****

**BUDT758T  
  
DATA MINING AND PREDICTIVE ANALYTICS**

**Project: Predicting the Performance Evolution of Facebook Posts**

**Mihir Gandhi**

**Rini Joshi**

**Spoorthi Gunuganti**

**Trisha Chandra**

1. **Executive Summary:**

Social media is on its way to become the most important channel for brand recognition. 85% of marketers using social media are unsure of how to use it. Thus, our project revolves around using various data mining techniques to predict the performance metrics of the posts published on Facebook. We analyzed how each of the seven input features (Page total likes, Type, Category, Post Month, Post Weekday, Post Hour, Paid) affect the impact of the post and we found out that performance variables comment, share, and like had the most influence by the given input features.

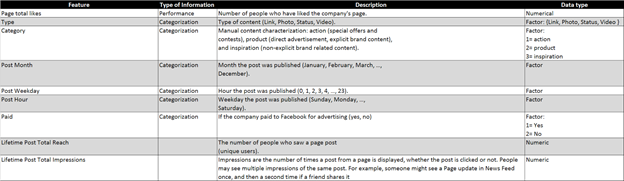
1. **Data source:**

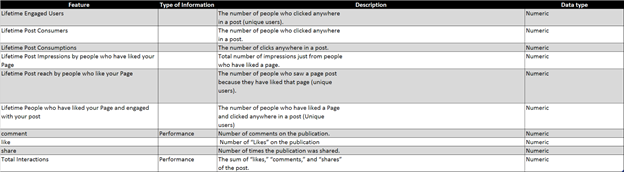
Our data set was sourced from “UCI Machine Learning Repository”, the UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. Below is the hyperlink to the dataset:

<https://archive.ics.uci.edu/ml/datasets/Facebook+metrics>

1. **Dataset Description:**

The data set consists of performance metrics of posts published on a cosmetic brand’s Facebook page in the year 2014. We split the dataset in 70 (training):30 (testing) ratio. The data set consists of 19 columns described as below:

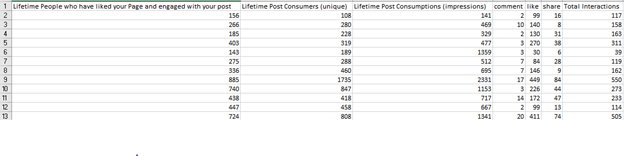
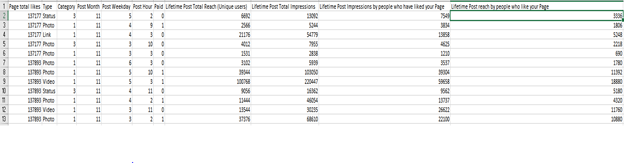




**Sample size:** n = 500 records

* 1. **Independent variables:** The total number of independent variables are 7 listed as follows:
* Page total likes
* Type
* Category
* Post Month
* Post Weekday
* Post Hour
* Paid
  1. **Dependent variables:** The total number of dependent variables are 12 dependent variables mentioned as follow:
* Lifetime post total reach
* Lifetime post total impressions
* Lifetime engaged users
* Lifetime post consumers
* Lifetime post consumptions
* Lifetime post impressions by people who liked Facebook page
* Lifetime post reach by people who liked Facebook page (unique users)
* Lifetime people who liked page and engaged with the post
* Comment
* Like
* Share
* Total Interactions (Sum of comment, like, and share)

* 1. **Sample Observations:**

1. Below is the snapshot of some 13 observations from the dataset.
2. The first entry in the first snapshot shows that the page had 137177 likes when the page was posted. The post is of type “status”, classified under “inspiration” category. It was posted on a “Thursday (5)” in the month of “November (11)” at 02:00 hours. It is an “unpaid” post (0). 6692 unique users saw the post. The post has 13092 impressions out of which 7549 impressions (number of time people saw the page whether clicked or not) are from people who liked the page. Out of those 7549 impressions, 3336 were unique. Similarly, the next snapshot shows entries of other 7 dependent variables. 
3. **Why is this data of interest?**

The dataset gives an insight on how the performance variables (like, comment, lifetime total reach, etc.) varies with respect to the independent variables mentioned above. Our main aim is to implement models that predicts the performance variables which is nothing but the impact of the posts. Using this prediction, manager can decide whether to publish a post or not.

1. **Methodology**

We explored the following regression based models to model the 12 performance metrics individually.

1. Linear Regression
2. k-Nearest Neighbors
3. Regression Trees
4. Support Vector Machines (SVM)

To find which performance metrics were being predicted well across all the models, we used the following error measures to compare the results of each of the 12 models for each modeling technique.

**RMSE:** root-mean-square error

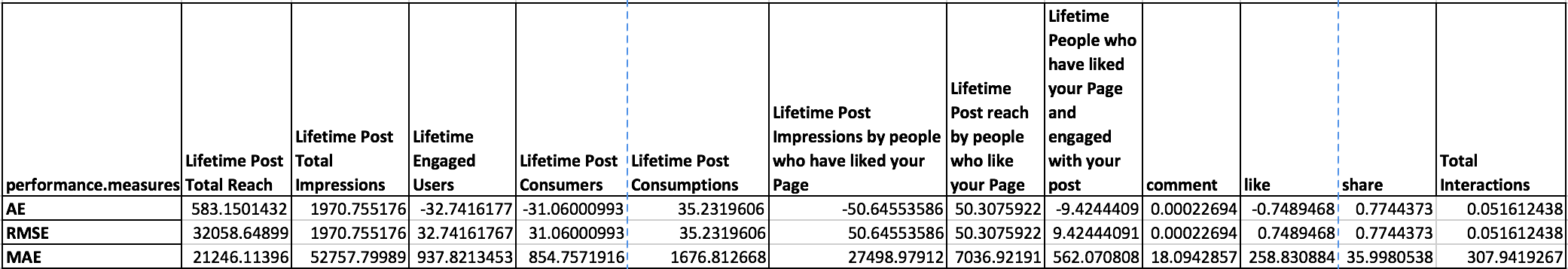
**AE:** Average Error

**MAE:** mean absolute error

Once we were able to narrow down on whicthe specific Tableau was used to perform exploratory analysis

* 1. **Linear Regression**

We performed Linear regression as a straightforward approach to train and consequently predict the quantitative response for our 12 dependent variables. Given below are the performance evaluation metrics for them:

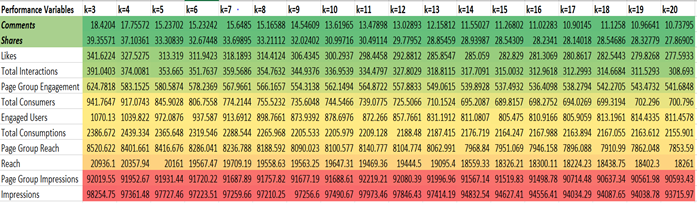


Total Interactions (a composite dependent variable; summation of number of comments, likes and shares) has the lowest RMSE.

* 1. **k- Nearest Neighbours**

The next in line of the regression models is the k-nearest neighbours. Since it uses a distance driven algorithm, the data was normalized before modeling. Below are the results of the

Model using 3-20 nearest neighbours. We can see that the error rate of the model kept decreasing as it included more and more neighbours into its algorithm making k-20 with the best results. We also observe that the error in prediction is lowest for the variables comments, shares and likes which collectively form the total interactions.



* 1. **Regression Tree**

We ran 12 regression trees for each of the 12 dependent variables against the 7 independent variables (See Appendix).

