STAT 4630 Project: Final Report

Group 32: Jeannette Jiang (jmj3bd), Elizabeth Lee (ewl3dv)

Xinyue (Chloe) Qiu (xq8zu), Emma Schloegel (eps7fa)

Executive Summary

The "Facebook Metrics" data set is from the UCI Machine Learning Repository. The data set contains 500 observations, and 19 variables that are a mix of quantitative and categorical variables regarding Facebook post characteristics and engagement metrics from a cosmetics company.

The two questions of interest that we look to answer through various analysis are as follows.

1. Can the number of likes on a Facebook post with at least one like be predicted based on certain post characteristics and engagement metrics?
2. Can we accurately predict if a Facebook post was paid for based on certain post characteristics and measures?

Our motivation for answering these two questions of interest stems from our observations that social media platforms have expanded beyond the casual purpose it once had of connecting family and friends. Facebook, among other social media platforms, is now part of the industry of advertising and customer outreach. Through different methods of analysis, we were able to answer our two questions of interest by discovering which variables could accurately predict the number of likes or accurately classify if the cosmetics company paid Facebook to advertise the post.

In light of the first question of interest, we look at the relationship between the number of likes that a post receives and different post characterizations and engagement metrics including but not limited to the number of likes, number of comments, the type of post, and day the post was published. Predicting likes on a post based on our factors would help businesses to maximize the factors that actually influence the number of likes they receive. More engagements with the post might mean more likes, which will help the company's post reach more consumers. This can help a company understand which engagement metrics are the most useful in expanding their client base.

From our analysis, we note that the predictor variables in the data set do not have a large impact in predicting the number of likes on a Facebook post. This was surprising at first as we expected as certain engagement metrics increase — such as number of shares a post receives — the number of likes on a post will increase as well. However, from our analysis, we saw that post characteristics had a relationship with each other that was different from expected.

Our motivation for our second question of interest is to predict which posts are promoted by Facebook. Promoted posts mean that Facebook is paid money to showcase the post to an audience that they use their own algorithm to target. By seeing if we can predict which posts are paid, we can investigate what is influenced when you promote a post and common metric trends in promoted posts. We can also provide insight to see if paid posts are actually more effective at targeting consumers which would be useful information for companies who spend large amounts of money paying for an advertising service they might or might not need.

The results of our analysis might be surprising, as when we test the different engagement metrics together, the relationship between certain post characteristics and predicting number of likes and whether a post was paid is not very strong. Throughout this report, we highlight possible reasons why the relationship between certain post characteristics and our variables of interest might not be as strong as intuitively predicted. We also take into account the nature of where the data set is collected from, the niche of the data set, and whether we can draw further conclusions from our analysis.

Data Cleaning and Processing

Data Cleaning

The data collection process on our end only required us to download the data set from the repository. The original data is related to posts' published during the year of 2014 on the Facebook page of a renowned cosmetics brand. The predictor variables that we chose to analyze are the number of likes and whether a post was paid, for the regression and classification setting, respectively. Our predictor variables are various engagement metrics chosen based on what was available in the data set and intuition of personal use of social media platforms such as Facebook.

We are using a combination of raw data and data that has been processed by previous researchers. Thus, the data set is ready for analysis. The raw data includes variables such as likes, comments, and shares that have not been manipulated but directly mined from Facebook. Other variables in the data set, such as Lifetime Post Consumption and Category, were determined by researchers through their own sensitivity analysis. The processed data gives us insight into different engagement metrics.

To answer our questions appropriately, we deleted rows in the dataset with zero likes from our data set as we wanted to include posts with a certain amount of interaction, and to stay consistent throughout our regression and classification analyses. We also deleted any rows with missing data (N/A) for each analysis.

Data Processing

Initial data processing for some predictor variables included coercing the Paid, Type, Category and Post Weekday variables into factors. For the variable Type, we created dummy variables for the category type of post. The dummy variables include Status, Photo, Video and Link as the 'base' indicator.

We read the paid variable as a factor to ensure classification would run smoothly. We did not have to do any data processing to the Paid categorical variable as it was already binary. The two categories for this variable included paid for advertising (1) or not paid for advertising (0).

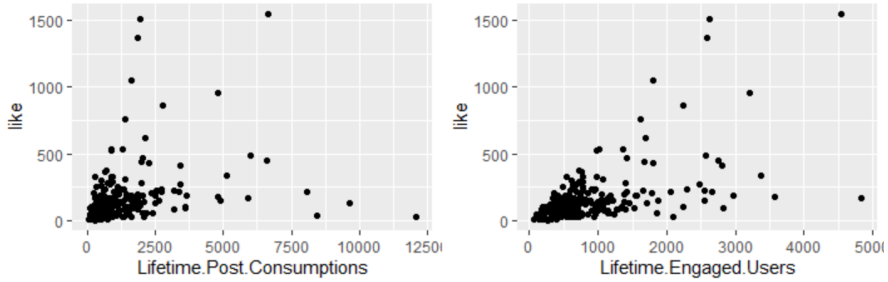
For the variable Category, we processed the categorical variable by creating a binary indicator for the variable: direct advertisement (1) and non-direct advertisement (0). The category direct advertisement contains the two initial categories 'action' and 'product' together. Action refers to posts categorized by researchers that contain special offers and contests. Product refers to posts categorized that contain direct advertisement and explicit brand content. We decided to combine the categories action and product as they pertain to the company directly advertising to consumers. The non-direct advertisement category contains the original category inspiration. Inspiration originally is described as non-explicit brand related content.

For the variable Post Weekday, we processed the categorical variable by creating a binary indicator on whether the post is posted on the weekend. We indicated this by identifying weekend (1) and weekday (0). For the weekend, we included Friday, Saturday, and Sunday. For the weekday, we included Monday, Tuesday, Wednesday, and Thursday. We decided to include Friday as part of the weekend because we note an increase in social media use on Fridays from personal experiences.

