```

Figure 11: Partial Interaction Model 2

Full Interaction Model:

Upon performing a full interaction model, it can be seen that most of the descriptors are insignificant at alpha<0.05. The R^2 value came to 66.09% with an F-statistic of 40.71.

```
Call:
lm(formula = log(LPConsumer) ~ (PTLike + Paid + typeP + typeS +
    log(TotalInterac) + typeV + category1)^2, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-2.23684 -0.26875 -0.02323  0.25290  1.98642

Coefficients: (4 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 8.817e-01  1.046e+00   0.843   0.3997
PTLike                      5.128e-07  8.635e-06   0.059   0.9527
Paid                        3.606e-01  5.179e-01   0.696   0.4867
typeP                       4.417e+00  7.787e-01   5.672 2.48e-08 ***
typeS                       1.649e+00  1.622e+00   1.017   0.3099
log(TotalInterac)           1.226e+00  1.881e-01   6.521 1.83e-10 ***
typeV                       7.242e+00  7.441e+00   0.973   0.3309
category1                   7.862e-01  5.268e-01   1.492   0.1363
PTLike:Paid                -3.029e-07  3.306e-06  -0.092   0.9270
PTLike:typeP               -1.075e-05  5.803e-06  -1.853   0.0644 .
PTLike:typeS               -7.493e-07  1.237e-05  -0.061   0.9517
PTLike:log(TotalInterac)   -2.779e-06  1.397e-06  -1.990   0.0472 *
PTLike:typeV               -5.457e-05  4.986e-05  -1.094   0.2744
PTLike:category1            2.641e-06  3.122e-06   0.846   0.3981
Paid:typeP                  3.217e-01  2.633e-01   1.222   0.2224
Paid:typeS                 -1.934e-02  3.295e-01  -0.059   0.9532
Paid:log(TotalInterac)     -1.026e-01  5.215e-02  -1.968   0.0497 *
Paid:typeV                  6.519e-01  4.836e-01   1.348   0.1783
Paid:category1             -1.827e-02  1.095e-01  -0.167   0.8676
typeP:typeS                        NA         NA      NA       NA
typeP:log(TotalInterac)    -4.627e-01  1.033e-01  -4.480 9.40e-06 ***
typeP:typeV                        NA         NA      NA       NA
typeP:category1            -3.673e-01  3.875e-01  -0.948   0.3437
typeS:log(TotalInterac)     8.329e-02  1.470e-01   0.567   0.5712
typeS:typeV                        NA         NA      NA       NA
typeS:category1            -3.837e-01  5.037e-01  -0.762   0.4466
log(TotalInterac):typeV     1.195e-01  3.115e-01   0.384   0.7014
log(TotalInterac):category1 -7.309e-02  4.859e-02  -1.504   0.1332
typeV:category1                    NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4843 on 465 degrees of freedom
Multiple R-squared:  0.6776, Adjusted R-squared:  0.6609
F-statistic: 40.71 on 24 and 465 DF,  p-value: < 2.2e-16
```

Figure 12: Full interaction Model 3

All the non-significant descriptors were removed, and the model was supported by a backward stepwise selection process. The R^2 is 65.98% which is less than the full interaction model, but the F-statistic score is 104 with all the descriptors being significant making this a better model.

```
Call:
lm(formula = log(LPConsumer) ~ PTLike + typeP + typeS + log(TotalInterac) +
    typeV + category1 + PTLike:typeP + PTLike:log(TotalInterac) +
    typeP:log(TotalInterac), data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-2.23186 -0.27005 -0.01573  0.26218  1.96499

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 9.930e-01  8.336e-01   1.191  0.23414
PTLike                      3.807e-06  6.754e-06   0.564  0.57328
typeP                       4.392e+00  5.991e-01   7.330 9.84e-13 ***
typeS                       1.561e+00  1.632e-01   9.568  < 2e-16 ***
log(TotalInterac)           1.218e+00  1.658e-01   7.349 8.65e-13 ***
typeV                       7.625e-01  2.422e-01   3.148  0.00174 **
category1                   3.893e-01  4.978e-02   7.821 3.35e-14 ***
PTLike:typeP               -1.194e-05  4.683e-06  -2.551  0.01106 *
PTLike:log(TotalInterac) -3.006e-06  1.255e-06  -2.395  0.01703 *
typeP:log(TotalInterac)  -4.884e-01  6.780e-02  -7.203 2.29e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4851 on 480 degrees of freedom
Multiple R-squared:  0.666,  Adjusted R-squared:  0.6598
F-statistic: 106.4 on 9 and 480 DF,  p-value: < 2.2e-16
```

Final Model