For all the models we calculated the performance measures which can be seen below:

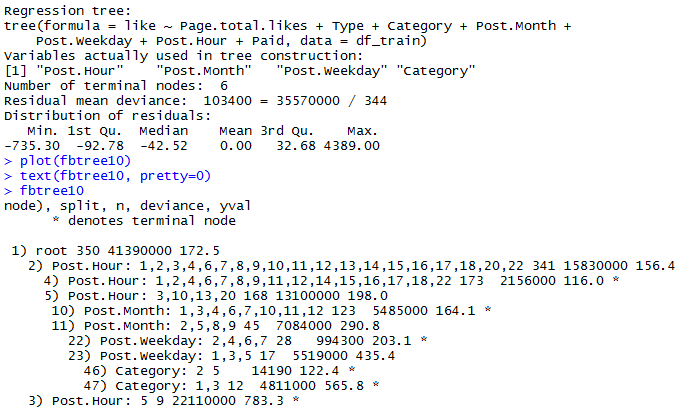


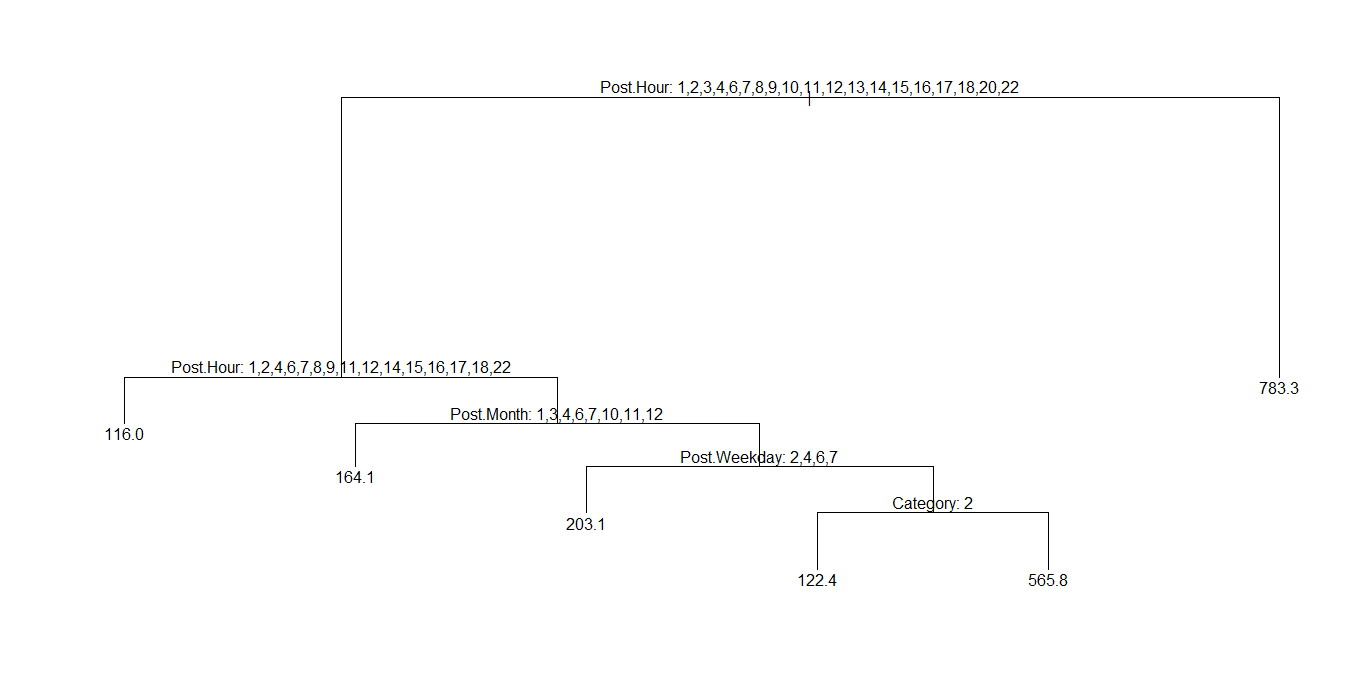
Models with dependent variables ‘Comment’, ‘Like’, ‘Share’, ‘Total Interactions’ tend to perform better than the other models. This led us to decide upon the most relevant dependent variables which are namely ‘Comment’, ‘Like’, ‘Share’, ‘Total Interactions’.

Considering the relative relevance and importance of these dependent variables, we delve further into explaining these four regression trees:

1. Dependent variable: Like

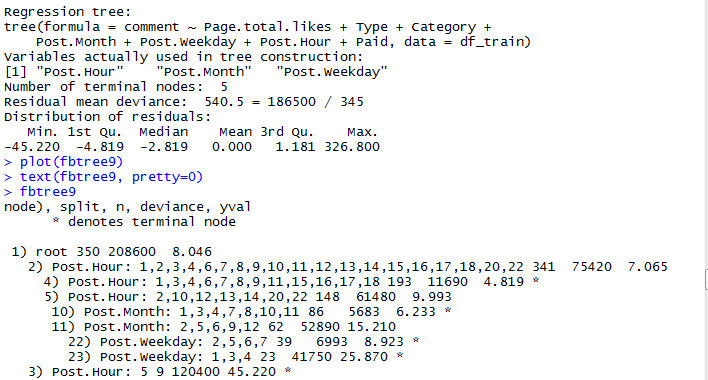
Here we see that if post hour is 5,19,21,23 the post will get an average of 783 likes. For post hours 3,5,10,13,19,20,21,23, Post months 2,5,8,9, weekdays 1,2,5 and Category 1,3 there will be an average of 566 likes