Regression Question

Our ordinary least squares regression model and regression tree analysis aims to accurately predict the number of likes on a facebook post for a cosmetic company. The specific question we are trying to answer is: can the number of likes on a Facebook post with at least one like be predicted based on certain post characteristics and engagement metrics?

Exploratory Data Analysis



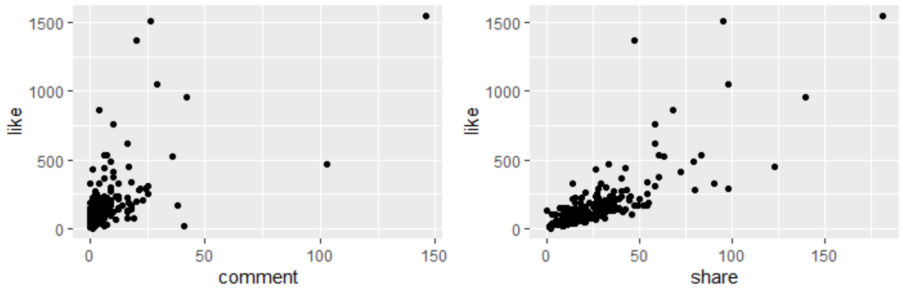


Figure A

When looking into predicting the number of likes on a Facebook post using regression analysis, the quantitative variables of Lifetime Post Consumption, Lifetime Engaged Users, Comments, and Shares ended up being significant predictors. As seen in figure A above, our quantitative variables tend to skew right and be concentrated around zero. These graphs show that the number of shares and comments a post received has a positive relationship with the number of likes. This follows what we found when performing our regression analysis, as seen in figure D of our Summary Output where shares and comments are both significant with positive coefficients. As seen in our EDA, Lifetime Post Consumption and Lifetime Engaged Users had a weaker relationship with likes with the points more scattered on the graphs. These predictors ended up having a slightly negative relationship with the number of likes on a post, as seen in figure D in the summary output.

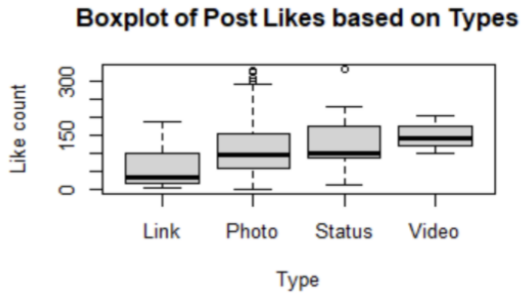


Figure B

Figure B displays the boxplot of the categorical variable Type, which consists of Link, Photo, Status, and Video. While all of the types of posts appear to differ, only status ended up being significant for predicting the number of likes on a post. You can see that this categorical variable follows the trend of the quantitative variables of being slightly skewed towards a higher number of likes while more concentrated around a lower number.

Regression Model

We decided to present our original regression model that included the variables type, lifetime post consumptions, paid, comments, shares, category, post weekday, and lifetime engaged users. Despite our attempts to improve the model, there was little change from our original regression model, and the test MSE was the lowest compared to the other model. Thus, with little to no differences in diagnostic plots, significant variables, or adjusted R-squared values, we decided to present the original regression model to stay consistent with the variables used in the regression tree model.

Diagnostic Plots

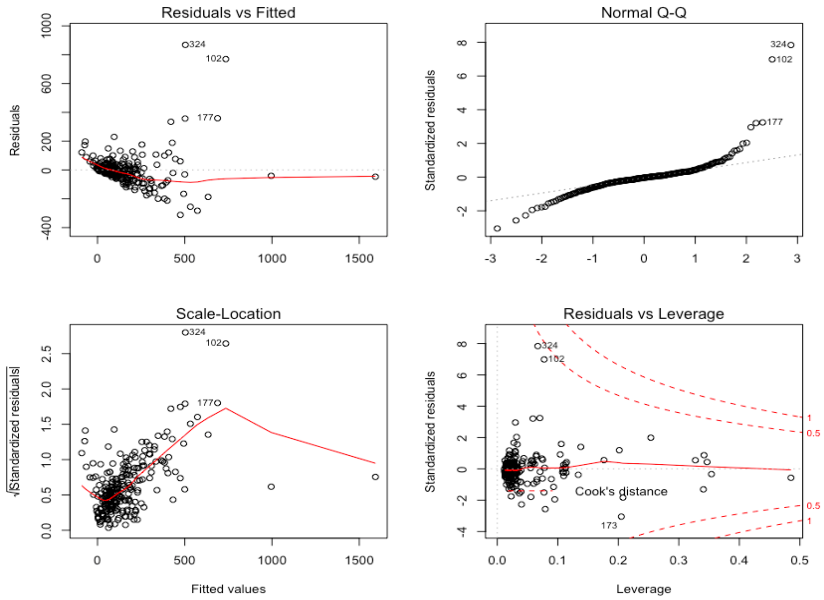


Figure C

Based on figure C above for our chosen regression model, most of the assumptions are met, but not all. The Residual Plot has a fairly constant scatter around the mean-zero line, so we conclude that the mean-zero residual assumption is met. The values on the Normal QQ Plot do not follow the diagonal line exactly, but since normality is robust, we conclude that this model meets the normality assumption. The Scale-Location Plot for the constant variance assumption is not perfect. There is a curvature pattern as shown on the plot indicating a heteroskedastic nature to the data, but after failed attempts to correct this violation including performing logarithmic transformation, this model ultimately resulted in the best model diagnostics. Finally, the Contour Plot shows no outliers beyond Cook's Distance meaning there are no significant outliers. Overall, our diagnostic plots for this regression model did not meet all regression assumptions. We speculate the violation in constant variance might provide inaccurate results.