```
Analysis of Variance Table

Response: log(LPConsumer)
                          Df  Sum Sq Mean Sq  F value    Pr(>F)
PTLike                     1  25.023  25.023 106.3189 < 2.2e-16 ***
typeP                      1  11.526  11.526  48.9739 8.772e-12 ***
typeS                      1  54.651  54.651 232.2032 < 2.2e-16 ***
log(TotalInterac)          1  84.587  84.587 359.3985 < 2.2e-16 ***
typeV                      1  15.182  15.182  64.5043 7.506e-15 ***
category1                  1  16.924  16.924  71.9087 2.803e-16 ***
PTLike:typeP               1   3.540   3.540  15.0421 0.0001198 ***
PTLike:log(TotalInterac)   1   1.661   1.661   7.0589 0.0081499 **
typeP:log(TotalInterac)    1  12.213  12.213  51.8900 2.286e-12 ***
Residuals                480 112.972   0.235
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 13: Final Full Interaction Model 3

From the Analysis of Variance and Variance Inflation Factor (VIF), all the variables are significant and there is no multicollinearity.

Result

Model Summary

In this dataset analysis, three models were constructed to identify the top-performing model that could explain the maximum variation of the dependent variable. The final model was chosen based on its ability to explain 65.98% variation in the dependent variable, considering the selected independent variables. The final model equation was presented, which provides a mathematical representation of the relationship between the dependent variable (log (LPConsumer)) and independent variables (PTLike, typeP, typeS, log(TotalInterac), typeV, category1, PTLike: typeP, PTLike:log(TotalInterac), and typeP:log(TotalInterac)). This equation can be used to make predictions about the dependent variable based on the values of the independent variables. The final model can be considered a useful tool for future analyses as it provides valuable insights into the relationship between the dependent and independent variables.

```
Call:
lm(formula = log(LPConsumer) ~ PTLike + typeP + typeS + log(TotalInterac) +
    typeV + category1 + PTLike:typeP + PTLike:log(TotalInterac) +
    typeP:log(TotalInterac), data = mydata)
```

Figure 14: Full Interaction Model 4

Review key metrics

Upon closer inspection of the model summary, the critical metrics presented in the given information are crucial indicators of the quality and validity of a linear regression model. For example, the R-squared value of 0.666 signifies that approximately 66.6% of the variation in the response variable can be explained by the predictor variables employed in the model. This implies that the model can effectively capture a significant portion of the data's variability, making it valuable in predicting and drawing conclusions regarding the association between the predictor variables and the response variable.

The adjusted R-squared value of 0.6598 is a modified version considering the number of predictor variables in the model. It offers a more precise estimate of the proportion of variation in the response variable that can be explained by the predictor variables while adjusting for the number of predictors in the model. A higher adjusted R-squared value implies that the model can better explain the variability in the response variable while controlling for the number of predictor variables. The F-statistic of 106.4 and its associated p-value of <2.2e-16 provide insights into the overall significance of the model. The F-statistic is computed by comparing the variation explained by the model to the variation not defined by the model. A high F-statistic value coupled with a low p-value indicates that the model is statistically significant and that the predictor variables possess a robust linear relationship with the response variable. In this instance, the p-value is significantly lower than the commonly used significance level of 0.05, signifying that the relationship between the predictor variables and the response variable is highly significant.

