ISM 6137: Data Analytics Project

Prediction of Online News Popularity in Social Networks

Submitted by:

Shravya Katukuri

Sai Anvitha Malapati

Amar Guru Dutta Aduri

Sri Sarada Nutakki

Mano Bhargav Reddy Dodla

Nov 9th, 2018

Dr. Anol Bhattacherjee

1. Executive Summary

With the unprecedented expansion of the World Wide Web in the recent years, getting information on the go has become quite easy as almost all the major newspapers and news and media houses have dedicated websites that provide constant and up-to-date information about the latest happenings around the world and in every sector.

This paradigm shift is really good but it undoubtedly renders paramount importance to the popularity of these websites and their published articles, as good traffic to these websites would be one of the major sources of revenue to these companies. Hence, we wanted to pursue this analysis and come up with a few actionable outcomes and reasonable recommendations to help online portals and website owners increase traffic to their website and thereby revenue.

For the analysis, our dataset called **"Online News Popularity Dataset"** was sourced from **University of California, Irvine (UCI)'s Machine Learning Repository**, one of the largest sources of machine learning datasets ever to be maintained. The dataset had attributes like Number of shares, type of Data channel (Business, Entertainment, Lifestyle, Social Media, etc), number of images/videos and number of words in the article's content, etc. The standard approach of Data cleaning, Descriptive analysis, forming hypotheses and running the relevant regression models on the data was followed. Modelling of data was done using 3 different models and the best model that predicts the data best has been chosen.

The key findings from our analysis suggest that different data channels have different threshold values of images, videos and number of words in content and also that weekends have an effect on particular data channels. The aim of this report is to make actionable recommendations to website owners who can act on them and increase their revenue.

2. Table of Contents

3. Problem Significance

With the expansion of the Internet, more and more people enjoy reading and sharing online news articles and hence every day, there is ubiquitous amounts of information that is produced both from media and individuals alike and posted online for readers/users of social media. With the increase of digital footprint, this information is only going to see an upward increase in the future.

Inspite of having great content which has the potential of going viral, many good articles do not get shared more and reach a lot of people because of a few strategic factors. In such a scenario, a lot of effort is being put into what makes an article a hit; the factors that affect its popularity, etc. The right combination of factors when used properly can be the single most differentiating factor in determining the profitability of an online website or news channel.

A popular article can help maximize the advertisement revenue or it can be set on top of the homepage so as to increase revenue and traffic to that particular website. Hence, identifying an article that may become popular is of strategic and financial interest to websites like mashable.com (in our case) and many others that are looking to make profits as well. Our work would also help online news companies predict news popularity before publication. Hence, our aim is to provide useful and actionable recommendations to our client (the website owners) that they can act on and improvise their content.

4. Data Sourcing/Preparation

4.1 Data Source

Since the credibility and reliability of data is of paramount importance in any analysis, the entire data for this project has been sourced from the **Machine Learning Repository** of **University of California, Irvine (UCI)**. 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. Since the time of its creation in 1987, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning data sets. As an indication of the impact of the archive, it has been cited over 1000 times, making it one of the top 100 most cited "papers" in all of computer science. The repository currently maintains 452 data sets as a service to the machine learning community.

4.2 Data Cleaning and Preparation

The original dataset contained 39,644 instances of 61 attributes of which 58 attributes were predictive, 2 were non-predictive (url of the article and timedelta) and 1 was a goal field. All our variables were quantitative and had numerical values or binary values (0's and 1's).

Our data was fairly clean and organized. So, we did not have to delete any rows or columns but had to add a few new columns and modify a few for the purpose of our analysis. The addition/modification has been described below:

a) Extra columns were added to accommodate squared terms for three variables – Number of images, Number of Videos and Number of Words in Content to the dataset to convert their distribution from a monotonic to a non-linear curve. This aided us in determining the optimal values of images, videos and word count an article needed to have for generating maximum Number of Shares.

b) A column called **"Data Channel"** was created to group 6 columns (one each for Lifestyle, Entertainment, Business, Social Media, Technology and World) of different data channel categories into one. For rows that had 0's for all the above 6 categories, a category called **"Other"** was considered.

c) The values in the column **"Text Sentiment Polarity"** (negative, positive and 0) were modified as: rows with 0 were considered **"Neutral"**; rows with negative/positive values were considered as having a **"Sentiment"**.

d) Similarly, the values in the column **"Title Sentiment Polarity"** (negative, positive and 0) were also modified as: rows with 0 were considered **"Neutral"**; rows with negative/positive values were considered as having a **"Sentiment"**.