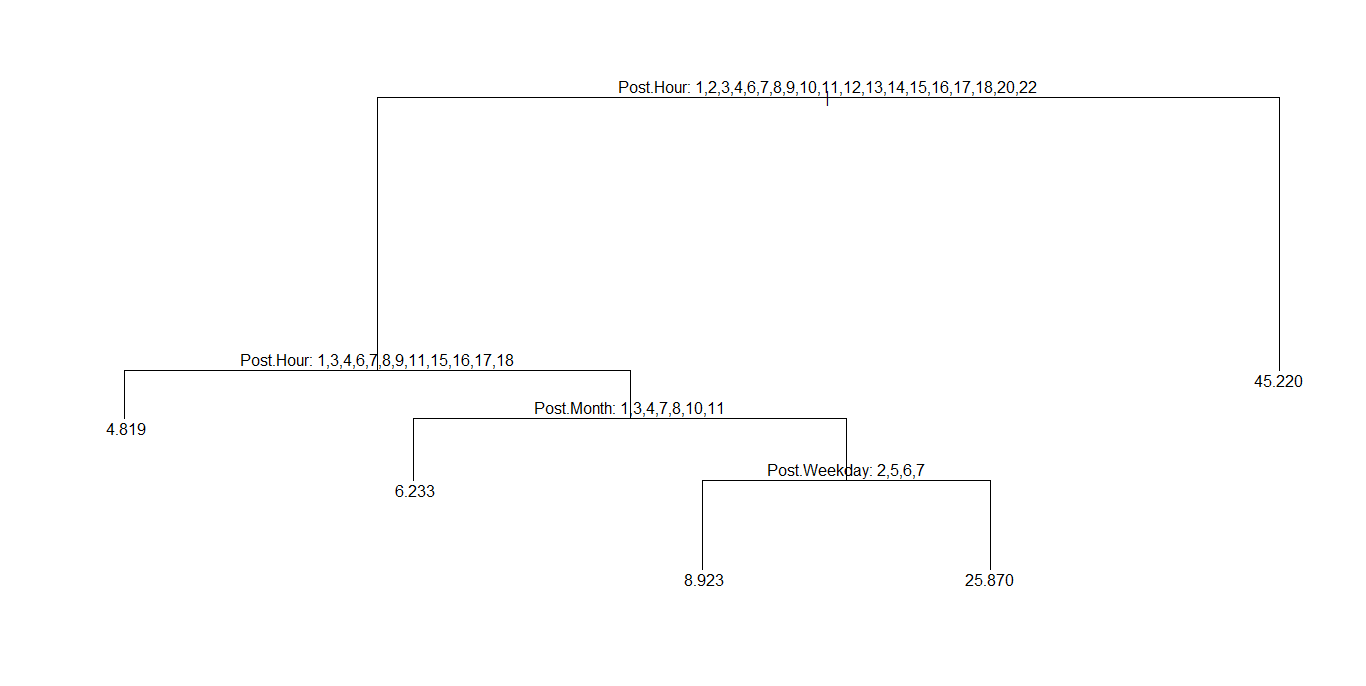




2. Dependent variable: Comment

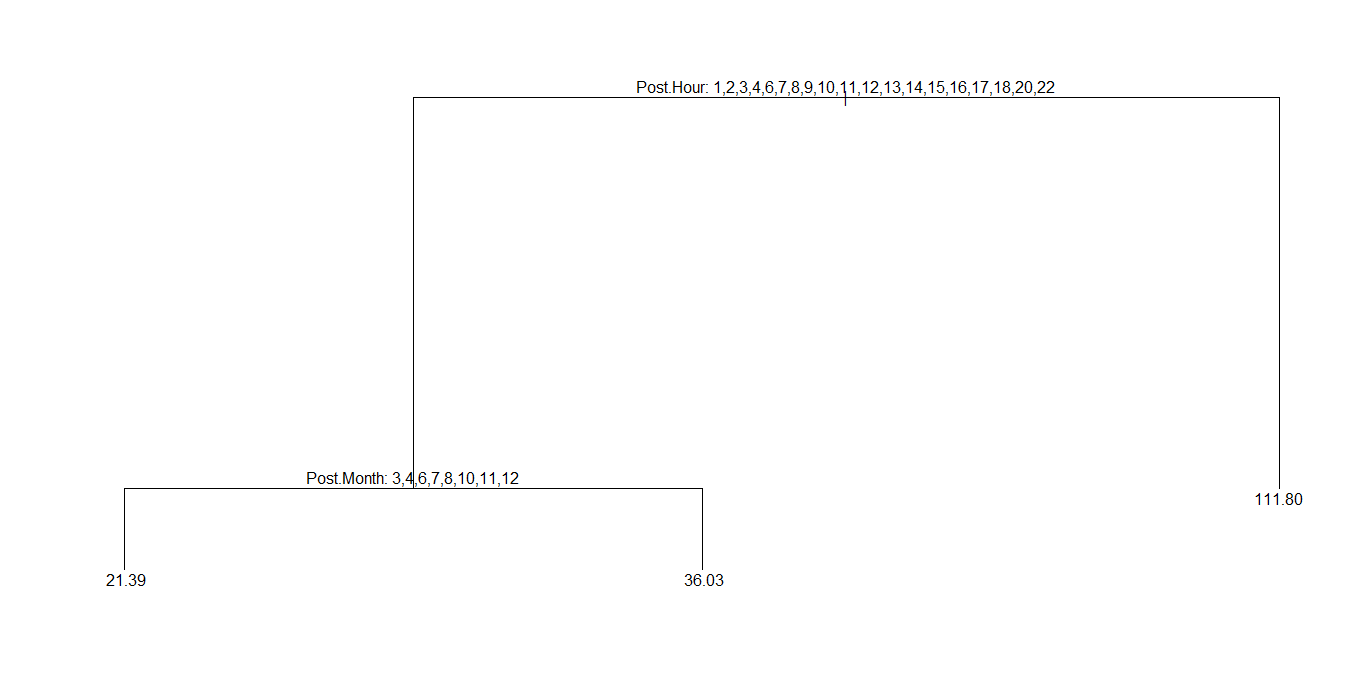
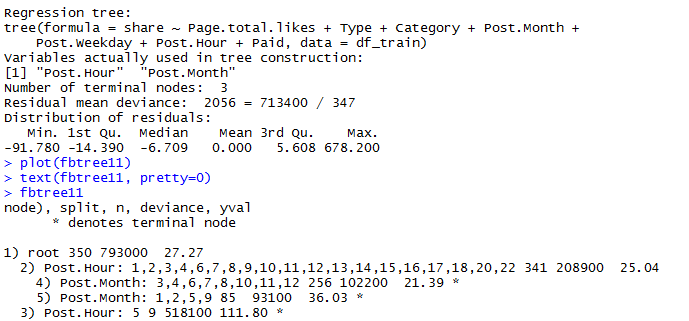
For hours 5,19,21,23 the post will get an average of 45 Comments. For hours 2,5,10,12,13,14,19,20,21,22,23, months 2,5,6,12 and weekdays 1,3,4 there will be an average of 26 comments





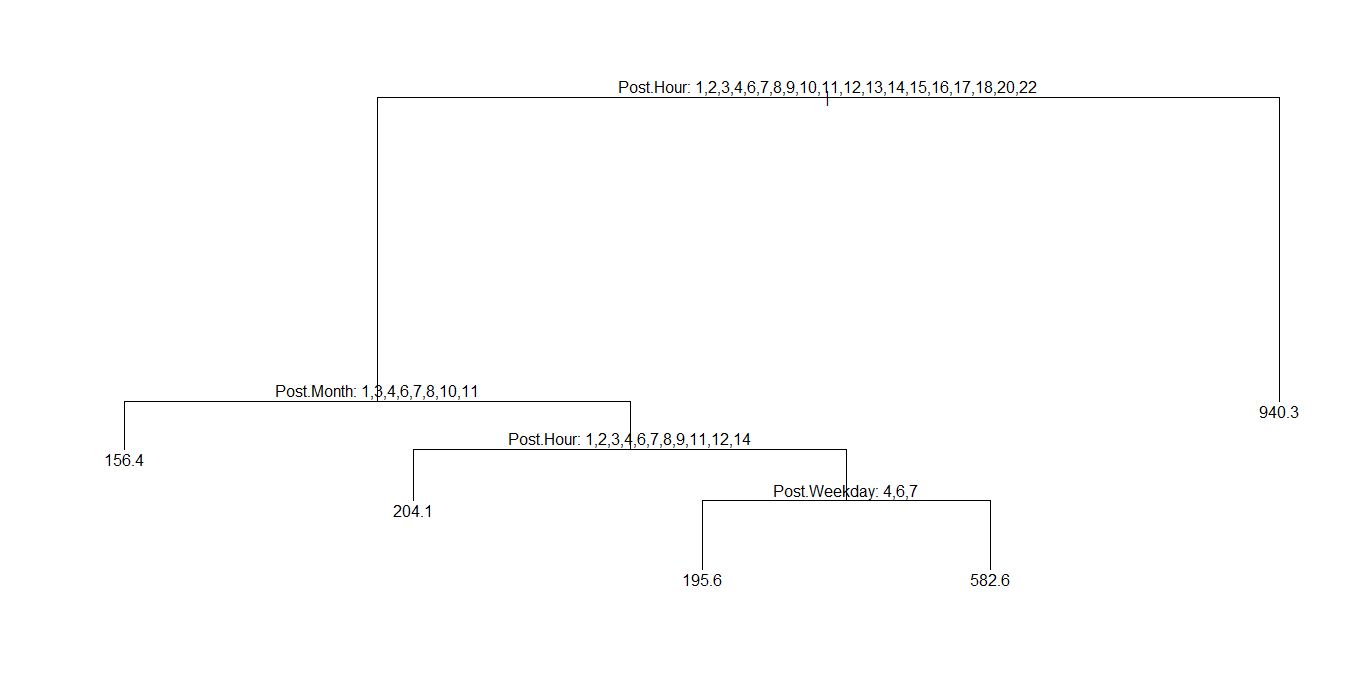
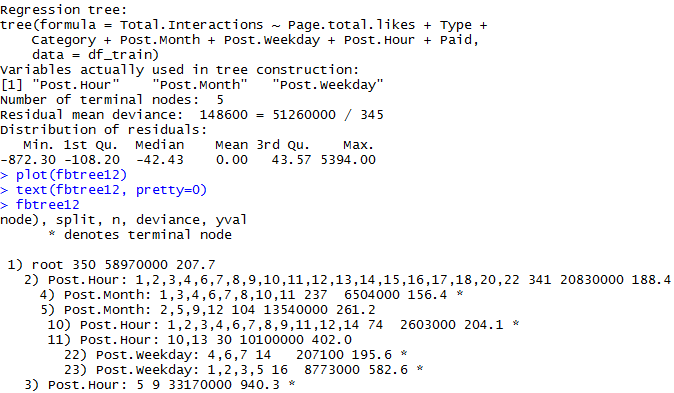
3. Dependent variable: Share

For hours 5,19,21,23 there will be an average of 112 shares, for hours 1,2,3,4,6,7,8,9,10,11,12,12,14,15,16,17,18,20,22 and months 1,2,5,9 there will be an average of 36 shares on the post