Summary Output

```
Call:
lm(formula = like ~ Type + Lifetime.Post.Consumptions + Paid +
    Post.Weekday + Lifetime.Engaged.Users + comment + share +
    Category, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-311.78  -39.35   -3.49   28.79  868.42
```

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -1.376e+01  4.235e+01  -0.325  0.74562
TypePhoto                   -2.084e+01  3.831e+01  -0.544  0.58699
TypeStatus                  -1.612e+02  4.900e+01  -3.289  0.00116 **
TypeVideo                   -7.500e+01  7.661e+01  -0.979  0.32859
Lifetime.Post.Consumptions -2.071e-02  7.073e-03  -2.928  0.00375 **
Paid                        -3.626e+00  1.767e+01  -0.205  0.83756
Post.Weekday                 1.309e+01  1.491e+01   0.878  0.38092
Lifetime.Engaged.Users       1.106e-01  1.843e-02   6.005 7.26e-09 ***
comment                      2.581e+00  7.274e-01   3.547  0.00047 ***
share                        4.818e+00  4.628e-01  10.410  < 2e-16 ***
Category                    -4.739e+00  1.689e+01  -0.281  0.77927
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 114.7 on 234 degrees of freedom
Multiple R-squared:  0.7055,    Adjusted R-squared:  0.6929
F-statistic: 56.06 on 10 and 234 DF,  p-value: < 2.2e-16
```

Figure D

The resulting regression model as shown in figure D is appropriate in answering the question of interest as it includes the most significant predictors when predicting the number of likes a post will receive. This model showed that whether the post was a status, lifetime engaged users, lifetime post consumption, comments, and shares significantly influenced the number of likes a post received. The number of shares, comments, and lifetime engaged users a post received influenced the number of likes positively. On the other hand, status and lifetime post consumption had a significant negative relationship with the number of likes in the presence of all other variables above, but this relationship seems weak as the coefficients in the model are smaller.

The variable Paid provides insights into our question of interest because we wanted to find out if paid posts had a relationship with the number of likes the post received. We believed that a post being paid for or not could have a significant effect on the overall number of likes. For instance, a paid post might be advertised more to consumers thus increasing the probability that a consumer interacts and likes the paid post versus an unpaid post. In the end, we find that the paid variable is insignificant and does not have a significant relationship with likes.

Regression Tree

Recursive Binary Splitting

We decided to present the regression tree built with recursive binary splitting because the predictors and number of nodes did not differ significantly with the pruned tree. The pruned tree just included shares and comments and essentially only removed the two nodes relating to lifetime engaged users as seen in figure E below. The recursive tree resulted in a lower test MSE than the pruned tree which we speculate to be related to the fact that including lifetime engaged users as a predictor leads to more accurate predictions. Therefore, with not much difference between the graphical outputs of the trees, we decided to go with the method that resulted in the lowest test MSE which was the regression tree built with recursive binary splitting.

Graphical Output

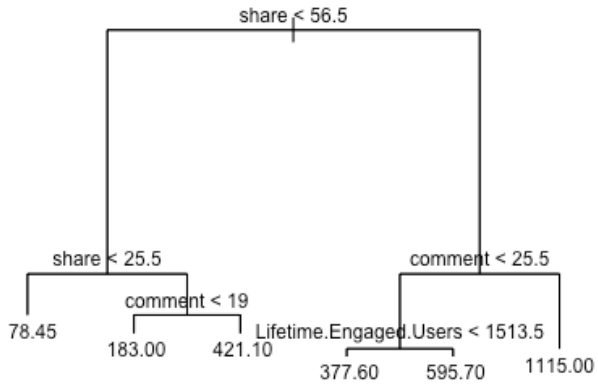


Figure E

The tree resulting from recursive binary splitting has 6 terminal nodes and uses 3 of the 8 predictors as seen in figure E. The three predictors used to predict the number of likes is shares, comments and lifetime engaged users. The resulting tree answers our question of interest because we have a better understanding of what variables are important in predicting likes. For this particular tree, the more shares (at least 56.5) and comments (at least 25.5) a post receives, the higher number of likes. Among posts that receive less shares (less than 56.5), shares and comments are important predictors. Specifically, if a post receives between 25.5 and 56.5 shares, the number of comments (less than or greater than 19) then determines how many likes the post will receive. However, if a post receives less than 25.5 shares, a post is predicted to have 78.45 likes. Lifetime engaged users factors in when the amount of shares is greater than 56.5 and amount of comments is less than 25.5. Overall, shares is the most important predictor in predicting likes for our recursive binary splitting tree-based model.