We chose the **Number of Shares (Number of times an article is shared)** as our dependent variable (DV) as online news popularity is determined/quantified by the number of times a news article is shared. More the number of shares, more viral the article gets and hence more popularity and vice-versa. In this project, we intend to find the best model and set of variables to predict the popularity of online news by using and comparing different Regression models.

The following set of Predictor and Target/Dependent variables have been selected:

Dependent Variable	Independent/Predictor Variables
Number of Shares (Target)	Number of Images
	Number of Videos
	Number of Words in the Content
	Text Sentiment Polarity
	Title Sentiment Polarity
	Is weekend
	Number of Words in the Content*Number of Images

5. Alternative Hypotheses (Rationale for selection)

The following hypotheses have been considered for this project:

1. H_{1a}: $\boldsymbol{\beta}_{NumberOfImages} > 0$
   Generally, an article with images tends to get shared more than an article without images due to heightened interest especially for data channels like Entertainment, Social Media, etc.

2. H_{2a}: $\boldsymbol{\beta}_{NumberOfVideos} > 0$
   Similar to images, videos in an article might also help generate interest and lead to increased shares.

3. H_{3a}: $\boldsymbol{\beta}_{NumberOfWordsInTheContent} > 0$
   The word count of the article (length) has an effect on the Number of Shares as people might be bored of reading if the article is too lengthy or might not get enough information if it is too short.

4. H_{4a}: $\boldsymbol{\beta}_{TextSentimentPolarity} > 0$
   Articles with a negative or a positive sentiment in their content tend to generate more hype and attract attention as opposed to neutral articles. This generated interest and a curiosity to know more might result in higher number of shares of such articles.

5. H_{5a}: $\boldsymbol{\beta}_{TitleSentimentPolarity} > 0$
   Title sentiment polarity would be similar in effect to text sentiment polarity. Articles with negative/positive titles would tend to catch the reader's attention quickly compared to a more neutral and general title which might end up in more number of shares.

6. H_{6a}: $\beta_{IsWeekend}>0$
   Weekends generally tend to see a spike in website traffic as opposed to weekdays as people tend to spend time online reading about their favourite topics even more so in the case of channels like Entertainment, Lifestyle, etc.

7. H_{7a}: $\beta_{NumberOfWordsInContent*NumberOfImages}>0$
   Having a few images in the article apart from content would be interesting/helpful (depending on the data channel) to the reader and might help in better understanding which might result in more number of shares.

6. Descriptive Analysis

The following parameters that are expected to have an impact have been graphically plotted to infer and explore any underlying trends in data.

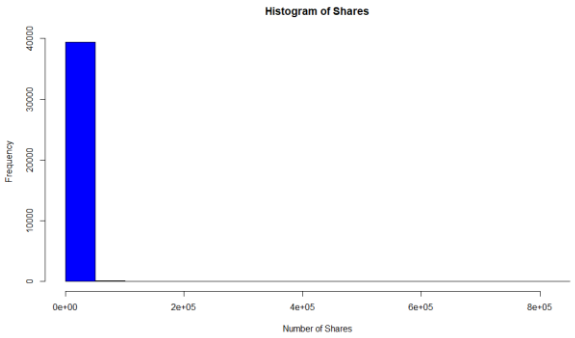


Figure 1- Histogram of Shares

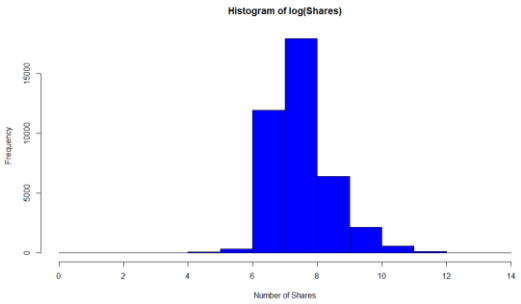


Figure 2-Histogram of log (Shares)

Since the histogram of Number of Shares (our DV) is highly skewed, a log scale has been used in all our models to estimate the number of shares. We can see that the histogram of log scale of Shares (Figure 2) also is not completely normally distributed but is better compared to the histogram of Shares (Figure 1).