4. Dependent variable: Total interactions

For hours 19,21,23 there will be an average of 940 interactions

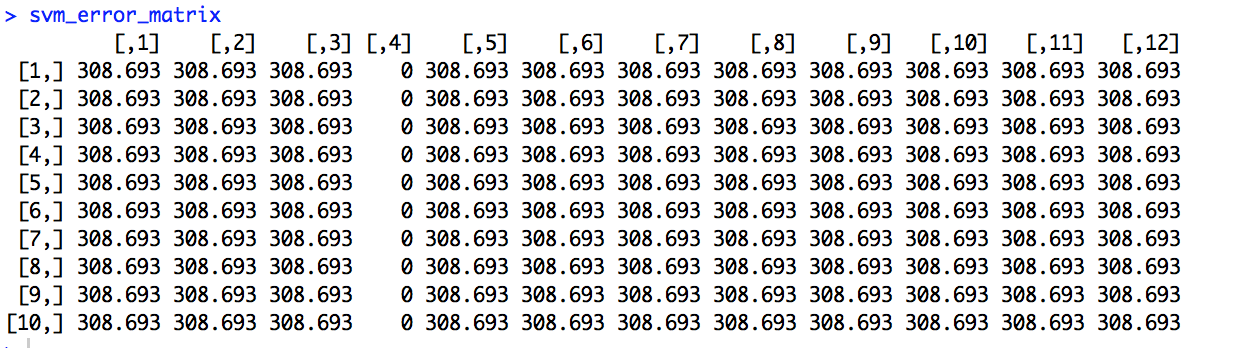


Taking a cumulative view at our regression trees we conclude that best hours to post to are 5,19,21 and 23, the best months are Feb, May, August and September and the best days are Sunday, Tuesday, Wednesday and Friday. This is to a great extend in coherence with our other models and exploratory analysis that we performed on tableau. Based on these statistics a manager can decide which hour, month and day of the week will likely be most suitable in terms of likes, comments, shares and total interactions.

* 1. **Support Vector Machines (SVM)**

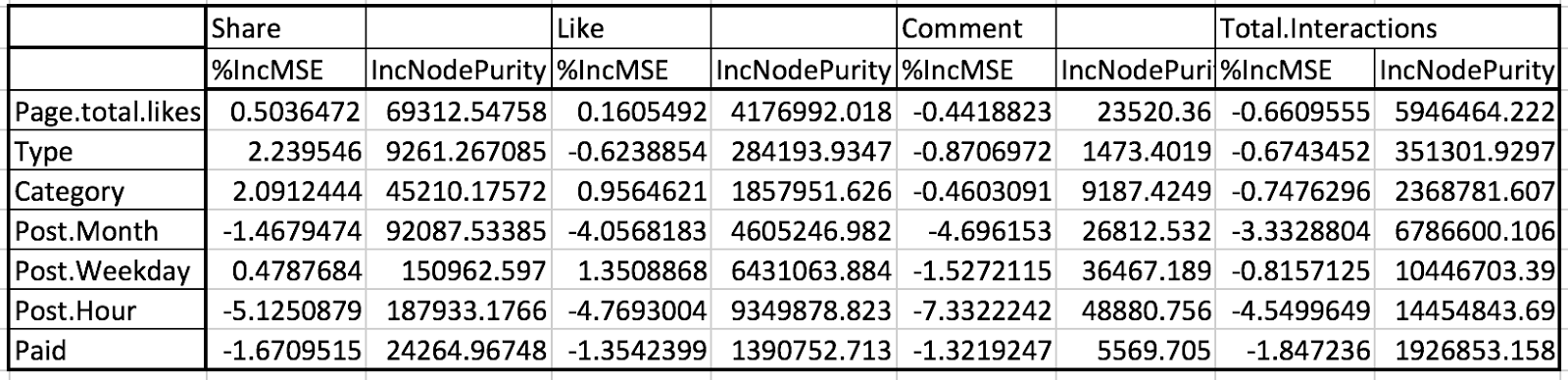
This regression was performed to enlarge the feature space in specific way (i.e penalizing by applying a cost of 100 for every 100th step in the range 100:1000)

The columns below represent the dependent variables. The error values are constant throughout the models and therefore we decided to not include this model in our analysis.

****

1. **Random Forest**

We implemented Random Forests which aggregates many decision trees to substantially improve the predictive performance and also to determine which features add considerable value to the output.  The first column ‘%IncMSE’ is based upon the mean decrease of accuracy in predictions on the out of bag samples when a given variable is excluded from the model. The latter column ‘IncNodePurity’ is the measure of total decrease in node impurity that results from each split.  All these results point towards ‘type’ being the greatest influencer.



1. **Business Perspective**

There is a saying used by many, often attributed to entrepreneur George Bradt, that states, "*People don't buy products, they buy brands*,". The phenomenon of social media has taken the world by storm and now online market presence has become a necessity.

If you’re doing it right, social media will lead to real relationship building; with already existing customers and also turn prospective consumers into buyers.

**But what defines right?**

According to one report by Social Media Examiner, 85 percent of marketers who use social media aren’t clear on which social media tools would work best for their business. While our analysis does not determine the content of the post, it illustrates the type and timing of the facebook posts that users are most receptive towards. Essentially, prediction of evolution of a post allows social media managers of the company to decide the ideal medium and timing for the post to garner maximum attention. Attention is determined through the number of interactions and engagements. The impressions statistics are derived from the number of times the post was loaded onto to the user’s browser. And engagements are variables that define actual user interaction feedback; such as comments, likes, shares, etc. While the post maybe available on the viewer’s homepage, it is highly likely the user scrolled through without noticing the post. Hence, we can say engagements are real determinants of the user’s interest in the post.

Amongst our 12 models developed for each dependent variables, prediction accuracy was highest for shares, followed by comments and likes. The results of these models are of maximum use as they will help forecast the popularity of each post.

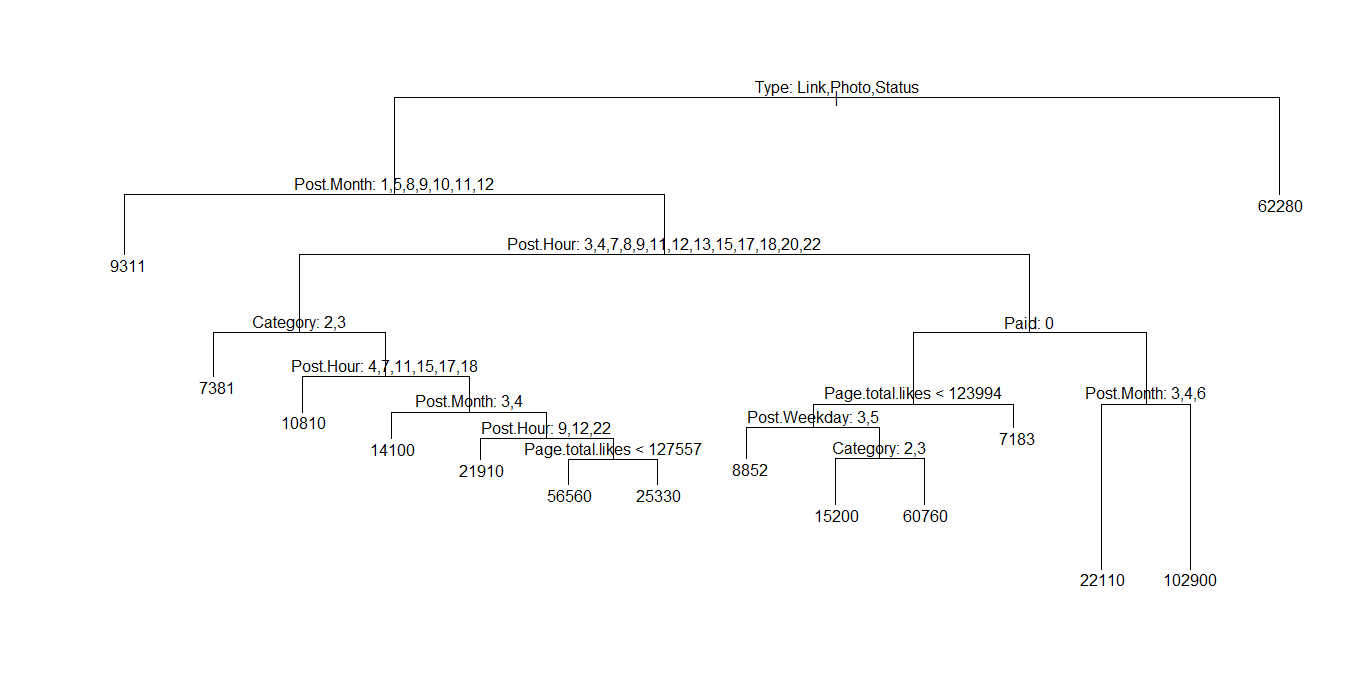
As per our analysis, the main decisive factor was the type of post, namely status, photo and video); with photo taking the first slot. Category was next in line amongst the causal factors followed by timing of the post. The prioritizing of these influencers helps social media marketing analysts to post so as to maximize impact.