Random Forest Important Variables

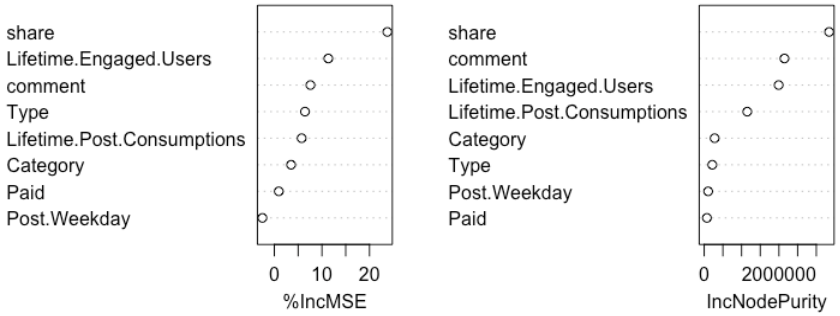


Figure F

	%IncMSE	IncNodePurity
Type	6.44	216831.42
Lifetime.Post.Consumptions	5.74	1152463.58
Paid	0.93	72350.15
Post.Weekday	-2.52	106281.28
Lifetime.Engaged.Users	11.34	1992218.84
comment	7.58	2147081.52
share	23.75	3343281.55
Category	3.52	279624.34

Figure G

Based on figures F and G, the most important predictors in predicting the number of likes is shares. This provides insight into our question of interest because we have a better understanding of which predictors are truly important when predicting likes. Figure G can provide some insight into why the recursive binary splitting tree resulted in a lower test MSE than the pruned tree since the percent that the variable lifetime engaged users increases MSE is the second highest after shares. Additionally, for both figures, paid is one of the least significant predictors of likes on a post which is interesting but not surprising based on the results of the regression model.

Summary

Test MSE Comparison

Type of Model	Linear Regression	Recursive Binary Splitting	Random Forest
Test MSE	21558.26	90466.93	82419.64

The test MSE was by far the lowest for our linear regression model. The highest test MSE resulted from our regression tree built with recursive binary splitting, then we saw some improvement with random forest producing a lower test MSE. We are unclear as to why the test MSE is much higher for the tree models and believe it could be due to the fact that our linear regression model assumptions are not fully met.

Overall, the results for the linear regression, recursive binary splitting and random forest models are fairly consistent. Shares and comments are two of the most important predictors in predicting the number of likes for all three models, while the paid variable was consistently insignificant. Our EDA is consistent with the results from linear regression especially when examining the positive or negative nature of the relationship between the predictors and the number of likes. Additionally, the variable lifetime engaged users ended up being fairly significant across all three models, showing a positive relationship with the number of likes. The biggest difference between the recursive tree model and the linear regression model is that the recursive tree model only showed three significant predictors- shares, comments and lifetime engaged users. The linear regression model included 2 additional significant predictors- lifetime post consumptions and status (one level of the type variable)- even though both of these variables had a negative relationship with the number of likes. When compared to the important variables outputted from

the random forest model, the results are expected as lifetime post consumptions and status (one level of the type variable) do not contribute to overall accuracy as much as shares.

Discussion on Question of Interest

In terms of answering our question of interest, the recursive binary splitting tree is better when compared to the other two models. We initially thought to choose our linear regression model since it resulted in the lowest test MSE, however, we think that the result might be misleading and not as accurate since our model had a violation in the constant-variance assumption. Because of this, we decided to select the recursive tree over the random forest model since we want to be able to interpret the most important predictors in predicting the number of likes. With random forest, we lose the ability to predict and interpret since no single tree is outputted despite the test MSE for random forest being the lower of the two. We realize that by selecting the recursive tree, we are selecting the model with the highest test MSE which leaves room for overfitting. However, based on the fact that our question is essentially a prediction problem, the recursive tree is the best choice based on the three models.