Our DV has been plotted against our predictor variables like Number of images, Number of videos, Number of words in content, etc. Below are the graphs:

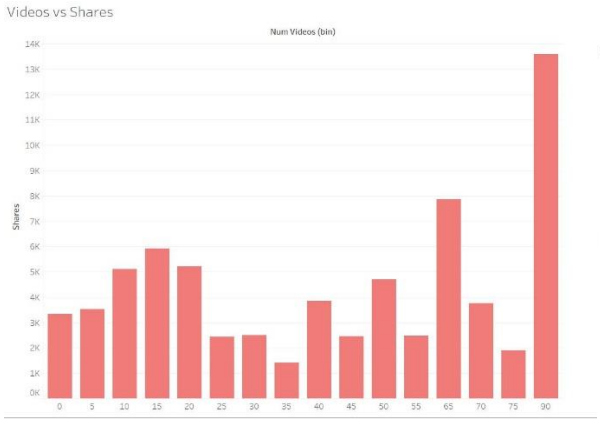


Figure 3- Number of videos vs Shares

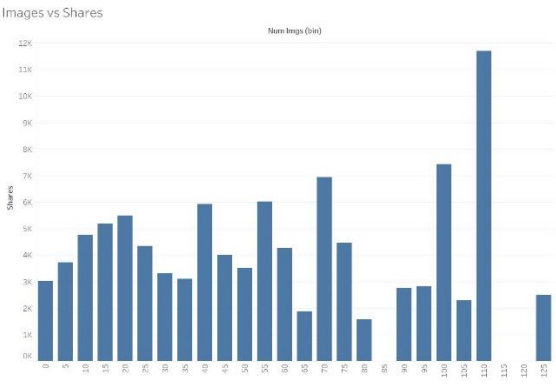


Figure 4- Number of images vs Shares

The above 2 bar graphs show that the number of shares is not linearly increasing with increase in images/videos. There are a few values for Number of images and videos that have extreme Number of Shares (either too low or too high).

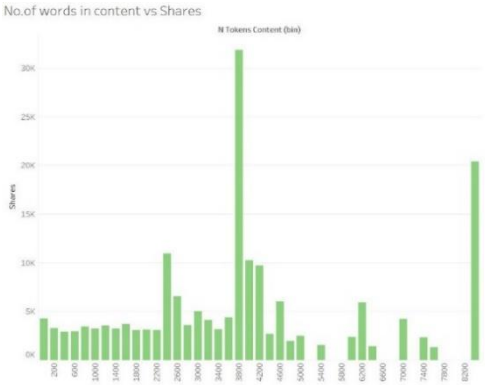
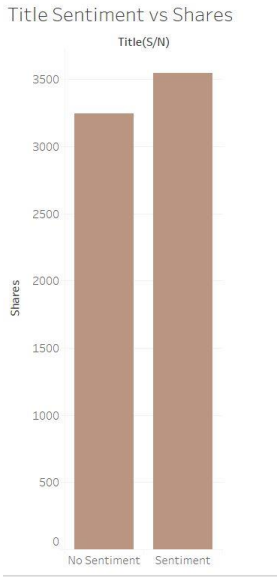


Figure 5-Number of words in content vs Shares

When "Number of Words in content" is plotted against Number of Shares, we can see that for articles exceeding a word count of 4200, the shares start dropping with an exception, however, for a word count of 8200.