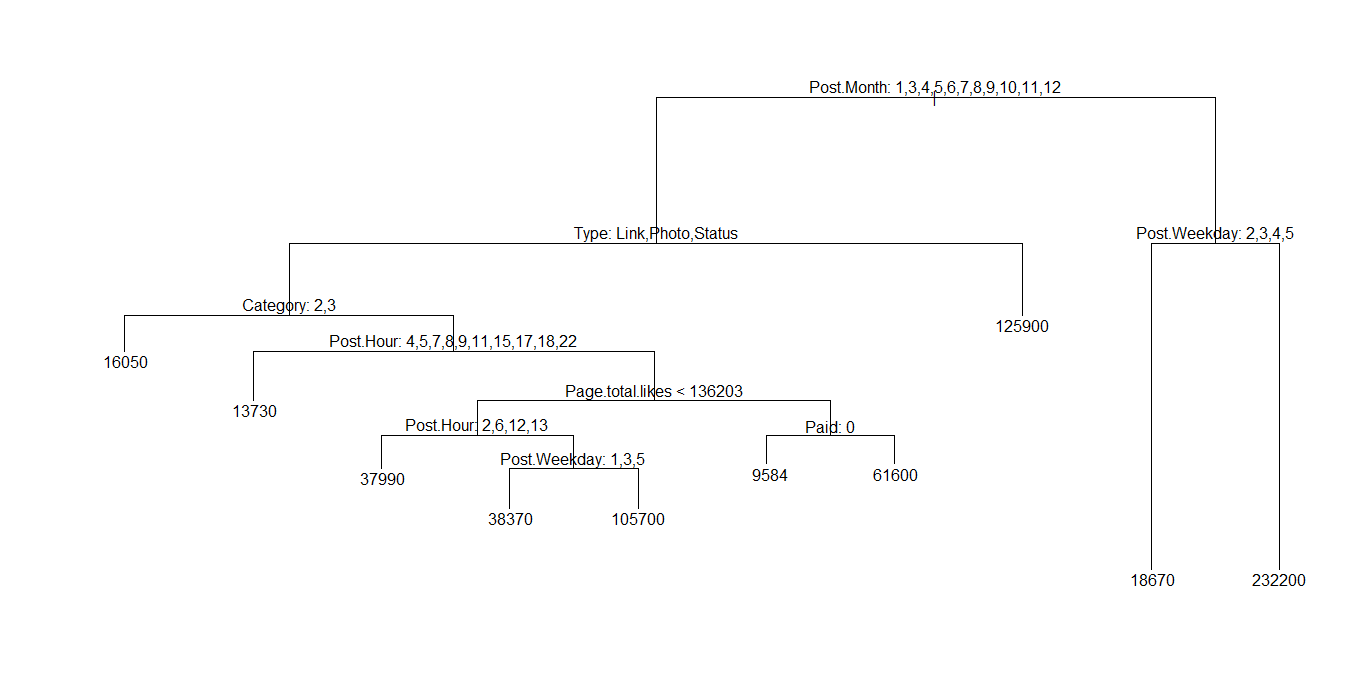
1. **Future Recommendations**

This research project primarily focused on modeling performance metrics (dependant variables) extracted from facebook posts of a company’s Facebook page using regression models through which we were able to identify the most important performance metric to monitor using some key features identified. However, this research is not the end as this does not provide an end-to-end predictive model that a company can use to evaluate its social media performance. With this project as a baseline, several ideas arise for future research.

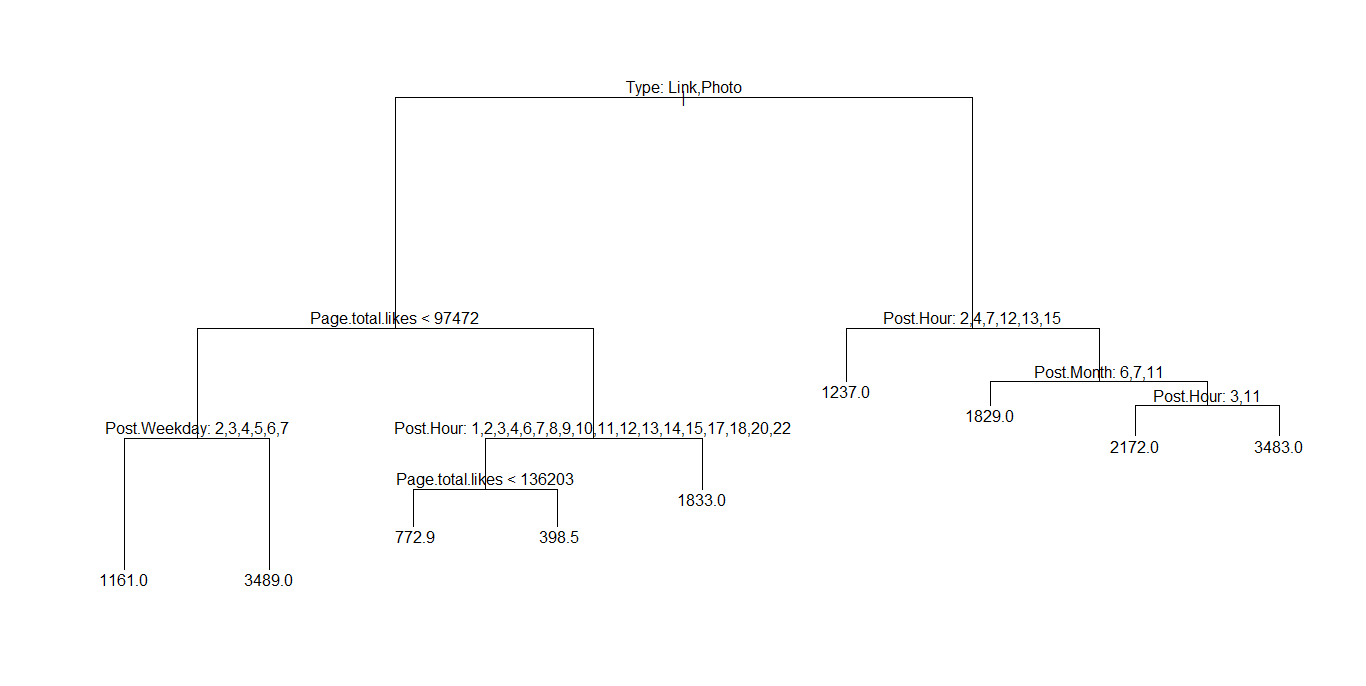
1. ***Refine Model:*** Our project is no exception to the saying that, no model is ever perfect and there is always some scope for improvement. While our project was based on the research done in (*S.Moro et all,2016),* this project produced fresh results with some different insights as compared to it. Hence, using our results as the baseline, the models used can be refined further to reduce the RMSE (using boosting) or experiment with different models (using neural networks). This can help build a more robust predictive model for total interactions.
2. ***Move up the conversion funnel:*** With a certain level of confidence in the predictions of total interactions, we can now use its correlation with various dependent variables to predict them by converting total interactions as an independent variable of the dataset. While the number of interactions represent potential conversions on the company website, we want to also be able to predict metrics from the top of the funnel as they reflect on the overall awareness of the brand. Predicting those metrics would help companies digital marketing strategies to improve on their post impressions reach which in turn can increase the number of total interactions of the post.
3. ***Text Mining*:** As the scope of this project, the dataset we worked on for this project was already aggregated and did not contain the actual text of the posts. However, it would be an interesting area for future research to mine the topics of the facebook using text mining models like LDA posts to create a predictive model for the topics to include in a post. This would help companies focus on creating posts that are guaranteed to attract larger attention on social media.
4. ***Revenue Model:*** A new tangent for research would be to create a predictive model for generating the Return on Investment of a post before it is actually posted. This however would involve developing a revenue model based on some existing/past data which we currently do not possess. A highly conceptual model with based on market statistics would be a long shot, but would make the research quite extensive and strenuous.
5. **Appendix:**

