The Impact of Facebook Post Features on Consumer Engagement

Lab Section: C3

Team Members: Jack Campbell, Zhan Ding, Ji Hoon Lee, Jarrett Petto, Wei Shao, Mengyan Sun

Abstract

This report offers a predictive analysis of user interactions with Facebook posts, using a dataset encompassing various engagement metrics from a cosmetics brand`s page. The study aims to construct a model to forecast the `Lifetime Post Consumers` count by examining eight specific input features related to post content and timing. Employing linear regression techniques, we analyze posts to identify significant predictors of consumer engagement. The results confirm that certain variables, such as post type and hour, significantly affect user interactions. The model demonstrates proficiency in predicting user engagement, providing insights for optimizing social media marketing strategies.

Introduction

In the evolving landscape of social media marketing, understanding consumer behavior is pivotal for crafting effective strategies. This report delves into a comprehensive analysis of Facebook metrics, aiming to construct a predictive model based on historical data. The data comprises user interactions with posts from a well-known cosmetics brand`s Facebook page, covering various performance metrics.

Our research is rooted in the dataset labeled `facebook\_updated.csv`, which separates observations into `Training` and `Validation` groups for model construction and evaluation. The primary goal is to predict `Lifetime Post Consumers`—a metric defining the number of unique users who have interacted with a post. To achieve this, we utilize eight variables: `TypePhoto`, `TypeStatus`, `TypeVideo`, `Paid`, `ln.Page.Total.Likes`, `ln.Page.Total.Likes:seasonspring`, `ln.Page.Total.Likes:seasonsummer`, and `ln.Page.Total.Likes:seasonwinter`. These predictors are selected based on insights from a prior analysis, including interaction terms between `Page Total Likes` and dummy variable `Season`.

This study aims to uncover the underlying patterns in user engagement and identify significant variables that influence consumer interaction on Facebook. By doing so, we aspire to provide actionable insights for marketers to optimize their content strategy and enhance user engagement.

Data

The dataset comprises posts published in 2014 on the Facebook page of a well-known cosmetics brand. It contains data points for 500 posts, represented across 20 columns—with one that divides the observations into `Training` and `Validation` subsets. The six categorical variables are `Type`, `Category`, `Post Month`, `Post Weekday`, `Post Hour`, and `Paid`: a binary feature indicating whether a post was paid to be promoted.. These variables are our primary predictors of interest and have been manually examined to ensure accuracy, i.e., eliminating erroneous entries such as '13' for `Post Month`. `Page Total Likes` is known before the post is published, and the remaining 12 are used to evaluate post impact. `Lifetime Post Total Reach` and `Lifetime Post Total Impressions` quantify visualizations, while `Lifetime Post Consumers`, the number of people who clicked anywhere on a post, measures user interaction and is also our response variable of interest.

Many of the variables included in the dataset are highly correlated and provide similar insights, giving rise to an issue of multicollinearity. `Lifetime Post Consumers` and `Lifetime Engaged Users` both measure the number of people who clicked anywhere in a post, for example, emphasizing the need for careful variable selection throughout the modeling process. It is discovered that some numerical variables cannot be normalized, such as the variable `Page Total Likes`, which may impede our normality assumption and modeling results.

**Preprocessing**

When preprocessing this dataset, the R function md.pattern() is utilized to display rows with missing values. We discover a missing value in the `Paid` variable and decide to remove the row containing this missing value. Since `Paid` is a dummy variable with values of '0' and '1', replacing the missing value with '0' would not be reasonable. Numerical values such as `Page Total Likes` and `Lifetime Post Consumers` are log transformed to reduce large values; the group sees improvement in the distribution for `Lifetime Post Consumers`, but no improvement in the distribution for `Page Total Likes`.

All relevant categorical variables, such as `Category`, are converted to factors. Following this, two new categorical variables, `Season` and `Weekday,` are introduced based on the `Post Month` and `Post Weekday` variables, respectively. The `Season` variable categorizes posts into seasons “winter”, “spring”, “summer”, “autumn”, while the `Weekday` variable distinguishes between “Weekday” and “Weekend” as binary values in that order. `Worktime` is created as well to denote whether a post was made during working hours, 9 am to 6 pm, assigning a value of 1 if true and 0 otherwise.

Furthermore, an investigation into the variation of user engagement across different seasons was conducted. The interaction term `season:ln.Page.Total.likes` was incorporated into the model to examine the effects of `Page Total Likes` on the outcome within each season.