Classification Question

Our classification question aims to explore the relationship between various post and engagement metrics and whether a post is paid for. The specific question that we are trying to answer is: can we accurately predict if a Facebook post was paid for based on certain post characteristics and measures? We are interested in the categorical variable paid (1) or not paid (0). We explore this question through a logistic regression model and a classification tree model.

Exploratory Data Analysis

Response Variable Distribution

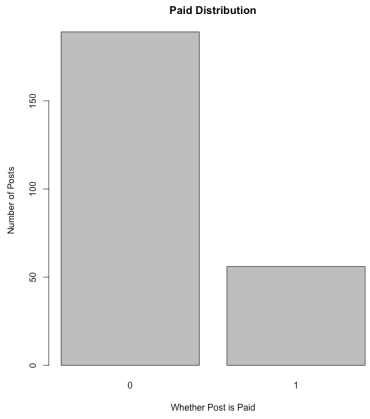


Figure H

Based on the barplot of the distribution of the categorical response variable, Paid, we note that there are a lot more posts that are not paid than paid. Intuitively, this makes sense as a company is more likely to post less advertised content. However, since we see a seemingly large difference in the number of paid and not paid posts, we might be dealing with unbalanced sample sizes between 2 classes. This might result in high accuracy but runs the risk of false negative rate or false positive rate being high. We would want to check the ROC curves of our model to verify our classifications.

Scatterplot

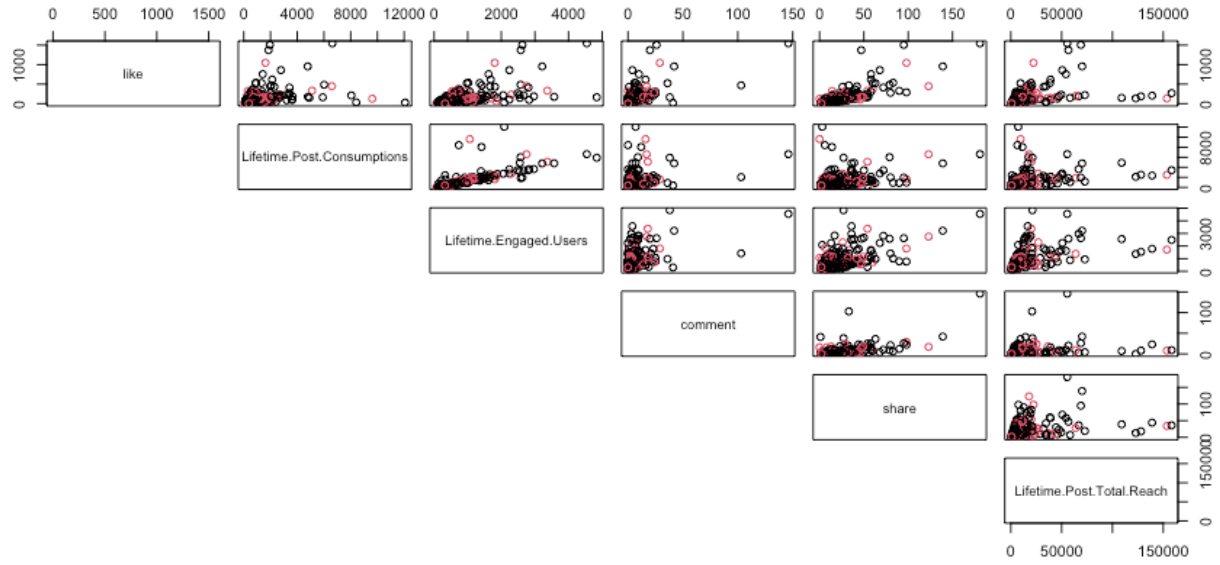
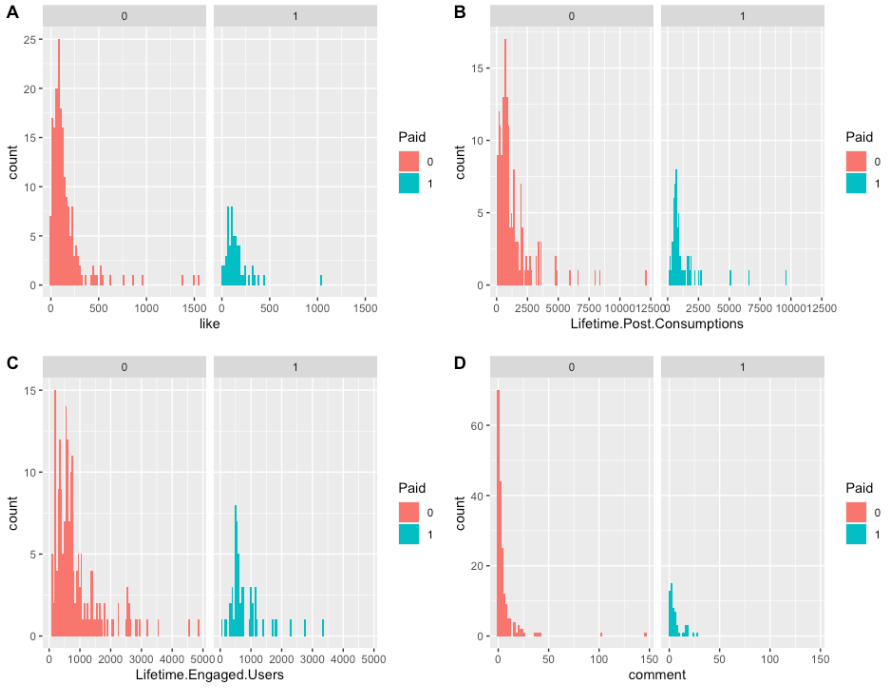


Figure I

The scatterplot matrix shows the relationship between predictor variables. Each point is in either black or red, which represents whether the post was paid or not paid. It is clear in all of the plots that there is a lot of overlap between black and red points. This means that paid and not paid posts are not well-separated at all. We believe that there will be difficulty in differentiating and classifying based on these predictor variables whether a certain post was paid.

Histogram grid for variables of interest and Paid variable



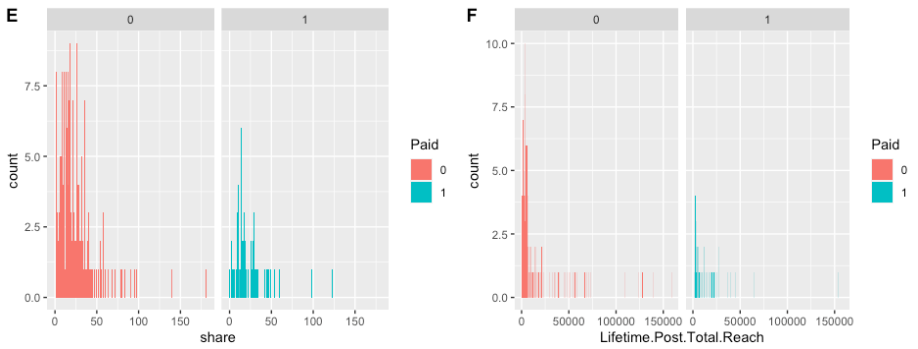


Figure J

Similar to the scatterplot matrix, the grid of histograms of the variables of interest also show no clear distinction between the distributions of paid and not paid. All the distributions are skewed to the right, indicating the potential for significant outliers. Based on the scatter plot and grid of histograms, classification between a paid advertisement and an unpaid one will be difficult given the number of likes, shares, comments, lifetime post consumptions, engaged users and post total reach. In both exploratory data analysis graphical summaries, it is evident that the predictor variables are not well separated depending on different categories of Paid.

Logistic Regression Model

The logistic regression model that our group is most satisfied with is our initial logistic regression model that does not include zero liked rows. This means that our classification question seeks to identify paid or not paid posts with a minimum amount of interaction — at least one like on the post. We choose the initial logistic regression model because by not including posts with zero likes, we are able to compare the logistic regression model to our classification tree analysis. Our classification tree analysis also only looks at posts with a minimum amount of likes.

Further, our "improved" regression model shows little improvement from our initial regression model. In an attempt to approve our initial logistic regression model, our second logistic regression model added categorical predictor variables to our model as well as zero like posts. While the second model performed slightly better, we are hesitant to conclude that the model is truly improved as interpretability might be more difficult and only one predictor is significant. We also choose a logistic regression model to compare to the classification tree analysis, because the linear discriminant analysis model does not meet the Multivariate Normality Assumption check as seen in the figure below.

Multivariate Normality Assumption Check

H0: distribution is consistent with Multivariate Normal
Ha: distribution is not consistent with Multivariate Normal