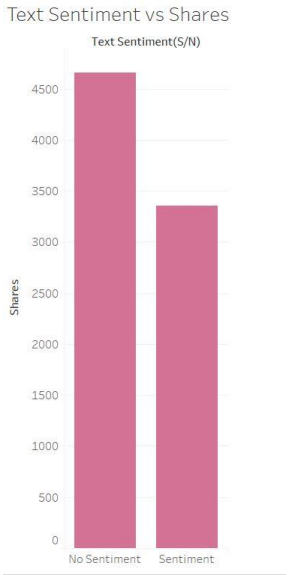


Figure 6-Title sentiment vs Shares Figure 7- Text sentiment vs Shares

The bar graph plotted for Title Sentiment (Negative or Positive) vs Number of Shares shows that articles with either a negative or a positive sentiment are shared more than articles that are generally neutral. But, the bar graph between Text Sentiment and Number of Shares shows that neutral articles tend to be shared more than negative/positive articles. This contradicts our assumption that articles with polarity in their title and text tend to be shared more.

7. Models

Based on the hypotheses formulated and the descriptive analyses, the following models have been considered:

Since our target variable is of count type (Number of Shares) and we want to determine which factors lead to more shares of an article, we used the Poisson distribution and generated the models using R.

7.1 Model 1

Number of Shares = *f* (Number of images, Number of videos, Number of images^2, Number of videos^2, Number of words in content, Number of words in content^2, Text sentiment polarity, Title sentiment polarity, Is weekend).

Where *f* = ***Poisson*** distribution.

7.2 Model 2

Number of Shares = *f* (Number of images, Number of videos, Number of images^2, Number of videos^2, Number of words in content, Number of words in content^2, Text sentiment polarity, Title sentiment polarity, Is weekend).

Where *f* = ***Quasi-Poisson*** distribution.

We have considered a Quasi-Poisson distribution for Model 2, as the Poisson model formulated in Model 1 does not consider the over-dispersion factor (the residual deviance being greater than degrees of freedom). This means that there is extra variance that is not accounted for by the model or by error structure.

7.3 Model 3

Number of Shares = *f* (Number of images, Number of videos, Number of images^2, Number of videos^2, Number of words in content, Number of words in content^2, Text sentiment polarity, Title sentiment polarity, Is weekend, ***Number of words in content*Number of images***).

Where *f* = ***Quasi-Poisson*** distribution.

In Model 3, we have considered an interaction term (the italicized and emboldened term in the function) to see its effect on other predictor variables and on the model.

7.3.1 Hypotheses rejection/acceptance

Hypothesis	Accepted/Rejected	Significance	Inference
H1	Accepted	High	Increase in number of images up to the optimal value increases the number of shares.
H2	Accepted	High	Increase in number of videos up to the optimal value increases the number of shares.
H3	Accepted	High	Increase in number of words in content up

			to the optimal value increases the number of shares.
H4	Accepted	Low	Sentiment in an article's text seems to affect the number of shares.
H5	Rejected	-	Sentiment in an article's title does not seem to affect the number of shares. Rather, neutral articles have a positive effect.
H6	Accepted	Medium	A few data channels seem to have more number of shares over the weekend.
H7	Accepted	Medium	Number of images increase as the word count of an article increases.

7.4 Comparison of Models

We have calculated the RMSE (Root Mean Square Error) for models 2 and 3 by splitting the entire dataset into training and test datasets (training->75% and test->25%). We used the observations in the training dataset to build the model required to predict the new values. Then, the observations in the testing data set were used to test the model built using the above approach.

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line, the data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells us how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

$RMSE_{Model2}$: 5809

$RMSE_{Model3}$: 6021

The values above show that Model 2 is statistically better as it has a lower RMSE value compared to Model 3. But, the difference is not huge. Our purpose of building different models was to compute optimal values for Number of images, videos and words in an article from Model 2 and to find the ratio of words to images from Model 3.

8. Quality Checks

The models have been built using the Quasi Poisson distribution. The following assumptions must be satisfied to ensure the quality and trustworthiness of our analysis:

8.1 Independence of Variables

Since all our variables are independent and influence the Number of Shares directly and no predictor variable is dependent on any other predictor variable, this assumption has been satisfied for our analysis.

8.2 Multi-Collinearity

All the predictor variables in our model, as such, are independent and not related to each other. But the addition of squared terms for Number of images, Number of videos and Number of words in content in our model introduces multi-collinearity. Therefore, this assumption has not been satisfied.