**Data Summary and Visualization**

**A B**

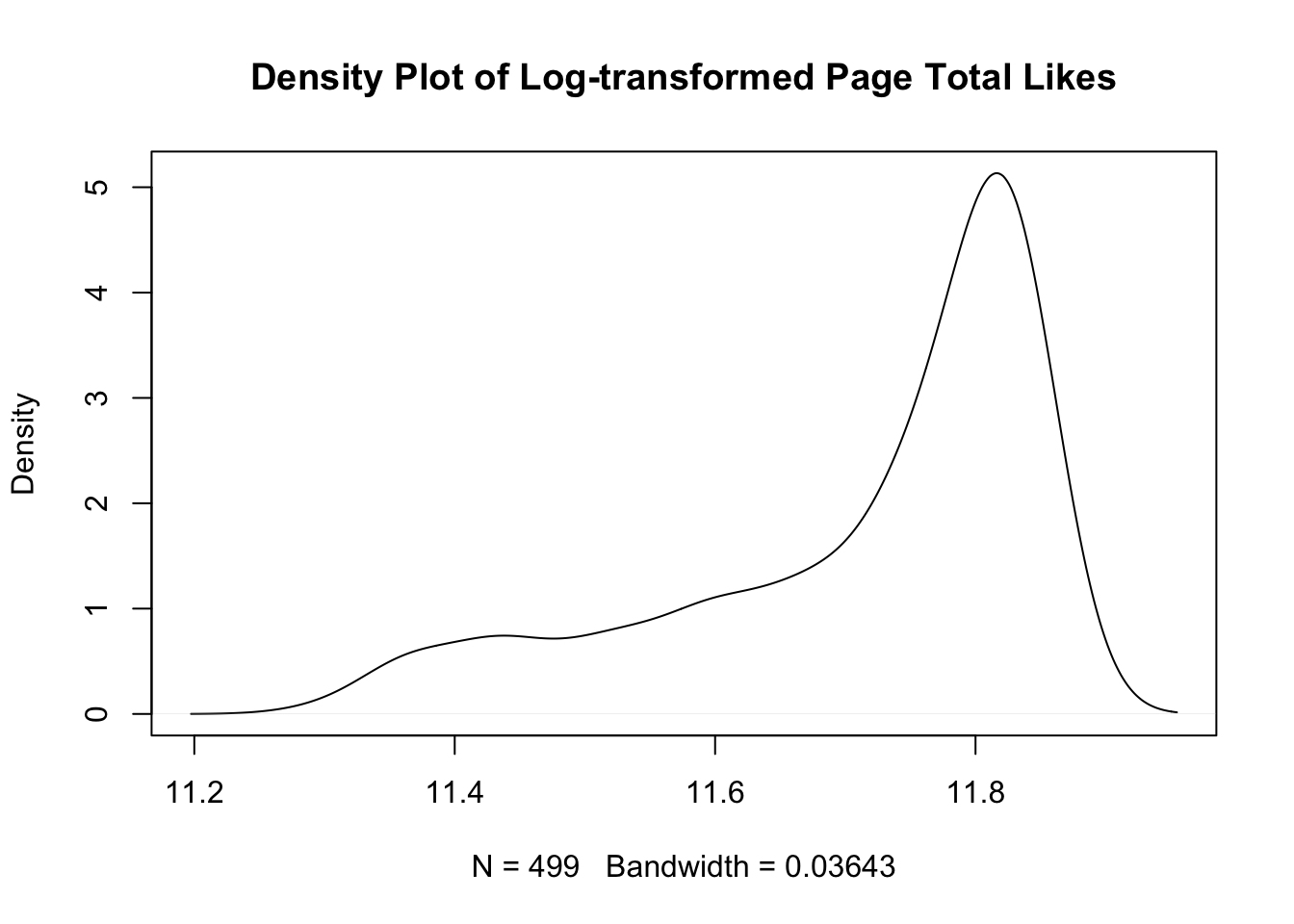
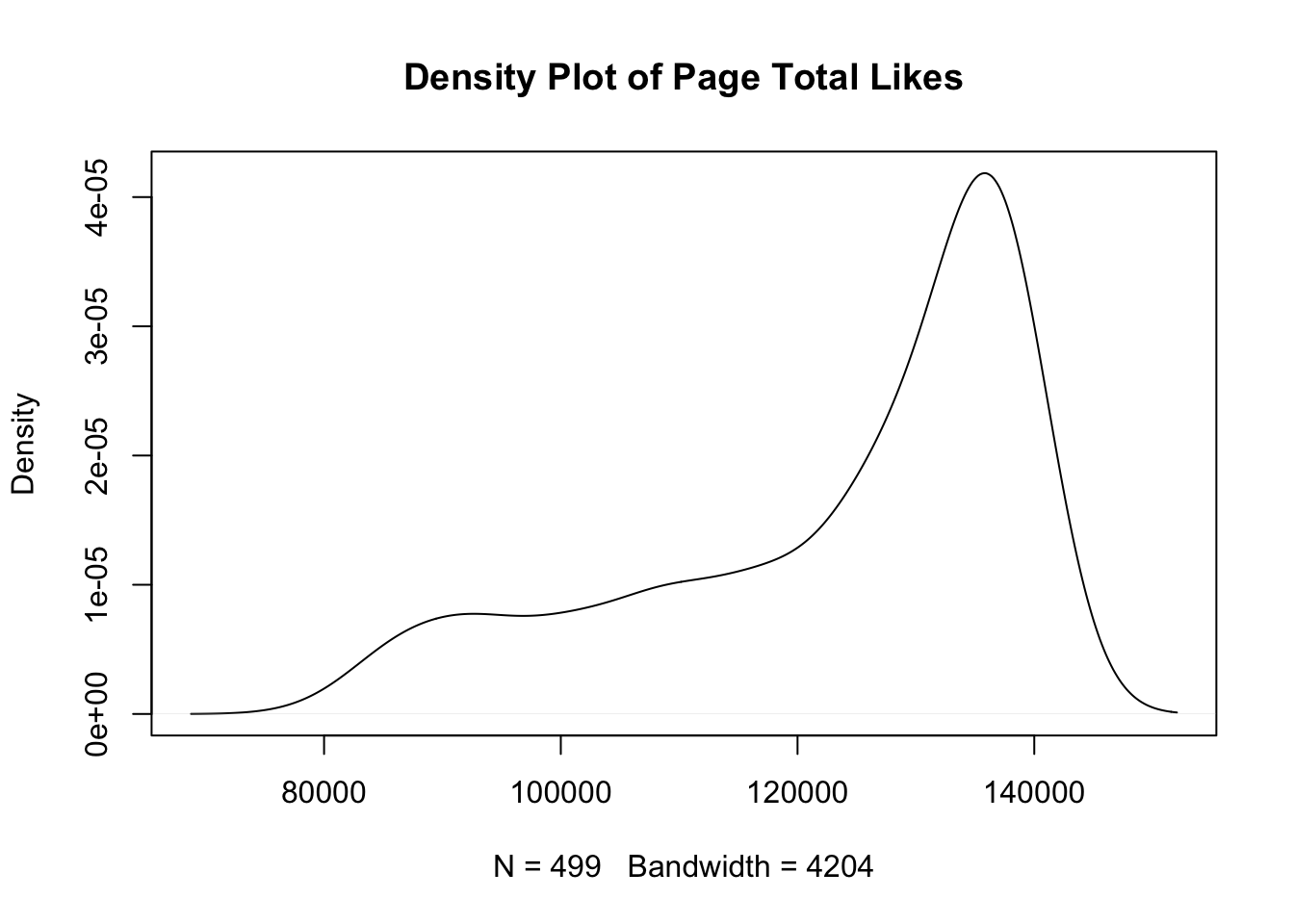