Overall, these critical metrics suggest that the linear regression model utilized in this analysis is a reliable and accurate representation of the data. In addition, the predictor variables employed in the model exhibit a strong linear relationship with the response variable. These metrics offer essential information for researchers, analysts, and decision-makers who rely on linear regression models to make predictions and conclusions regarding the relationship between variables.

```
Residual standard error: 0.4851 on 480 degrees of freedom
Multiple R-squared:  0.666,  Adjusted R-squared:  0.6598
F-statistic: 106.4 on 9 and 480 DF,  p-value: < 2.2e-16
```

Confidence Interval

To better understand the relationship between the independent variables in the model, please refer to the appropriate table that illustrates the 95% confidence interval for each attribute. This table will provide the confidence intervals for each independent variable in the model.

	2.5 %	97.5 %
(Intercept)	-6.449017e-01	2.630957e+00
PTLike	-9.464860e-06	1.707850e-05
typeP	3.214333e+00	5.568783e+00
typeS	1.240404e+00	1.881561e+00
log(TotalInterac)	8.925037e-01	1.543929e+00
typeV	2.866074e-01	1.238406e+00
category1	2.915339e-01	4.871637e-01
PTLike:typeP	-2.114688e-05	-2.743537e-06
PTLike:log(TotalInterac)	-5.472391e-06	-5.392596e-07
typeP:log(TotalInterac)	-6.216273e-01	-3.551799e-01

Figure 15: Confidence Interval Model 3

Residual Plot Analysis

The residual plot analysis provides insight into the effectiveness of the model's predictions. In this case, the residual output indicated that the residuals were relatively symmetrical, and the model was making predictions equally well at both the high and low ends of the dataset. This implies that the model's predictions were reasonably accurate and unbiased. Additionally, the QQ plot suggested linearity, indicating that the model's assumptions regarding the normality of the residuals were reasonable. In short, the results of the residual plot analysis suggest that the linear regression model used in this analysis was effective and reliable in predicting the relationship between the predictor variables and the response variable.

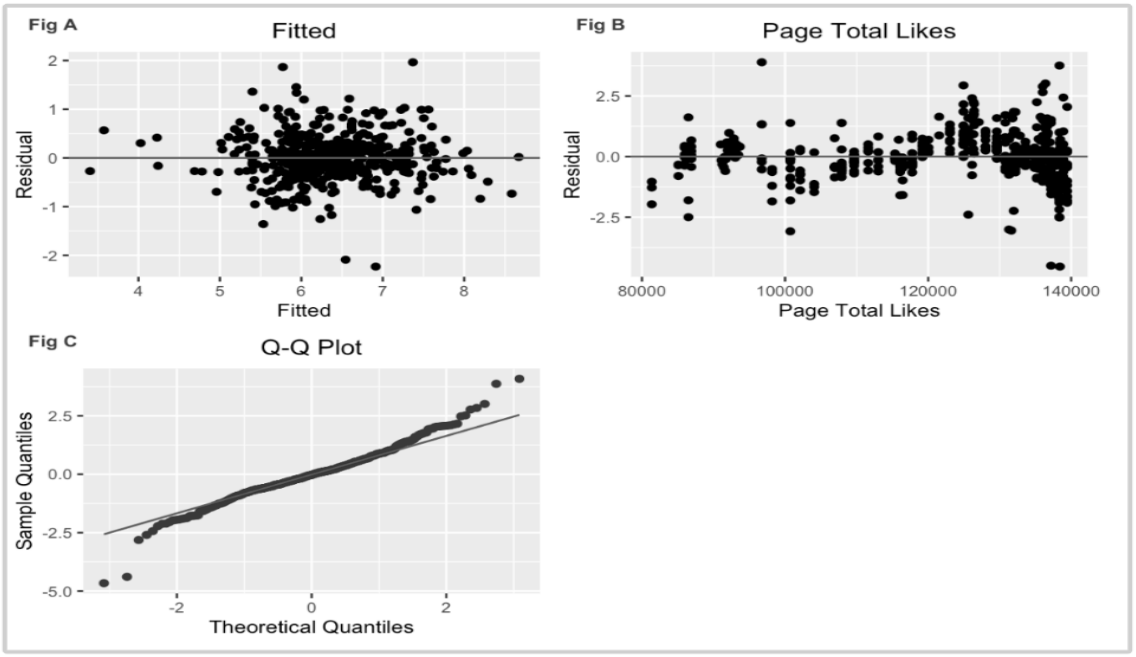


Figure 16: Residual Plot (Model 3)

Model Validation

The key metrics used in the model validation process are RMSE, MAE, and MAPE. Model 3 (fit_full_interaction_6) was found to minimize all three of these validation metrics. As a result, it can be concluded that this model provides more accurate predictions closer to the actual values.