8.3 Auto-Correlation

We don't have any panel or time-series variables in our dataset. So, there is no auto correlation. Hence, this assumption is also satisfied.

8.4 Bias of the model (Stability of β-estimates)

Our β-coefficients across the 3 models are similar with not much variation i.e., they are stable. Hence, this assumption also has been satisfied.

9. Recommendations

Since our model had numeric variables, we ran regression models to calculate the optimal values of Number of images, Number of videos and Number of words in the content that an article needed to have to ensure maximum number of shares and increased popularity. This would, in turn, increase traffic to the corresponding websites and help website owners generate revenue through advertisements and offers, etc.

We have tabulated the optimal values for the above mentioned 3 factors below from Model 2 (without interaction) and Model 3 (with an interaction term) separately.

9.1 Table 1 (From Model 2):

Type of Data Channel	Number of words	Images	Videos
Overall	1840	27	24
Entertainment	1750	35	17
Business	233	35	35
Lifestyle	300000*	40	24
Social Media	2140	5	40
Technology	7000	54	13
World	2500	90	31
Other	1500	31	24

This table gives the optimal values for Number of words, images and videos for each data channel separately for maximum number of shares and an overall estimate for all the data channels overall.

* -- We got a really large value for the Number of words for Lifestyle data channel. It could be because of error in data.

9.2 Table 2 (From Model 3):

Type of Data Channel	No of words (500)	No of words(1000)	No of words (2000)
Overall	48	43	33
Entertainment	50	45	35
Business	67	54	28
Lifestyle	130	120	98
Social Media	-33	-70	-141
Technology	7	25	63
World	90	80	60
Other	36	30	20

This table gives the number of images that an article in any one of the Data channels with a word limit of 500, 1000 and 2000 should have for maximum number of shares.

Based on our analysis and findings, our recommendations would be:

- If the Data Channel is **'Social Media'**, for every 10 words, there should be 2 images i.e., a ratio of 1:5 (calculated from Model 3). The ratio is very high compared to other data channels and understandably so, as Social Media platforms generally tend to generate a lot of viral and trending content and a lot of it is graphically pleasing because of images. Hence, having images in the above ratio would ensure maximum number of shares.
- If the Data Channel is **'Technology'**, as the word count keeps increasing, the number of images needed increases drastically as people might find it difficult to understand and interpret latest technological trends just by reading a lot of tech jargon without proper images or representations. Hence, the number of images should be proportional to the content, the values for which are shown in Table 2.
- We can see from Table 2 that the Data Channel **'Lifestyle'** has the highest number of images for word count compared to other categories and understandably so, as this category is associated with a wide variety of topics like fitness, fashion, health and beauty that require a lot of images. Hence, images should be taken as per the ratio given in Table 2.
- If the Data Channel is **'Entertainment'** or **'World'**, there is a significant increase in number of shares during the weekend as opposed to weekdays. So, it would be profitable to publish articles of these categories during weekends. For other categories, the articles can be published either on a weekday or weekend, but the above categories are more profitable during weekends.
- **'Text sentiment polarity'** has a significant effect on Number of shares. But, articles which are neutral (having no sentiment) have a positive effect whereas articles with sentiment have a negative effect which is contrary to the general notion that articles with sentiment tend to be more popular. Hence, we recommend that website owners increase articles that are neutral in nature to maximize revenue.

These are a few recommendations based on our analysis which we think would be of substantial importance to the website owners in increasing the popularity of their articles and revenue from those.