Figure 1. (A) Density Plot of `Page Total Likes`

(B) Density Plot of `ln.Page.Total.likes`

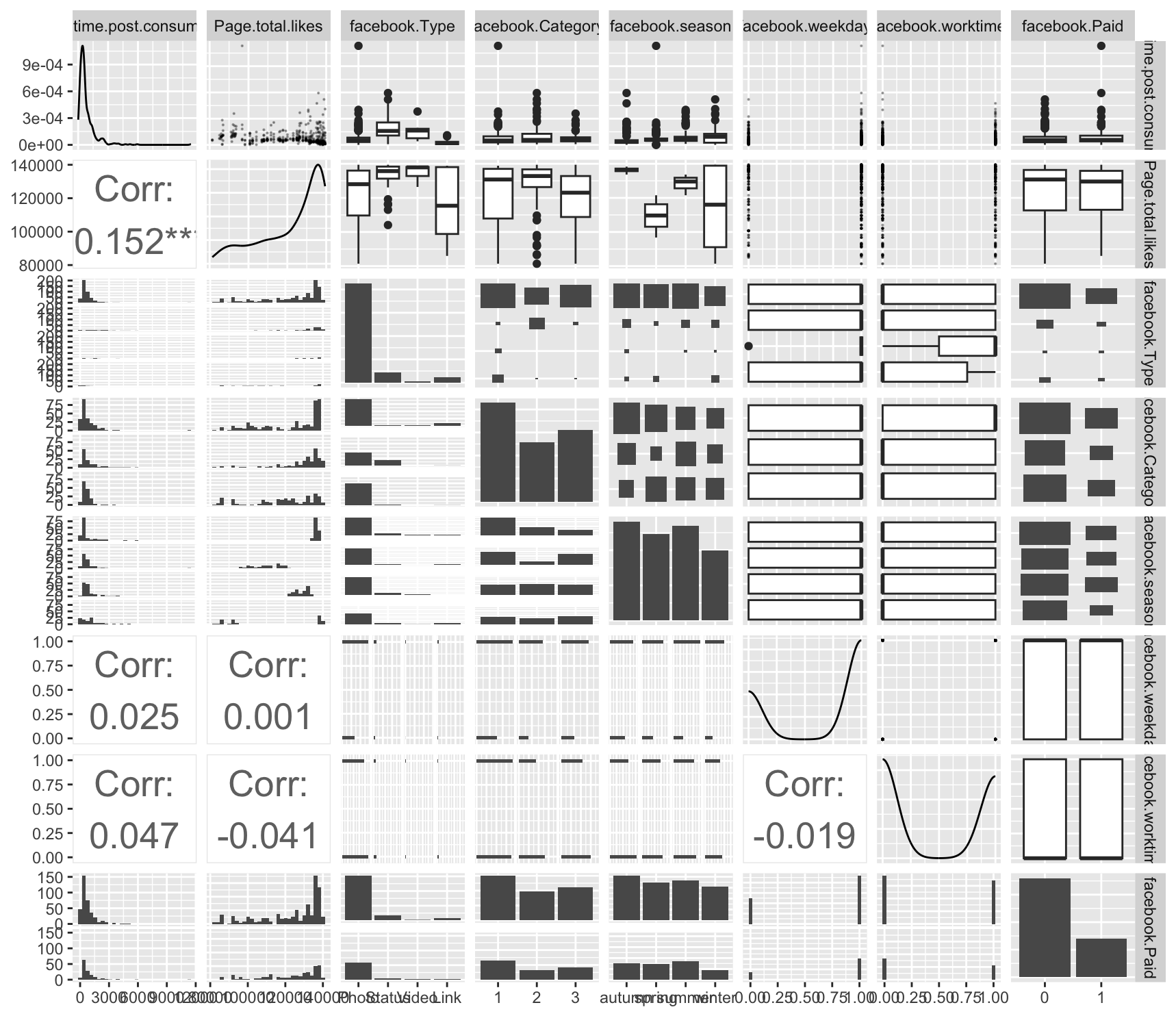
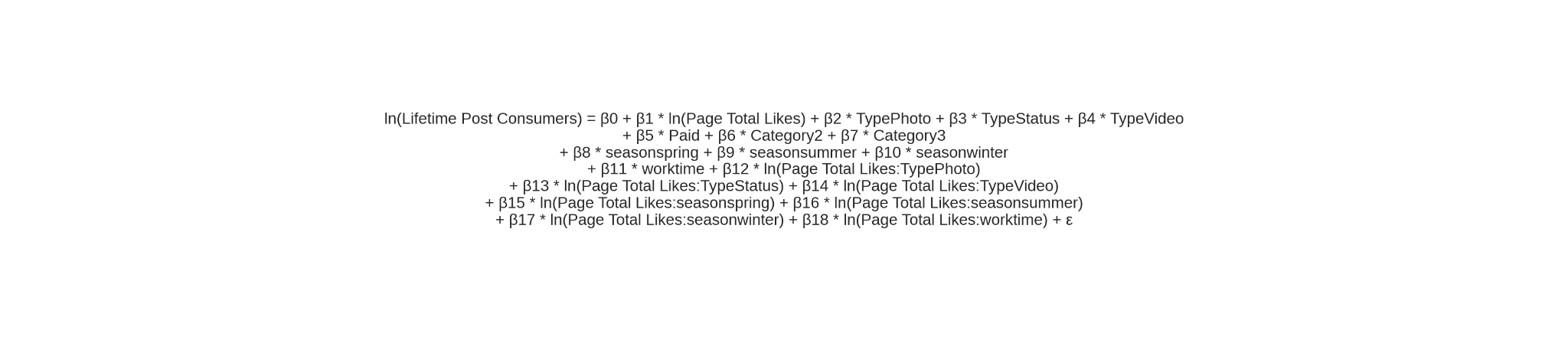


Figure 2. Scatterplot Matrix

Modeling and Analysis

The stepwise function, along with AIC and BIC scores, guided the model selection process. As shown in the table, the model selection process identifies the model with the lowest AIC. After eliminating statistically insignificant variables and addressing multicollinearity issues, the final model is derived. The AIC of this final model is comparable to that of the original model, with the added advantage of a lower BIC.

**Full Model**

****

**Reduced Model**

****

**Analysis**

The p-values associated with the coefficients indicate whether the estimated effect of each predictor is statistically significant. For instance, all post types, “Photo”, “Status”, “Video”, are found to be statistically significant, as their p-values are much smaller than the conventional significance level of 0.05. The p-value for `ln.Page.Total.likes:seasonsummer` is significant, suggesting that the interaction between `ln.Page.Total.likes` and the summer season has a significant effect on `ln.Lifetime.Post.Consumers`. The small p-value, < 2.2e-16, for the entire model also suggests that the model is statistically significant. The residual standard error is 0.7458, providing a measure of the typical difference between observed and predicted values. Adjusted 0.3314, is adjusted for the number of predictors, indicating that approximately 33.14% of the variance in `ln.Lifetime.Post.Consumers` is explained by the model. The VIF outputs are around one, suggesting that there is no multicollinearity between the variables.

The Residual vs. Fitted Values Plot and vs. Fitted values plots reveal that the residuals are not randomly scattered, but demonstrate a vertical pattern. This observation suggests a challenge to the assumption of constant variance, also known as homoscedasticity. The standard Q-Q plot demonstrates a generally fitted line against the prediction, spanning two standard deviations despite an exponential increase and decrease beyond two standard deviations. The leverage plot shows that most of the residuals fall within the -4 to 4 range, except for data point 442, indicating the model is nicely fitted. In fact, all four plots also identified points 422 and 233 as suspected outliers with high residuals and leverage that affected normality. However, upon inspecting the data, these points should not be considered outliers.