```
             Multivariate Normality Test Based on Skewness

        data:  notadvertised[, 2:7]
        U = 3863.9, df = 6, p-value < 2.2e-16

      Multivariate Normality Test Based on Kurtosis

data:  notadvertised[, 2:7]
W = 51916, w1 = 0.625, df1 = 20.000, w2 = 1.000, df2 = 1.000, p-value < 2.2e-16
```

Figure K

Since the p-values of the MVN test based on skewness and kurtosis are very small, we reject the null hypothesis. We do not have enough evidence to say that the distribution of our model is consistent with the Multivariate Normality assumption. Since the assumption is not met, we choose to use a logistic regression model to interpret our results. However, if we were to truly want to proceed with LDA and our main goal is to classify whether posts are paid or not, we can still proceed with LDA with caution. We proceed with interpreting our initial logistic regression model.

Summary of Logistic Regression Model

```
Call:
glm(formula = Paid ~ like + comment + share + Lifetime.Post.Consumptions +
    Lifetime.Engaged.Users + Lifetime.Post.Total.Reach, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8363  -0.7345  -0.7204  -0.6197   1.9032

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.118e+00  2.556e-01  -4.375 1.21e-05 ***
like                       -1.873e-05  1.380e-03  -0.014    0.989
comment                    -2.095e-04  1.623e-02  -0.013    0.990
share                       8.636e-04  1.170e-02   0.074    0.941
Lifetime.Post.Consumptions  4.902e-05  1.413e-04   0.347    0.729
Lifetime.Engaged.Users     -2.536e-04  4.064e-04  -0.624    0.533
Lifetime.Post.Total.Reach   2.654e-06  7.380e-06   0.360    0.719
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 263.40  on 244  degrees of freedom
Residual deviance: 262.84  on 238  degrees of freedom
AIC: 276.84

Number of Fisher Scoring iterations: 4
```

Figure L

From the Wald test p-values for each predictor, none of the predictors were significant in classifying if a post was paid for or not. When we ran the delta g-squared test on this model to test this model against an intercept-only model, our conclusion was supported. We failed to reject

the null hypothesis with a p-value of 0.997, indicating that the predictors were insignificant at the alpha level of 0.05. We expected this insignificant result as the initial EDA showed virtually no difference in any of the interested predictor variables in terms of their distributions for paid and not paid.

Multicollinearity Check