10. Appendix

```
setwd("C:/Users/Anvitha/Desktop/SDM/Project")

pop<-read.csv("Online_news.csv")

pop

#Linear Regression method (model1)


model1<-lm(log(pop$shares)~
pop$num_imgs+pop$num_imgs.2+pop$num_videos+pop$n_tokens_content+pop$timedelta
+pop$Absolute_Text_Sentiment)

summary(model1)


#Assumptions check


plot(model1$residuals)

plot(model1)


#Poisson Model


glmmodel1<-glm((pop$shares)~
pop$n_tokens_content.2+pop$n_tokens_content+pop$num_imgs+pop$num_imgs.2+pop$nu
m_videos+pop$num_videos.2+
as.factor(pop$Title.S.N.)+as.factor(pop$Text.Sentiment.S.N.)+ as.factor(pop$is_weekend,
family = poisson)

summary(glmmodel1)


#Quasi-Poisson Model
```

```
glmmodel2<-glm((pop$shares)~
pop$n_tokens_content.2+as.factor(pop$Title.S.N.)+as.factor(pop$Text.Sentiment.S.N.)+pop
$n_tokens_content+pop$num_imgs+pop$num_imgs.2+pop$num_videos+pop$num_videos.2
+as.factor(pop$is_weekend),family = quasipoisson(link = log))

summary(glmmodel2)


#Quasi-Poisson Model with interaction


glmmodel3<-glm((pop$shares)~
pop$n_tokens_content*pop$num_imgs+pop$n_tokens_content.2+as.factor(pop$Title.S.N.)+
as.factor(pop$Text.Sentiment.S.N.)+pop$n_tokens_content+pop$num_imgs+pop$num_imgs
.2+pop$num_videos+pop$num_videos.2+as.factor(pop$is_weekend),family =
quasipoisson(link = log))

summary(glmmodel3)


residuals(glmmodel1,type="deviance")

predict(glmmodel1,type="link")


#For checking if variance is proportional to mean


p1Diag <- data.frame(pop,
                     link=predict(glmmodel1, type="link"),
                     fit=predict(glmmodel1, type="response"),
                     pearson=residuals(glmmodel1,type="pearson"),
                     resid=residuals(glmmodel1,type="response"),
                     residSqr=residuals(glmmodel1,type="response")^2)

install.packages("ggplot2")

library(ggplot)

ggplot(data=p1Diag, aes(x=fit, y=residSqr)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1) +
  geom_abline(intercept = 0, slope = summary(glmmodel2)$dispersion,color="green")+
  stat_smooth(method="loess", se = FALSE) + theme_bw()
```

```
#Entertainment data

Pop_Entertainment=pop[pop$Data_Channel == 'Entertainment',]

Pop_Entertainment


#Entertainment without interaction


glm_Entertainment_1<-
glm(Pop_Entertainment$shares~Pop_Entertainment$n_tokens_content.2+as.factor(Pop_Entertainment$Title.S.N.)+as.factor(Pop_Entertainment$Text.Sentiment.S.N.)+Pop_Entertainment$n_tokens_content+Pop_Entertainment$num_imgs+Pop_Entertainment$num_imgs.2+Pop_Entertainment$num_videos+Pop_Entertainment$num_videos.2+as.factor(Pop_Entertainment$is_weekend),family = quasipoisson(link = log))

summary(glm_Entertainment_1)


#Entertainment with interaction

glm_Entertainment<-
glm(Pop_Entertainment$shares~Pop_Entertainment$n_tokens_content*Pop_Entertainment$num_imgs+Pop_Entertainment$n_tokens_content.2+as.factor(Pop_Entertainment$Title.S.N.)+as.factor(Pop_Entertainment$Text.Sentiment.S.N.)+Pop_Entertainment$n_tokens_content+Pop_Entertainment$num_imgs+Pop_Entertainment$num_imgs.2+Pop_Entertainment$num_videos+Pop_Entertainment$num_videos.2+as.factor(Pop_Entertainment$is_weekend),family = quasipoisson(link = log))

summary(glm_Entertainment)


# Business data

Pop_Business=pop[pop$Data_Channel == 'Business',]