**Regression Trees:**

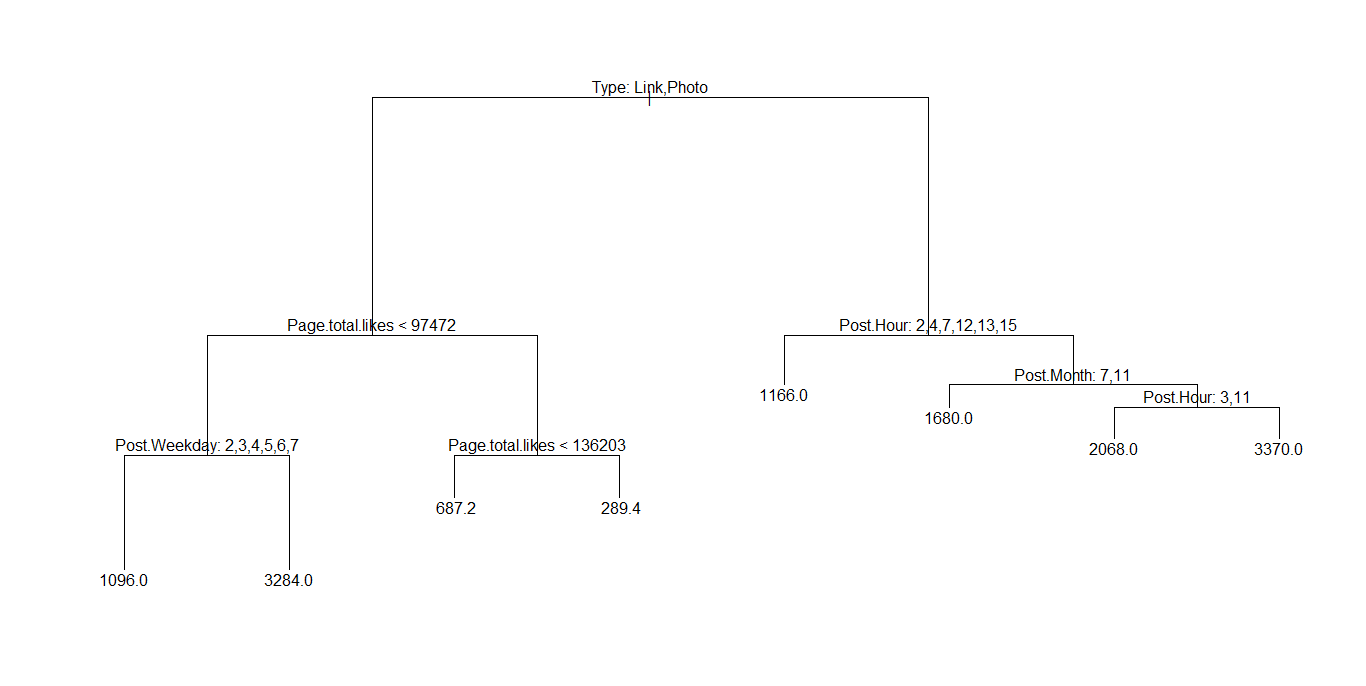
1. Dependent variable: Lifetime Post Total Reach 2. Dependent variable: Lifetime Post Total Impressions



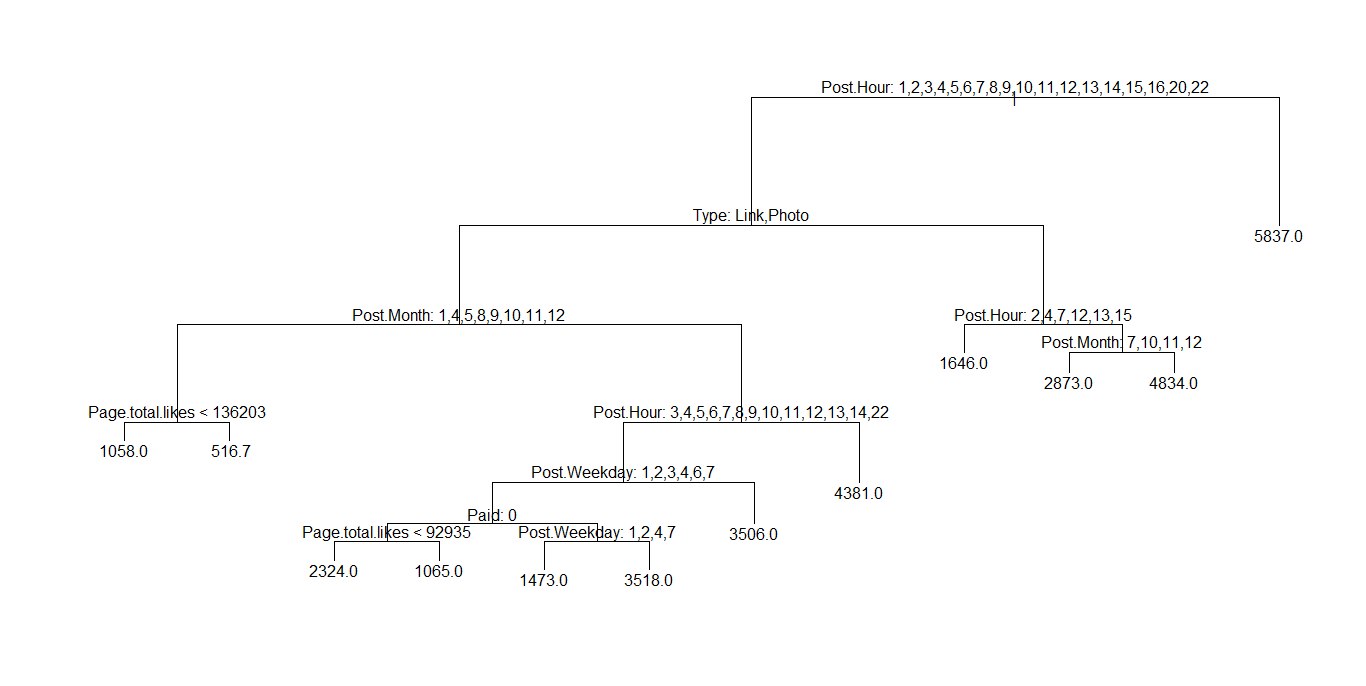
3. Dependent variable: Lifetime Engaged Users



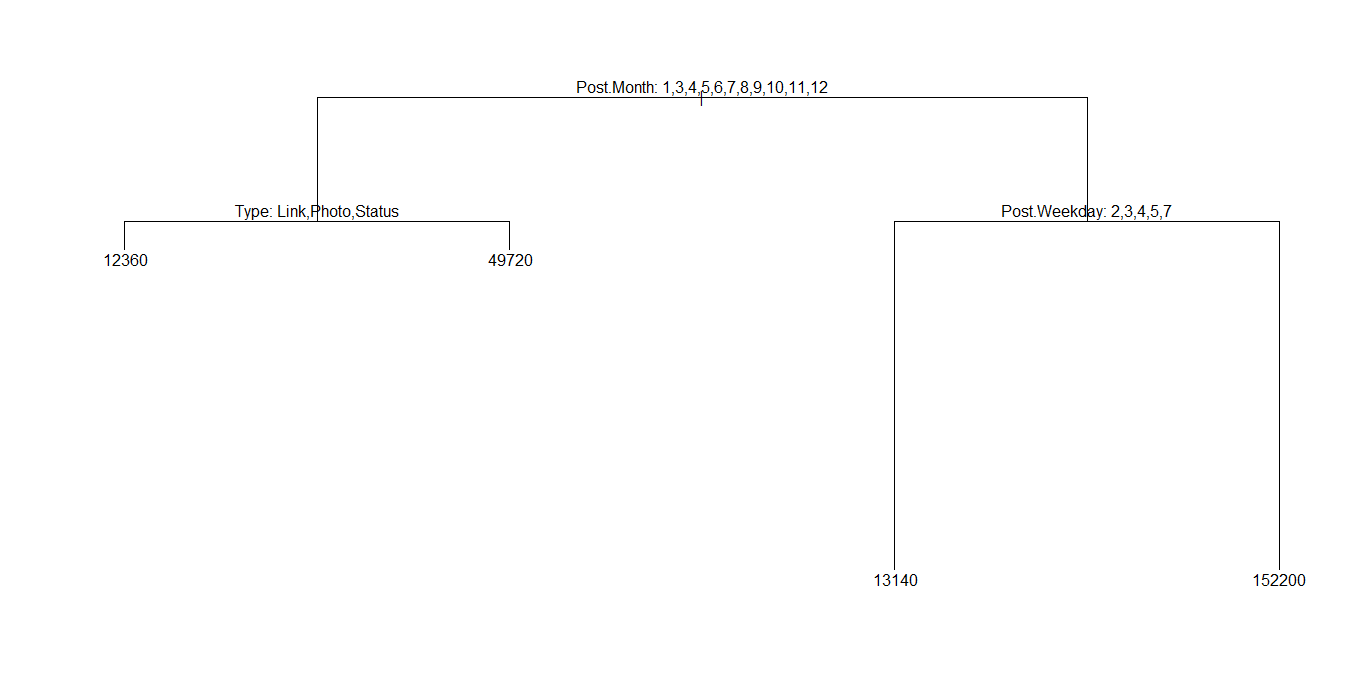
4. Dependent variable: Lifetime Post Consumers