	Model	RMSE	MAE	MAPE	Diff_R2
1	fit_full_3	0.4944894	0.3766650	6.094994	0.008043312
2	fit_inter_Paid_3	0.4944894	0.3766650	6.094994	0.008043312
3	fit_full_interaction_6	0.4663610	0.3629355	5.868759	0.004332479

Figure 17: Model validation

Standardized Beta Coefficients

The standardized beta coefficients derived from our model indicate that the total interaction term had the most significant effect, with a coefficient of exp(1.611329) = 5.009464 on Lifetime Post Consumers.

Total interaction is a crucial metric for measuring the success of a post, as it reflects the number of likes, shares, comments, and other forms of engagement. Analyzing engagement data can provide valuable insights into the effectiveness of posts and help users identify areas for improvement. Additionally, analyzing the consumer lifetime value of posts can give useful information about their effectiveness in generating long-term engagement and loyalty. By considering these metrics, users can optimize their social media strategy and improve the performance of their posts.

PTLike	typeP	typeS	log(TotalInterac)
7.407307e-02	1.892543e+00	5.425615e-01	1.611329e+00
typeV	category1	PTLike:typeP	PTLike:log(TotalInterac)
1.089008e-01	2.313110e-01	-2.324298e-01	-1.295361e-06
typeP:log(TotalInterac)			
-1.697578e-01			

Figure 18: Standardized Beta Coefficients

Conclusion

Through our analysis of Facebook metrics data, we have shown that engagement is a crucial metric for measuring the success of a post. By examining the number of likes, shares, comments, and other forms of interaction, users can gain valuable insights into the effectiveness of their social media strategy. Furthermore, by considering the consumer lifetime value of posts, users can gain a deeper understanding of their audience's engagement and loyalty over time.

Our analysis involved creating and evaluating three models, each providing valuable insights into the relationship between post engagement and other key metrics. The best-performing model included interaction terms and had an R-squared value of 0.666, indicating that it explained a significant portion of the variance in the data. The model also had an Adjusted R-squared value of 0.6598, which suggests that it is a reliable and robust predictor of post engagement.

Overall, our analysis demonstrates the value of data-driven decision-making in social media marketing. By analyzing Facebook metrics data, users can better understand their audience's engagement, loyalty and develop more effective strategies for maximizing their reach and engagement. We hope our findings will be helpful to social media marketers and others seeking to optimize their online presence.

Division Of Work:

Category of work Done.

- Jiten Mishra: Data Cleaning, R Code, Power Point, Final Report
- Erik Pak: R Code, Power Point, Final Report
- Denvir Gama: Power Point, Final Report
- Liang Tse Hsu: Power Point, Final Report
- Martello Pollock: Audio presentation, Final Report

Future Work:

In Our Analysis we performed Linear Regression model analysis on only Lifetime post Consumers however there are other metrics which are part of the data set, but we excluded it keeping in mind the timeframe of the project.
We would like to go ahead with this data set and try to create new models for other metrics and analyze their effects on Brand building on Facebook platform.

We would like to find some additional data to see how other brand's post's impact the brand building with the current model that we have created.

APPENDIX

Figure Table

References:

Dataset Reference: https://archive.ics.uci.edu/ml/datasets/Facebook+metrics

Research Reference:

Ballantine, P. W., et al., 2015
P.W. Ballantine, Y. Lin, E. Veer

The influence of user comments on perceptions of Facebook relationship status updates

Computers in Human Behavior, 49 (2015), pp. 50-55

View PDFView articleView in ScopusGoogle Scholar

Bianchi, C. and Andrews, L., 2015
C. Bianchi, L. Andrews

Investigating marketing managers' perspectives on social media in Chile

Journal of Business Research, 68 (12) (2015), pp. 2552-2559

View PDFView articleView in ScopusGoogle Scholar