#Business without interaction

glm_Business_1<-
glm(Pop_Business$shares~Pop_Business$n_tokens_content.2+as.factor(Pop_Business$Title.S.N.)+as.factor(Pop_Business$Text.Sentiment.S.N.)+Pop_Business$n_tokens_content+Pop_Business$num_imgs+Pop_Business$num_imgs.2+Pop_Business$num_videos+Pop_Business$num_videos.2+as.factor(Pop_Business$is_weekend),family = quasipoisson(link = log))

summary(glm_Business_1)
```

```
#Business with interaction

glm_Business<-
glm(Pop_Business$shares~Pop_Business$n_tokens_content*Pop_Business$num_imgs+Pop
_Business$n_tokens_content.2+as.factor(Pop_Business$Title.S.N.)+as.factor(Pop_Business$
Text.Sentiment.S.N.)+Pop_Business$n_tokens_content+Pop_Business$num_imgs+Pop_Busi
ness$num_imgs.2+Pop_Business$num_videos+Pop_Business$num_videos.2+as.factor(Pop_
Business$is_weekend),family = quasipoisson(link = log))

summary(glm_Business)


#LifeStyle data


Pop_Lifestyle=pop[pop$Data_Channel == 'Lifestyle',]


#Lifestyle without interaction

glm_Lifestyle_1<-
glm(Pop_Lifestyle$shares~Pop_Lifestyle$n_tokens_content.2+as.factor(Pop_Lifestyle$Title.
S.N.)+as.factor(Pop_Lifestyle$Text.Sentiment.S.N.)+Pop_Lifestyle$n_tokens_content+Pop_
Lifestyle$num_imgs+Pop_Lifestyle$num_imgs.2+Pop_Lifestyle$num_videos+Pop_Lifestyl
e$num_videos.2+as.factor(Pop_Lifestyle$is_weekend),family = quasipoisson(link = log))

summary(glm_Lifestyle_1)


#Lifestyle with interaction


glm_Lifestyle<-
glm(Pop_Lifestyle$shares~Pop_Lifestyle$n_tokens_content*Pop_Lifestyle$num_imgs+Pop
_Lifestyle$n_tokens_content.2+as.factor(Pop_Lifestyle$Title.S.N.)+as.factor(Pop_Lifestyle$
Text.Sentiment.S.N.)+Pop_Lifestyle$n_tokens_content+Pop_Lifestyle$num_imgs+Pop_Life
style$num_imgs.2+Pop_Lifestyle$num_videos+Pop_Lifestyle$num_videos.2+as.factor(Pop_
Lifestyle$is_weekend),family = quasipoisson(link = log))

summary(glm_Lifestyle)


#Social_Media data


Pop_Social_Media=pop[pop$Data_Channel == 'Social Media',]
```

```
#Social_Media with interaction

glm_Social_Media<-
glm(Pop_Social_Media$shares~Pop_Social_Media$n_tokens_content*Pop_Social_Media$num_imgs+Pop_Social_Media$n_tokens_content.2+as.factor(Pop_Social_Media$Title.S.N.)+as.factor(Pop_Social_Media$Text.Sentiment.S.N.)+Pop_Social_Media$n_tokens_content+Pop_Social_Media$num_imgs+Pop_Social_Media$num_imgs.2+Pop_Social_Media$num_videos+Pop_Social_Media$num_videos.2+as.factor(Pop_Social_Media$is_weekend),family = quasipoisson(link = log))

summary(glm_Social_Media)


#Social_Media without interaction

glm_Social_Media_1<-
glm(Pop_Social_Media$shares~Pop_Social_Media$n_tokens_content.2+as.factor(Pop_Social_Media$Title.S.N.)+as.factor(Pop_Social_Media$Text.Sentiment.S.N.)+Pop_Social_Media$n_tokens_content+Pop_Social_Media$num_imgs+Pop_Social_Media$num_imgs.2+Pop_Social_Media$num_videos+Pop_Social_Media$num_videos.2+as.factor(Pop_Social_Media$is_weekend),family = quasipoisson(link = log))

summary(glm_Social_Media_1)


#Technology data

Pop_Technology=pop[pop$Data_Channel == 'Technology',]


#Technology with interaction

glm_Technology<-
glm(Pop_Technology$shares~Pop_Technology$n_tokens_content*Pop_Technology$num_imgs+Pop_Technology$n_tokens_content.2+as.factor(Pop_Technology$Title.S.N.)+as.factor(Pop_Technology$Text.Sentiment.S.N.)+Pop_Technology$n_tokens_content+Pop_Technology$num_imgs+Pop_Technology$num_imgs.2+Pop_Technology$num_videos+Pop_Technology$num_videos.2+as.factor(Pop_Technology$is_weekend),family = quasipoisson(link = log))