The added-variable plots correspond to the coefficient estimates and statistical significance in the output, signifying that `TypePhoto`, `TypeStatus`, `TypeVideo` all have positive linear contribution while `ln.Page.Total.likes` has a negative linear contribution. The variable `Paid` and the interaction term between `ln.Page.Total.likes` and `Season` do not significantly contribute.

**Summary**

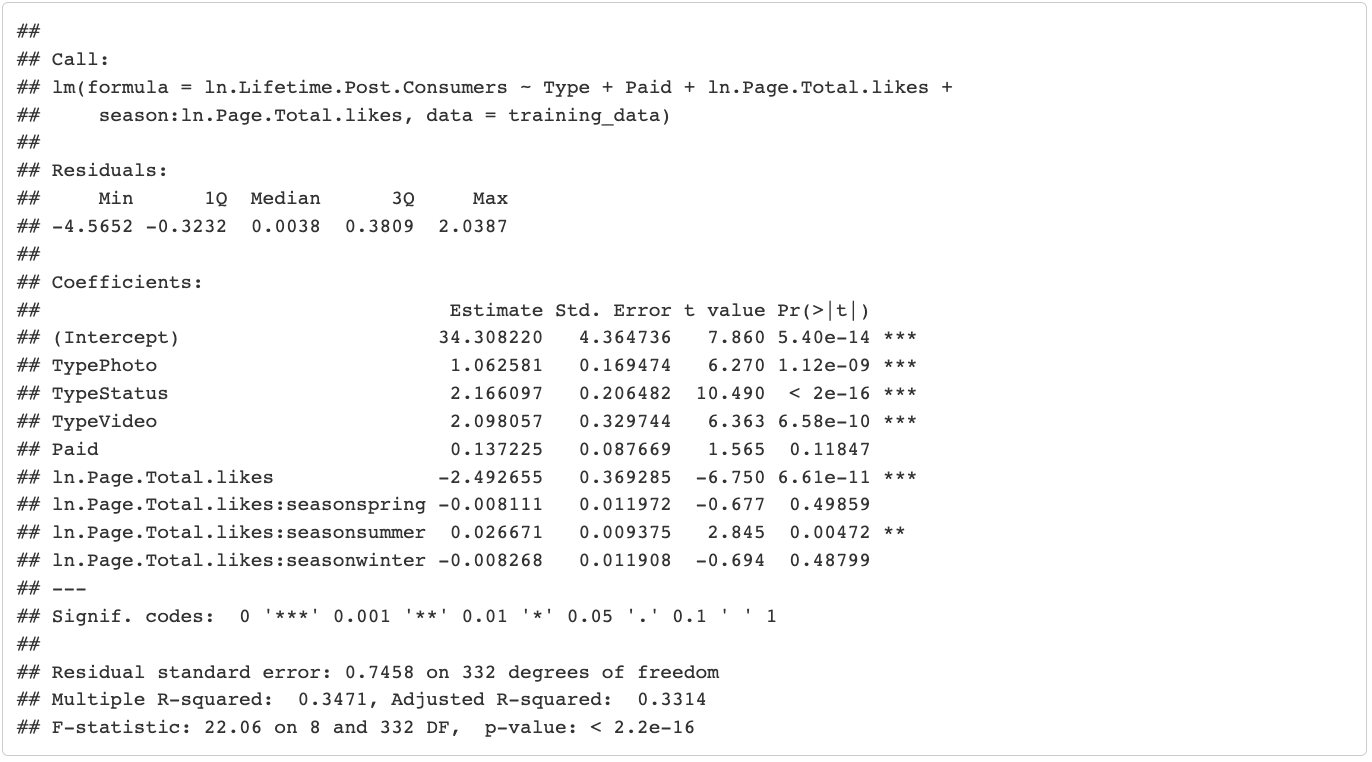
****

Table 1. Summary of Linear Regression Model

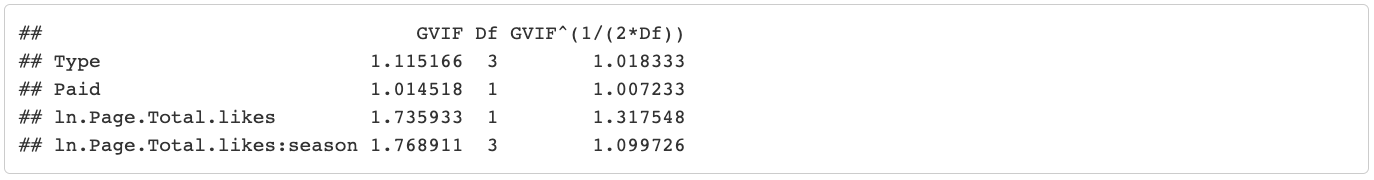


Table 2. Variation Inflation Factor

|  |  |
| --- | --- |
|  |  |

Figure 3. Diagnostic Plots

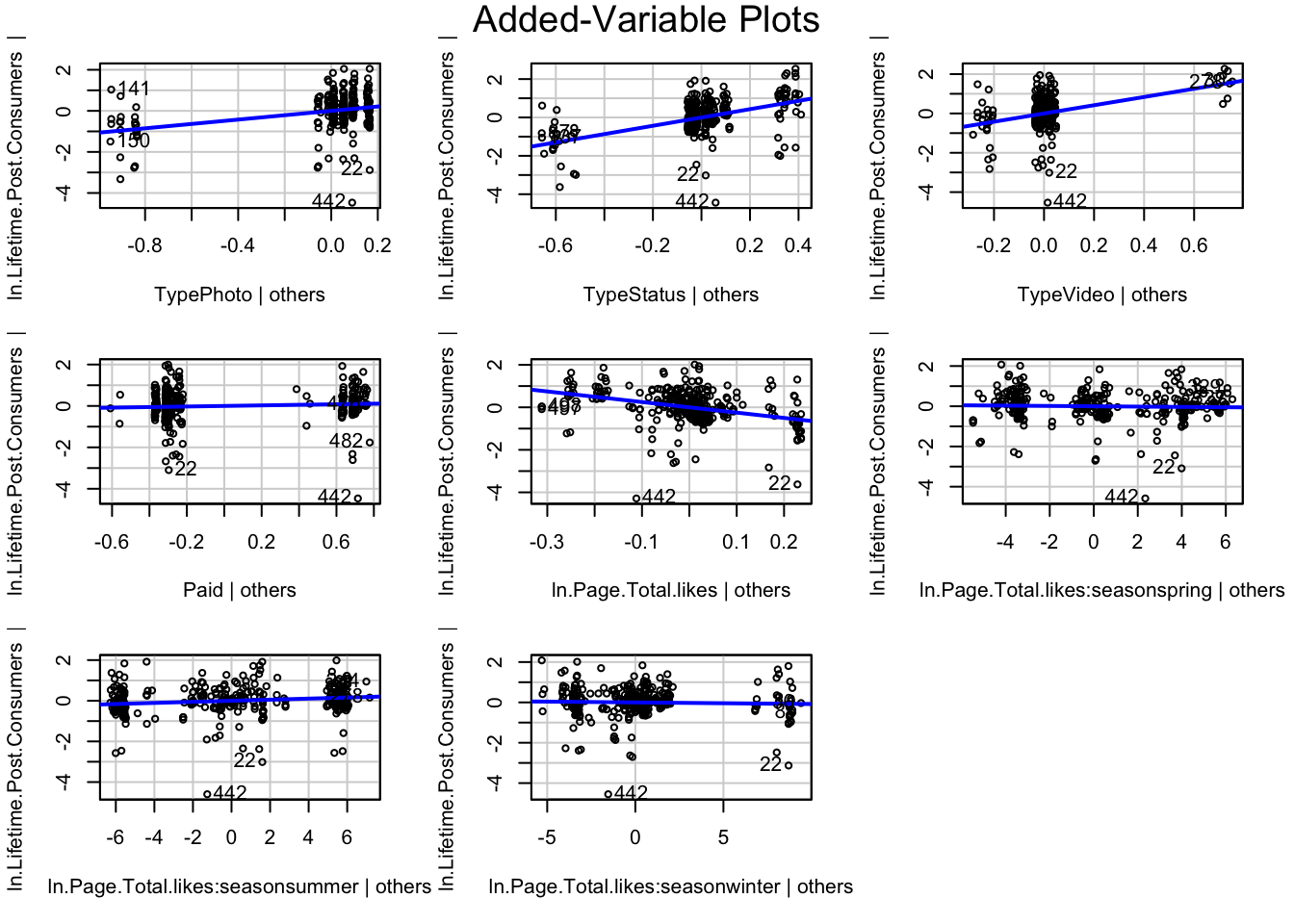


Figure 4. Added-Variable Plots

**Lasso Regression and Ridge Regression**

The lasso regression model identified the best regularization parameter (lambda) is 0.00802, which is used to minimize overfitting while retaining predictive power.

From the graph, it is clear that variable `Type`is the best predictor for `Lifetime Post Consumers`; variables `Page Total Likes` and its interaction term with `season` have negative values, posing a more complex picture for the two variables.`TypeStatus` has the strongest potential to explain the variation in `Lifetime Post Consumers`, which is coherent with our analysis of our model.

Best Lambda ( : 0.008025502

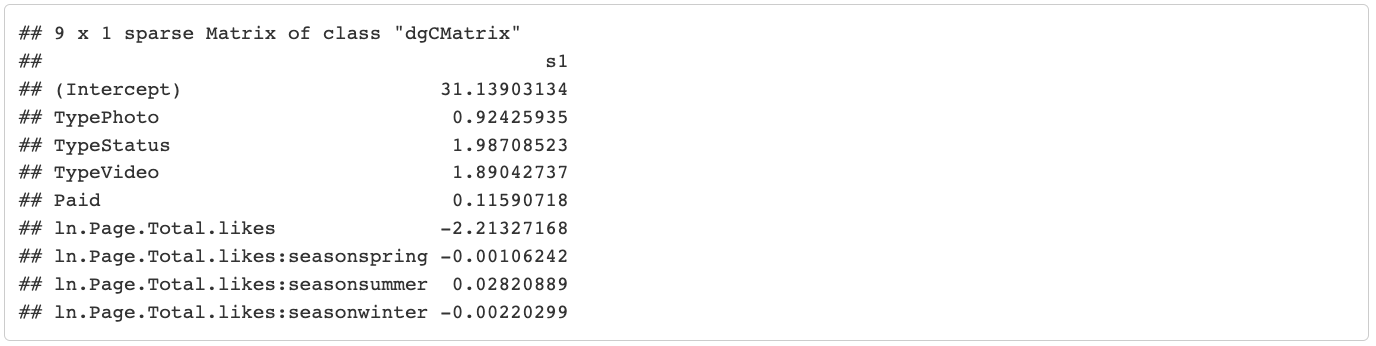


Table 3. Lasso Regression

| Red Line (2) = `TypeStatus`  Green Line (3) = `TypeVideo`  Black Line (1) = `TypePhoto`  Blue Line (5) =`ln.Page.Total.likes` |
| --- |

Figure 5. Lasso Regression Plot

Prediction

For the training data plot, the points are closely lined up with the red dashed line, indicating that the predicted values closely match the observed values. Moreover, the points do not deviate significantly from the line, indicating that overfitting is unlikely For the validation data plot, the points are more spread out than the training data point randomly, suggesting a less predictive power compared to the training data. Also, since the validation data plot performs worse than the training data, implying that the training data and not capturing underlying patterns The residuals plot, the residuals spread out randomly across the predicted line, indicates that heteroscedasticity is unlikely.

This is the result of using our model to predict the validation dataset. We have MAE = 0.4942, suggesting that the difference between predicted and observed values is around 0.4942. The R2 = 0.31, representing 31.0% of the variance in the observed values explained by the predicted values from the model. From the Observed vs. Predicted Value Plot, the model predicts well, as most points are scattered around the line without systematic deviations. However, there is some variance in the predictions, as indicated by the spread of points, especially for higher observed values. The Residual vs. Observation Index Plot shows that residuals appear to be randomly distributed around the zero line, with no obvious patterns or trends. This suggests that the model does not have systematic errors in its predictions across the range of observations.