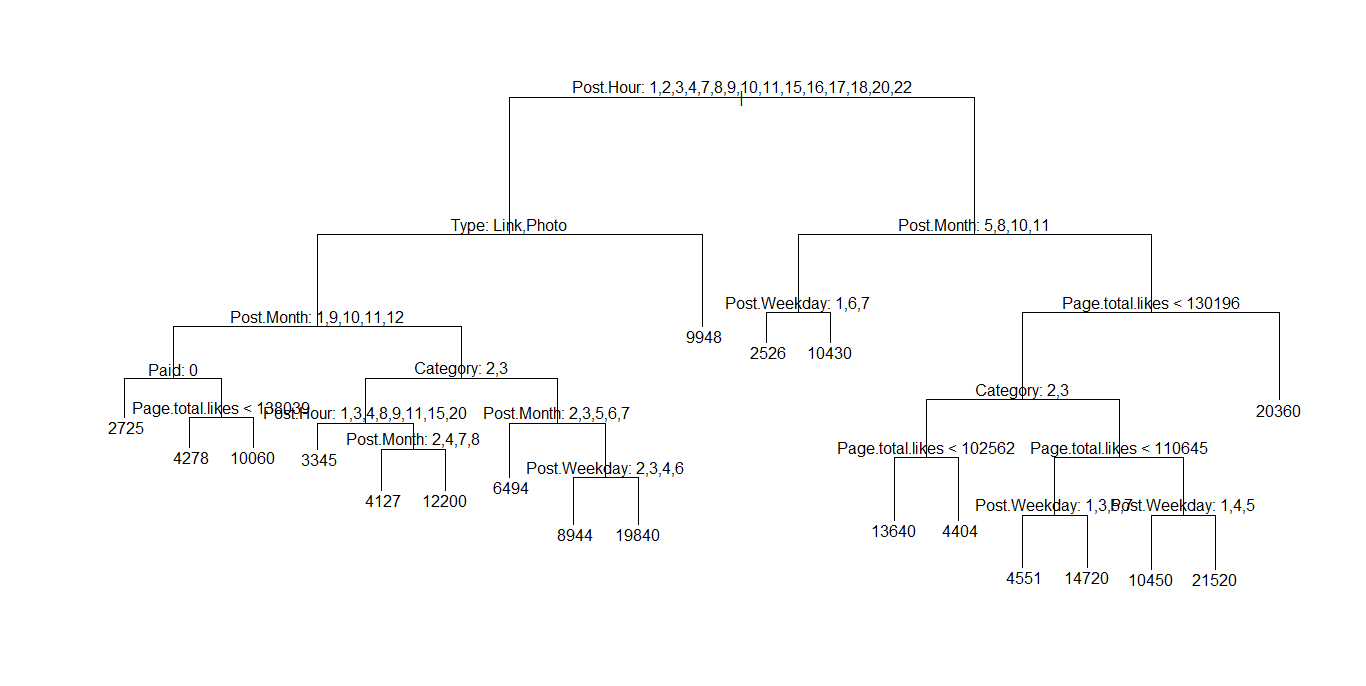
5. Dependent variable: Lifetime Post Consumptions



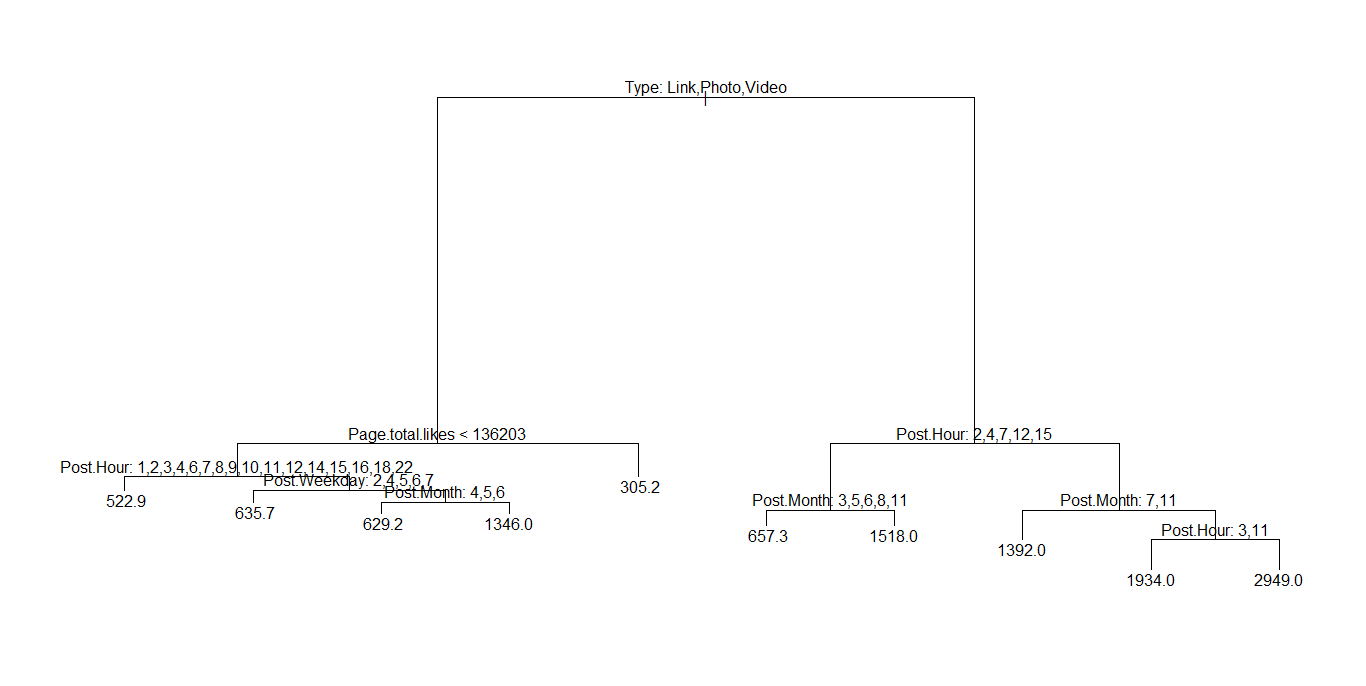
6. Dependent variable: Lifetime Post Impressions by people who have liked your Page