```
            like Lifetime.Post.Consumptions     Lifetime.Engaged.Users
        3.156194                    2.069797                   3.382742
         comment                       share Lifetime.Post.Total.Reach
        2.278992                    3.278659                   1.367804
```

Figure M

Since all of our predictors in our logistic regression model turned out to be insignificant in relation to the Paid response variable, we wanted to check for multicollinearity. All variable VIF values are lower than 10, indicating that multicollinearity is not a concern. We can rule out that multicollinearity is the reason why all of the predictor variables for our initial logistic regression model are insignificant.

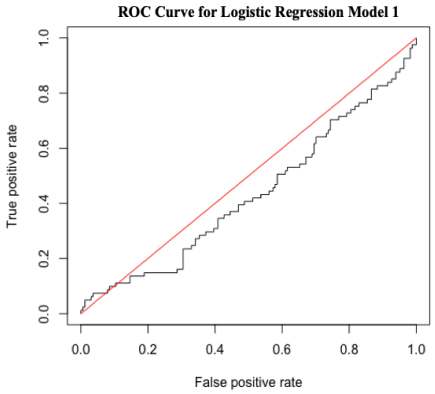


Figure N

The ROC curve for the logistic regression model is roughly along the random assignment line. The curve for the logistic regression model hovers mostly below the red line indicating that the model is performing worse than random guessing. Ideally, we would like the ROC curves to be curved towards the northwest quadrant of the plot. However, our models are performing worse than random guessing.

The logistic regression model does not perform well. In future sections, we compare the error rates of this model to other models of our classification question. We also seek to explain reasons why our logistic regression does not perform well.

Classification Tree

The classification tree we are most satisfied with is the tree from the recursive binary model. We chose the tree from recursive binary splitting instead of the tree from pruning because the pruned tree resulted in only one terminal node. It was not able to be plotted and the entire predictor space was not split. If we could only visually present one of the two trees, we would have very little choice and would have to go with the recursive binary tree for the data set that we have on

hand. Furthermore, the tree from recursive binary splitting also gives a lower overall error rate compared to the pruned tree. The only weak benefit of the pruned tree is that it has a 0 false positive rate, but it comes at the expense of the false negative rate being 1. A disadvantage of the recursive binary tree is that it is very large and complex, making it hard to interpret.

	Overall Test Error Rate	False Positive Rate	False Negative Rate
Recursive Tree	0.3510	0.1220	0.8148
Pruned Tree	0.4939	0	1

Recursive Binary Tree Model

The classification tree created from recursive binary splitting has 24 terminal nodes. The resulting tree uses 7 of the 8 predictors that we used to run the model. The predictors that are used in the tree are Lifetime Post Total Reach, likes, shares, Lifetime Engaged Users, Lifetime Post Consumption, comments, and Post Weekday.

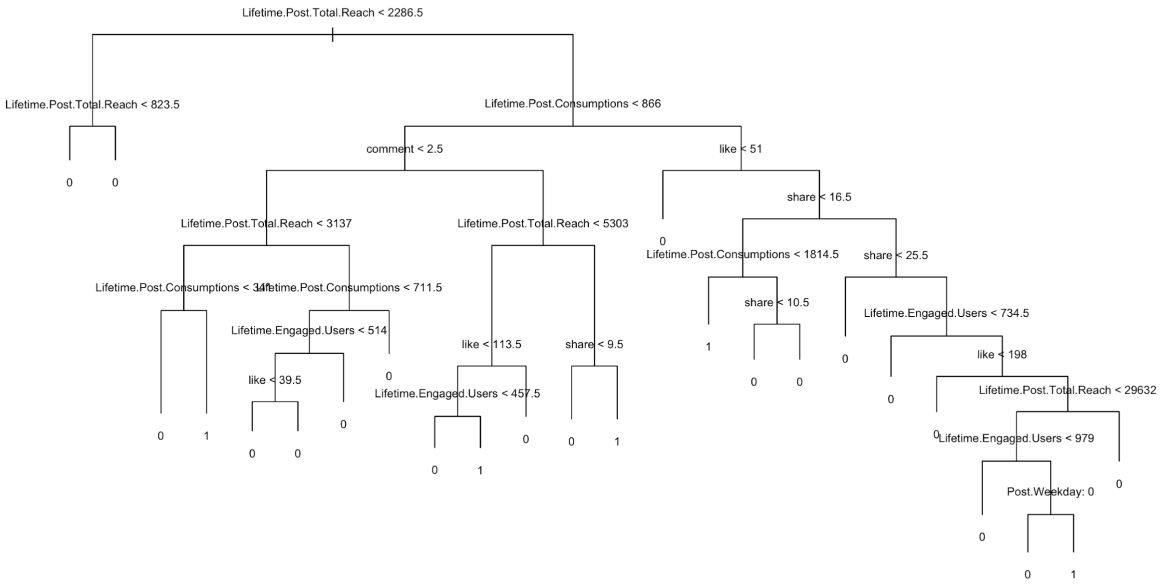


Figure O

The output of the classification tree is fairly hard to interpret. There are many nodes and terminal nodes, making the graphical output messy and very difficult to interpret. The most prominent variable in classifying each post is Lifetime Post Total Reach. Based on the classification tree, we can note that if Lifetime Post Total Reach is less than 2286.5, each post is classified as not paid for advertising. When Lifetime Post Total Reach is greater than 2286.5, the classification becomes a lot more complicated. An example of a post being classified as being a paid advertisement are posts with Lifetime Post Total Reach greater than 2286.5, Lifetime Post Total

Consumption less than 886, comments less than 2.5, Lifetime Post Total Reach not being more than 3137, and Lifetime Post Consumption not more than 341. The interpretation of one node is complicated, so understanding the whole tree will take a lot of time.

The tree created from recursive binary splitting answers our classification question of interest by saying which values of different variables will result in paid or not paid advertisements. While the tree is fairly large and complicated, if one was truly interested in determining whether a post was paid or not paid, they could use that post's specific characteristics to see which terminal node the post belongs to.

Random Forest

Next, we use random forest as a method to decrease the variance of the predictions. By using random forests, we hope to form trees by bootstrap samples that are less correlated with each other to reduce the variance of decision trees. Since our classification problem has 8 predictors in the model, we use a random selection of 2 predictors as split candidates for each tree. We determine that we should use a random selection of 2 predictors by taking the square root of the original number of predictors.

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
like	9.175401	-6.80061461	6.222838	13.964182
Lifetime.Post.Consumptions	8.377398	-6.47244713	5.410921	14.432469
Lifetime.Engaged.Users	14.303375	-8.29713082	10.995613	14.670617
comment	5.947286	-4.29333676	3.566665	9.343394
share	11.041307	-8.99847636	6.649684	12.883387
Lifetime.Post.Total.Reach	8.090665	-3.56511640	6.348291	15.602549
Post.Weekday	-1.740815	0.02591907	-1.672943	1.949131
Type	-1.260197	-0.45481169	-1.482160	2.104430

Figure P

The larger the mean decrease accuracy and the mean decrease Gini index, the more important the variables are. We note that the relative most important variables in this model are Lifetime Post Total Reach, Lifetime Engaged Users, Lifetime Post Consumption, like, share, and comment.

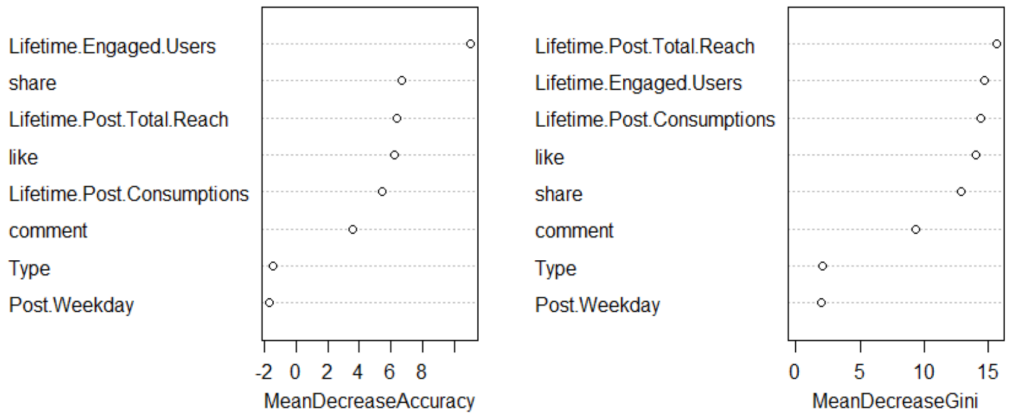


Figure Q

The figure above gives a graphical representation on what the most important predictors in the model are. The results from the important function output as well as the graphical output are fairly consistent with what we see in the recursive binary splitting model summary output in terms of variables that are important to the model.

The results of the classification tree models tell us which variables are most important in predicting whether a post was paid for or not. In the next section, we will compare all of our classification models to weigh on which analysis best answers our question of interest.

Summary
Confusion Matrices

Logistic Regression Model	Recursive Binary Splitting	Random Forests
FALSE 0 164 1 81	tree.pred.test pred.test 0 1 0 144 20 1 66 15	yhat test.y 0 1 0 161 3 1 80 1

All of the confusion matrices in the table above use a 0.5 threshold level. We use actual data in the rows, and predicted classes in the columns of the matrices.

	Overall Test Error Rate	False Positive Rate	False Negative Rate
Logistic Regression	0.3306	0	1
Recursive Tree	0.3510	0.1220	0.8148
Random Forest	0.3387	0.0122	0.3265

For our logistic regression model, at a 0.5 threshold level, we get a false negative rate of 1. Our original logistic regression model incorrectly classifies advertisements that are actually paid 100% of the time. This suggests that our model is very poor in classifying posts that are paid based on the predictors that we are using. The overall error rate at a 0.5 threshold is 0.3306. This error rate does not seem too bad, but it is important to take into account the very high false negative rate.

For the recursive binary splitting model, we have an overall error rate of 0.3510. The false positive rate and the false negative rate are 0.1220 and 0.8148, respectively. Compared to the logistic regression model error rates with a 0.5 threshold, the balance in error rates for recursive binary splitting is better. However, it is still not too ideal in that the false negative rate is fairly high at 0.8148. This means that 81.49% of the time while a post is actually paid for, it is predicted as not paid for.