**A B**

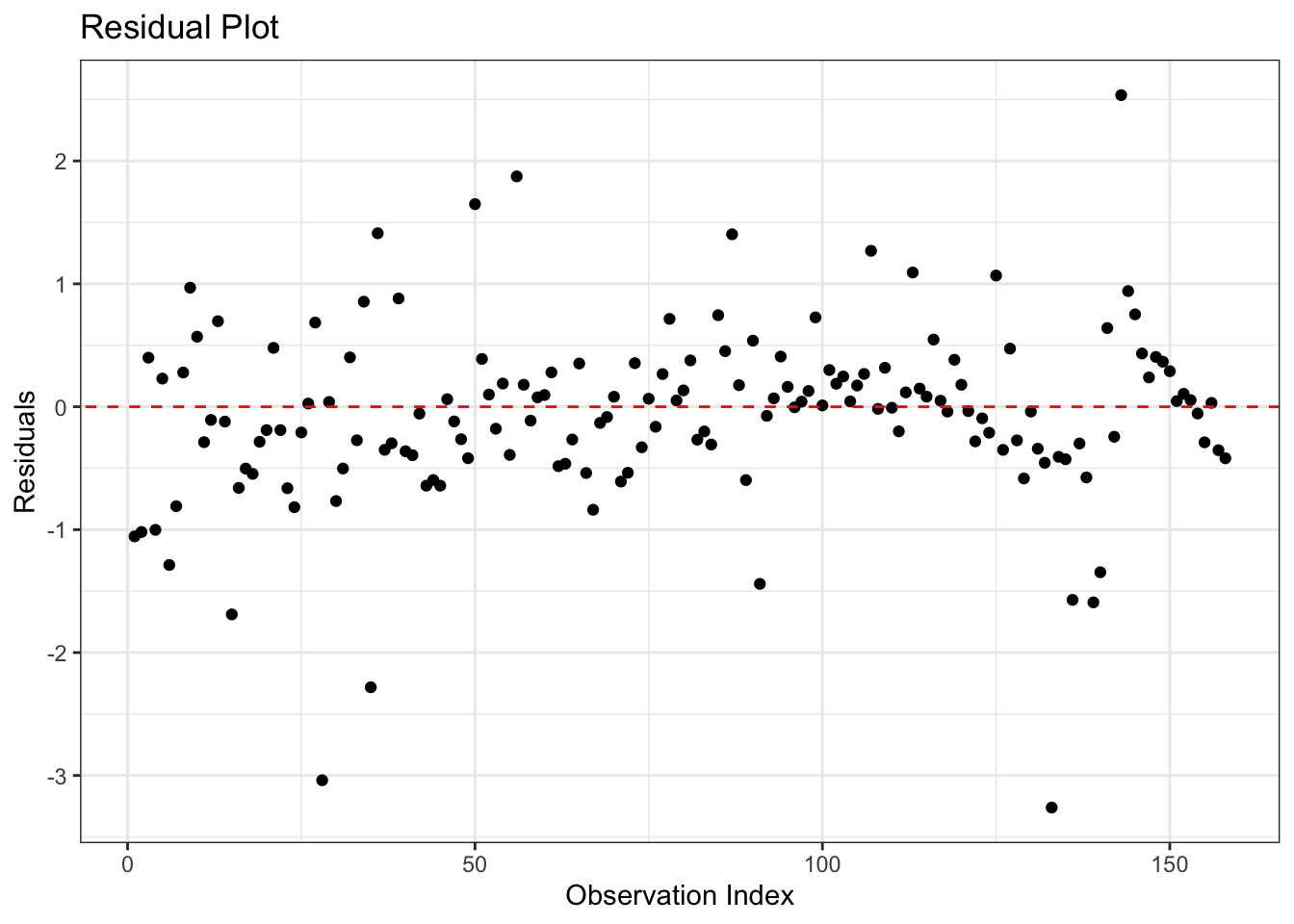
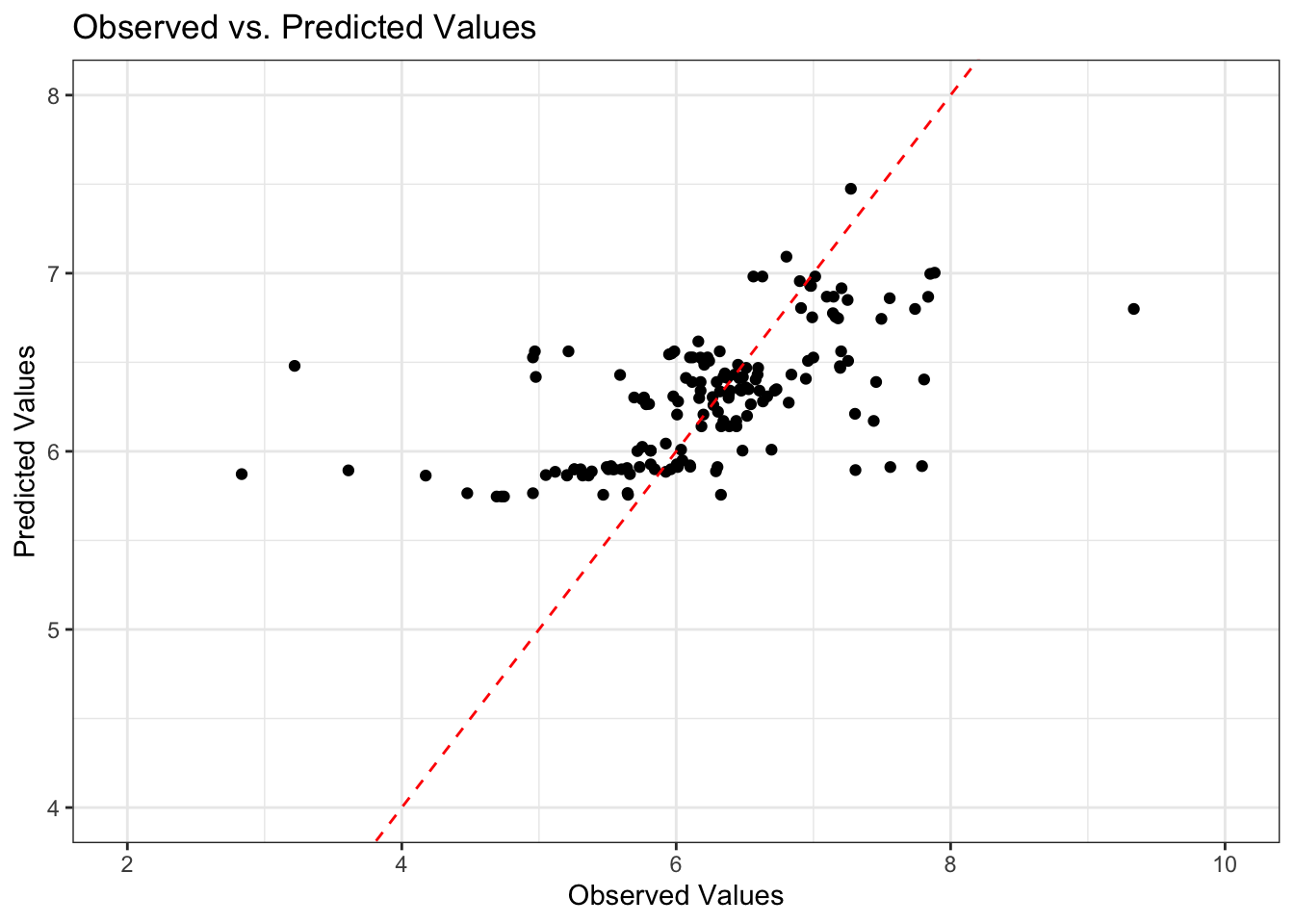