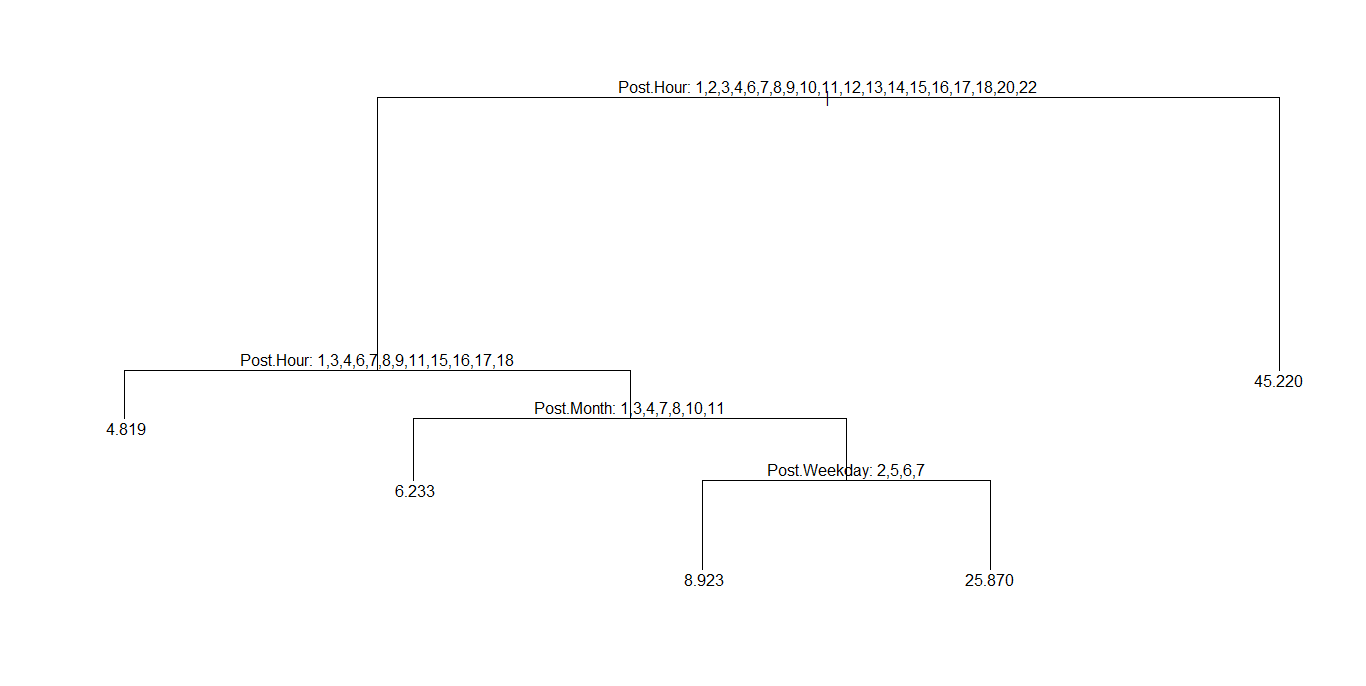
7. Dependent variable: Lifetime Post reach by people who like your Page



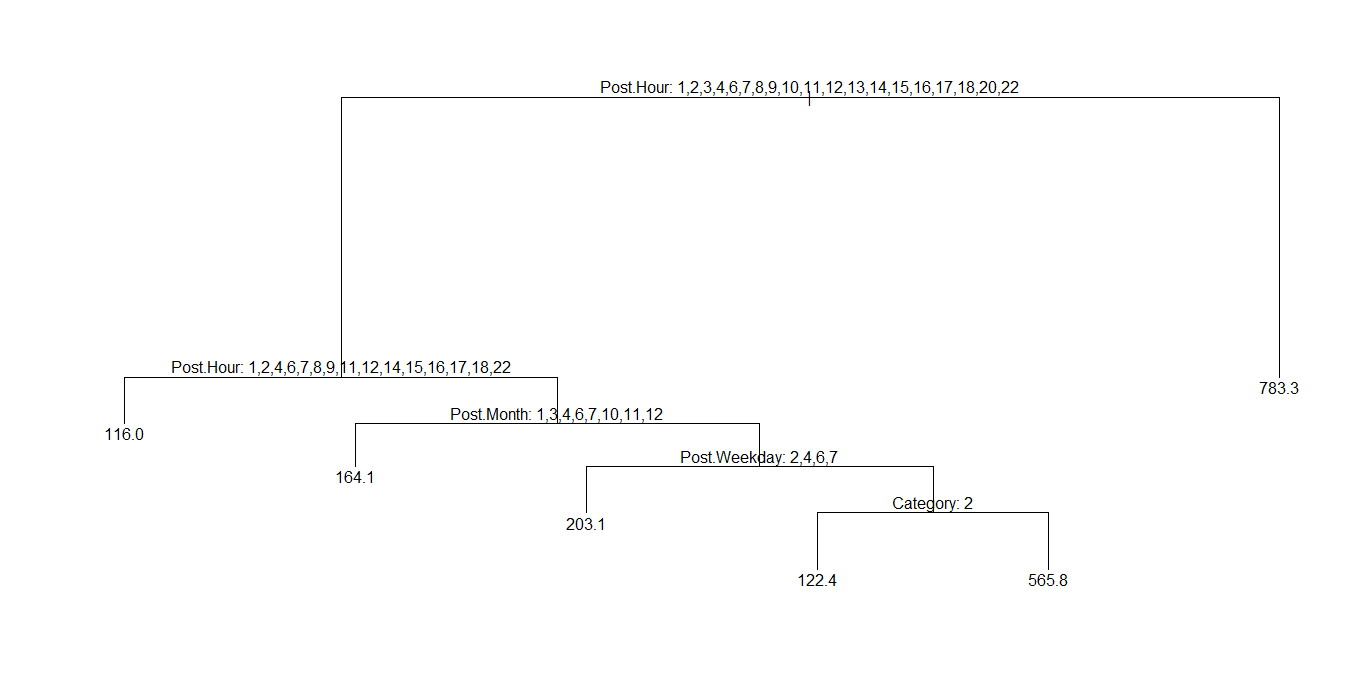
8. Dependent variable: Lifetime People who have liked your Page and engaged with your post



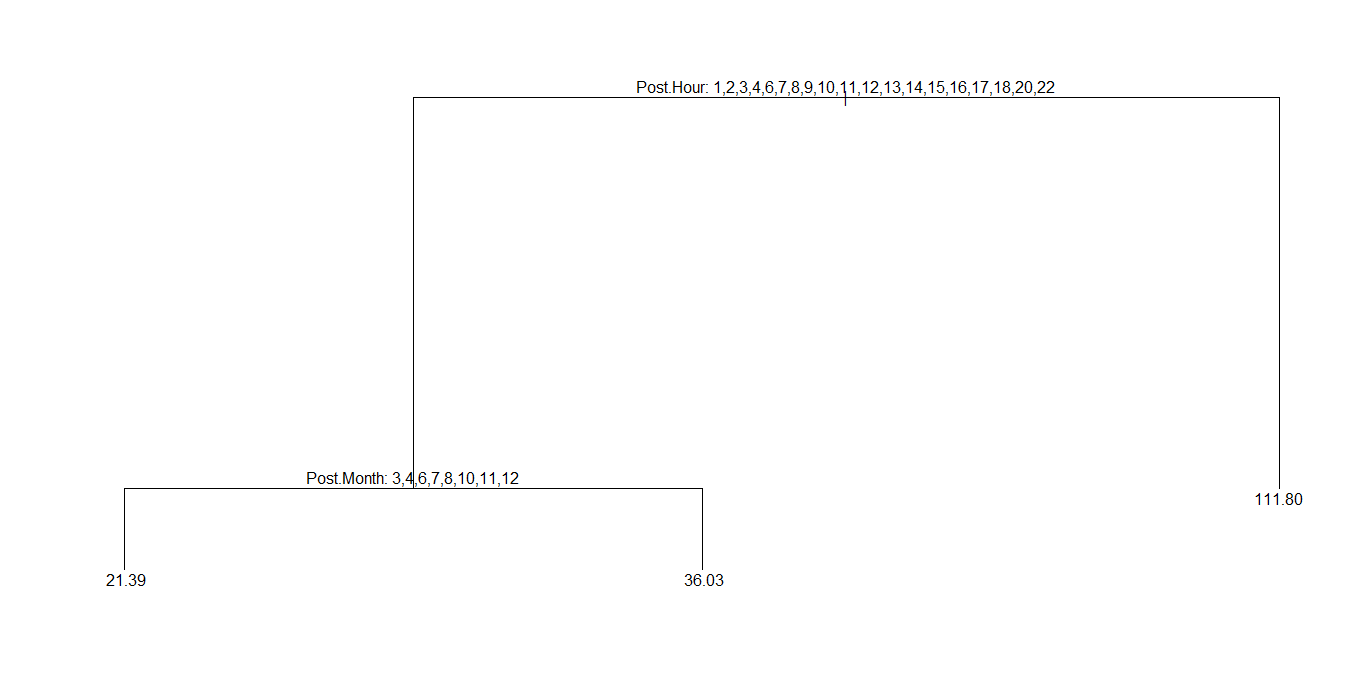
9. Dependent variable: comment



10. Dependent variable: like



11. Dependent variable: share



12. Dependent variable: Total interactions