Compared to the recursive binary tree model, the random forest model has improved in overall test error rate. This makes sense intuitively since random forest seeks to decrease variance by creating trees from each bootstrap sample. The false positive rate and false negative rate are also improved in comparison to the tree produced in recursive binary splitting. However, a downside to random forest is the lack of visualization and interpretation of the model, since the model is generated through the bootstrapping of trees.

Discussion on Thresholds

In the context of our classification problem, changing the threshold would not benefit or help the error rates. We are trying to determine the error rates of being classified as a paid post or not, so it is not too alarming like it would be in a medical setting. If we were to increase the threshold for both models, it would be harder to classify as paid, so values inside the actual paid column would decrease in both rows. As a result, the false negative rate would increase and the false positive rate would further decrease in the recursive tree. Similarly to the error rates in recursive binary splitting, changing the threshold in this scenario would be limited in terms of the context of Facebook metrics data. We are mainly interested in how well the model can classify. In the case of social media engagements, it is not intuitively important to decrease the false negative rate of false positive rate that would be important in another setting like drug testing. Overall, there would be no real benefit in increasing the threshold. If we were to decrease the threshold, we might better balance the false positive rate and false negative rate from the recursive tree.

Logistic Regression Model (0.3 Threshold)

	FALSE	TRUE
0	163	1
1	79	2

When testing if lowering our threshold levels would be beneficial to the model, the logistic regression model with a lowered threshold at 0.3 still classifies poorly. At the 0.3 threshold level for logistic regression, we get a false negative rate of 0.9875 and a false positive rate of 0.0061. As expected our false negative rate goes down as we decrease the threshold, but it still remains very high. Our model still incorrectly classifies paid advertisements 98.75% of the time.

Overall, the classification model classifies poorly. There is a lack of predictors in the model that have a significant relationship in classifying whether a post is paid or not paid.

Discussion on Findings from Models

The logistic regression model and the classification tree models all reach a general consensus that the predictors in the data set perform poorly in predicting whether a post is paid or not paid. The logistic regression model with a minimum amount of likes per post performs very poorly and even worse than random guessing.

Differences in our findings from the logistic regression model and classification tree model stem from the differences of each statistical learning method. As expected, depending on which method is used, the interpretations and results differ. For example, with the logistic regression model, we were able to see the summary of the levels of significance for each predictor in relationship to paid. From there, we can perform ANOVA tests and other analysis to try to improve the model. With recursive binary splitting we were at least able to produce an interpretable tree and relationship to predict whether a post was paid. While the tree was large and hard to interpret, there was a producible visualization. From random forests, we were able to lower the error rates and see which were the most important predictors. The results were slightly different from each learning method used. However, overall the results were consistent in that our classification models performed poorly.

Discussion on Question of Interest

Comparing three different models, we conclude that the random forests model is best in interpreting our question of interest. Given our logistic regression model found no predictor significant in classifying if a post was paid for or not, the model is not contributing to interpretation of our question of interest. From our random forests model, we were able to determine the most important predictors in predicting whether a post from a cosmetic company's Facebook page was paid for advertising or not paid for advertising. From our recursive binary splitting tree model, we were able to note that the variable Lifetime Post Total Reach was the top split in determining whether a post was paid or not. This is in line with the important variables output from the random forest model. Lifetime Post Total Reach in the random forest model is also one of the most important random predictors.

Overall, both random forests model and the recursive binary tree models have performed well in providing insights into predictors useful in predicting paid posts. As the random forests model has the lowest overall error rates and false negative rates, the performance of the random forests model is the best out of all three models. Even though we could not plot out the random forests model, the model is providing enough insights into our question of interest and the degree of interpretability lost was not significant enough to impede its interpretation. We have thus concluded that the random forests model is better for interpretation, when compared to the recursive tree model and logistic regression model.

Further Work

Through various analysis of a certain cosmetic company's Facebook page data set, we were able to draw conclusions through regression and classification analysis on the relationship between various post characteristics and engagement metrics. It is important to note that while our analysis is thorough in terms of statistical learning methods used, the nature of the data set might be limited. Our data set comes from posts' published during the year of 2014 on the Facebook page of a renowned cosmetics brand. Since we are analyzing data from a specific cosmetic brand, there might be underlying reasons as to why these metrics do not significantly predict Paid. When using a data set with limited background, we don't want to extrapolate the data and draw too general conclusions for all social media usages. If we could include a broader data set extracted from a less biased database such as Facebook and Twitter, we could reduce the magnitude of background bias in our data set. Thus, if we have more time to do this project, we would want to include a more general data set.

If we have more resources, we could also incorporate more engagement metrics for predictors. Intuitively, other predictors that could predict Paid better might be measures of engagement such as clicks. More clicks on a certain post might mean consumers are more engaged with the post due to the advertisement nature of the post. On the other hand, consumers might have insight of which posts are paid, and might be less likely to interact with these paid posts. If the end goal of the cosmetic company were to increase sales, we can also track sales coming from a link that is from the specific post. Furthermore, we gained insight on how different engagement metrics that were not previously thought about could better predict paid posts from our models. Therefore incorporation of other available engagement metrics into our model could also contribute to better performance in models and the answering of our questions of interest.