Figure 6. (A) Observed vs. Predicted Values

(B) Residual Plot

Discussion

There are multiple results from this study leading us to conclude that we’ve come up with a useful model that can predict the `ln.Lifetime.Post.Consumers`. Our AV plots show that many of our variables either have a positive or negative relationship with the `ln.Lifetime.Post.Consumers` and even though some of the variables don't show a positive or negative relationship, they can still contribute to our model and help avoid overfitting. We distinguished the variable `Type` as the most influential in building our model and predicting `ln.Lifetime.Post.Consumers`. Other diagnostics such as our VIF values, p-values of the variables, and the p-value of the model all lead us to conclude that we have a useful model to predict `ln.Lifetime.Post.Consumers`.

Although we've created a model that meets most of our standards as far as the constant variance assumption and having significant predictors, there are still a few limitations of the model. Our R2 value is fairly low, telling us that our model only represents 33.14% of the data. Additionally, our Normal Q-Q plot has tails that after two standard deviations vary greatly from the expected values and do not follow the linear path we expect it to. This prompts a closer examination of the normality assumption. Despite this, the belief persists in the model's utility. This is attributed to the statistical significance of almost all predictors, the overall statistical significance of the model, and the proper appearance of plots and prediction of the validation data for the most part.

Incorporating knowledge about the timezone and location of the brand’s user base could substantially enhance our model’s predictive accuracy, particularly concerning the `Worktime` and `Weekday` variables. Understanding when users are most active online due to regional work hours and cultural habits could allow better prediction and capitalization on peak engagement time, thus refining the strategy for post timing and potentially increasing the reach and interaction with the content.

Acknowledgements

Our sincerest appreciation goes to Professor. Julio Castrillon for his invaluable mentorship, astute guidance, and the profound impact on our academic journey through the course of our project.

Our team appreciates the support and insightful feedback from our lab instructor, Shariq Mohammed, which played a crucial role in the successful completion of our project.

Appendix

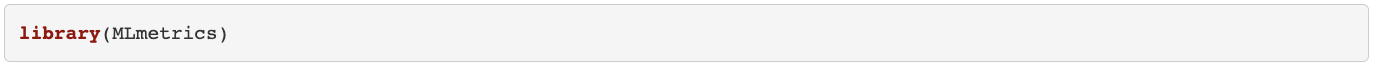










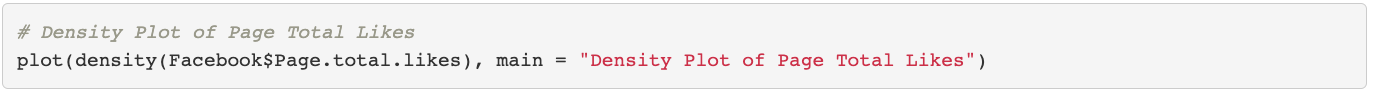








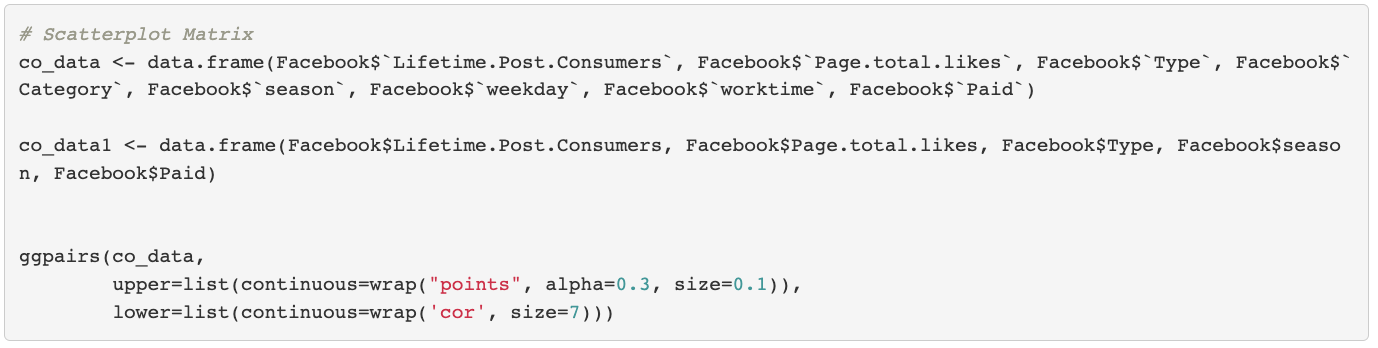
Code for Figure 1A



Code for Figure 1B



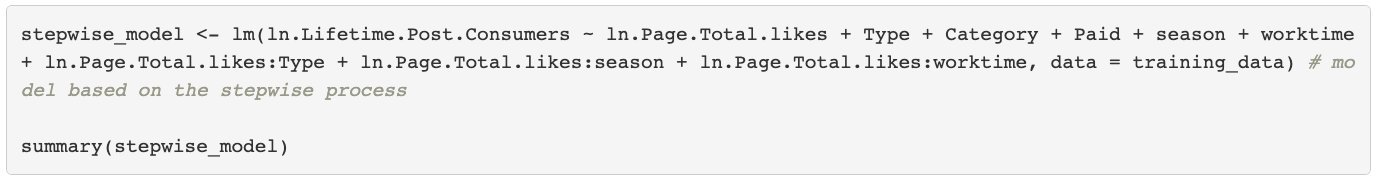
Code for Figure 2

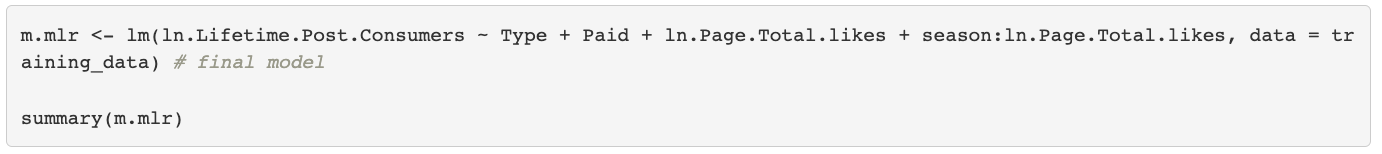




Code for Table 1







Code for Table 2



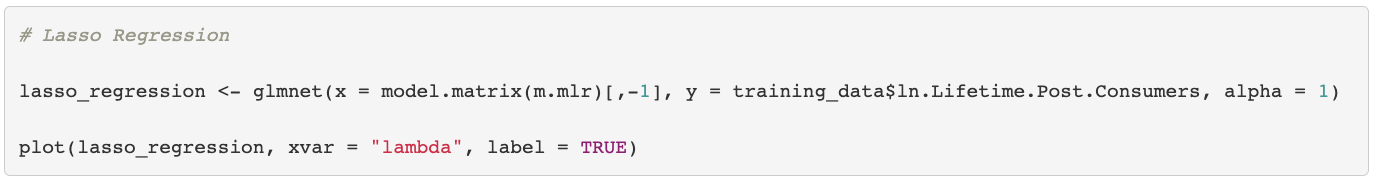
Code for Figure 3

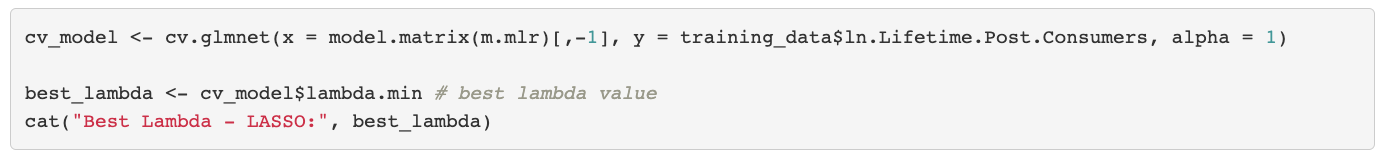


Code for Figure 4



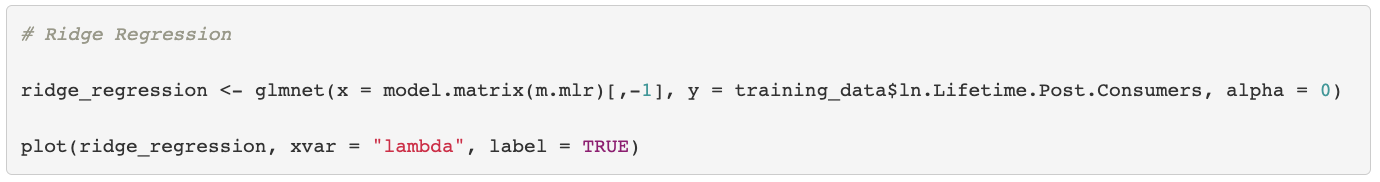
Code for Figure 5

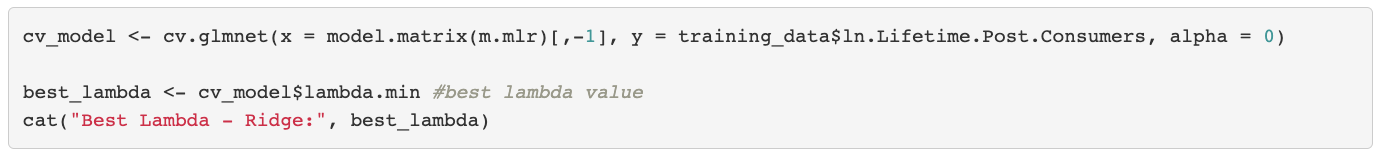




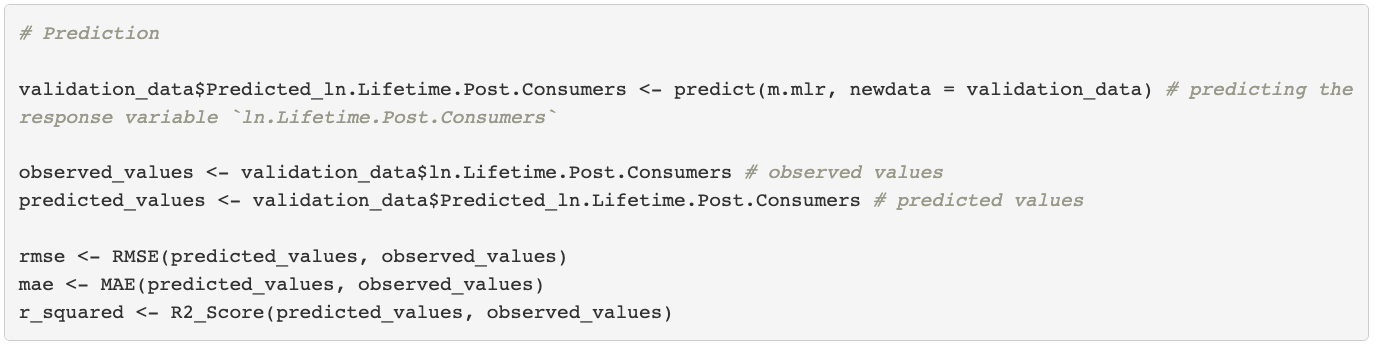
Code for Table 3



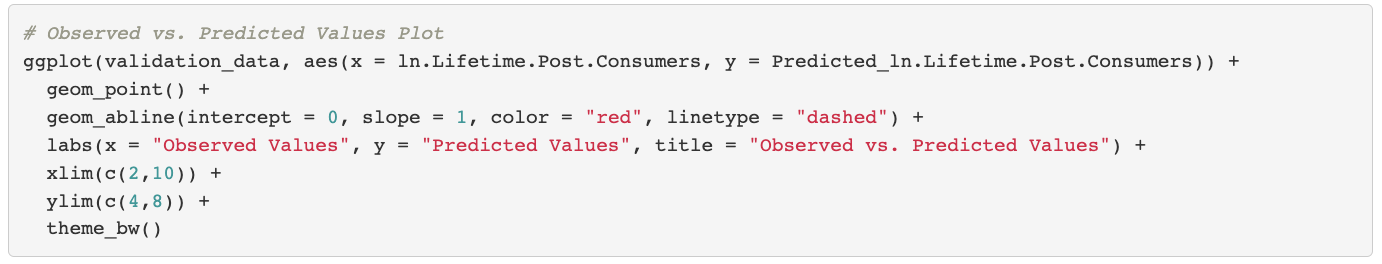








Code for Figure 6A



Code for Figure 6B