summary(glm_Technology)


#Technology without interaction

glm_Technology_1<-
glm(Pop_Technology$shares~Pop_Technology$n_tokens_content.2+as.factor(Pop_Technology$Title.S.N.)+as.factor(Pop_Technology$Text.Sentiment.S.N.)+Pop_Technology$n_tokens_content+Pop_Technology$num_imgs+Pop_Technology$num_imgs.2+Pop_Technology$nu
```

```
m_videos+Pop_Technology$num_videos.2+as.factor(Pop_Technology$is_weekend),family = quasipoisson(link = log))

summary(glm_Technology_1)


#World data

Pop_World=pop[pop$Data_Channel == 'World',]


#World with interaction

glm_World<-
glm(Pop_World$shares~Pop_World$n_tokens_content*Pop_World$num_imgs+Pop_World$n_tokens_content.2+as.factor(Pop_World$Title.S.N.)+as.factor(Pop_World$Text.Sentiment.S.N.)+Pop_World$n_tokens_content+Pop_World$num_imgs+Pop_World$num_imgs.2+Pop_World$num_videos+Pop_World$num_videos.2+as.factor(Pop_World$is_weekend),family = quasipoisson(link = log))

summary(glm_World)


#World without interaction

glm_World_1<-
glm(Pop_World$shares~Pop_World$n_tokens_content.2+as.factor(Pop_World$Title.S.N.)+as.factor(Pop_World$Text.Sentiment.S.N.)+Pop_World$n_tokens_content+Pop_World$num_imgs+Pop_World$num_imgs.2+Pop_World$num_videos+Pop_World$num_videos.2+as.factor(Pop_World$is_weekend),family = quasipoisson(link = log))

summary(glm_World_1)


#Other data

Pop_Other_Data=pop[pop$Data_Channel == 'Other_data',]


#Other data with interaction

glm_Other_data<-
glm(Pop_Other_Data$shares~Pop_Other_Data$n_tokens_content*Pop_Other_Data$num_imgs+Pop_Other_Data$n_tokens_content.2+as.factor(Pop_Other_Data$Title.S.N.)+as.factor(Pop_Other_Data$Text.Sentiment.S.N.)+Pop_Other_Data$n_tokens_content+Pop_Other_Data$num_imgs+Pop_Other_Data$num_imgs.2+Pop_Other_Data$num_videos+Pop_Other_Data$num_videos.2+as.factor(Pop_Other_Data$is_weekend),family = quasipoisson(link = log))

summary(glm_Other_data)
```

```
#Other data without interaction

glm_Other_data_1<-
glm(Pop_Other_Data$shares~Pop_Other_Data$n_tokens_content.2+as.factor(Pop_Other_Data$Title.S.N.)+as.factor(Pop_Other_Data$Text.Sentiment.S.N.)+Pop_Other_Data$n_tokens_content+Pop_Other_Data$num_imgs+Pop_Other_Data$num_imgs.2+Pop_Other_Data$num_videos+Pop_Other_Data$num_videos.2+as.factor(Pop_Other_Data$is_weekend),family = quasipoisson(link = log))

summary(glm_Other_data_1)


install.packages("lmtest")

library(lmtest)

R<-coeftest(glmmodel1)

R

plot(glmmodel1)

cor(pop)

install.packages("car")

library("car")


#Divide into Training and Testing data in the ratio 75% to 25%


require(caTools)

set.seed(123)

sample=sample.split(pop,SplitRatio = 0.75)

train=subset(pop,sample==TRUE)

View(train)

testing=subset(pop,sample==FALSE)

View(testing)

mod1<-glm((train$shares)~
+train$n_tokens_content*train$num_imgs+train$n_tokens_content.2+as.factor(train$Title.S.N.)+as.factor(train$Text.Sentiment.S.N.)+train$n_tokens_content+train$num_imgs+train$num_imgs.2+train$num_videos+train$num_videos.2+as.factor(train$is_weekend),family = quasipoisson(link = log))

train<-testing

prediction<-predict(mod1, newdata=train, type="response")
```

prediction

rmse<-sqrt(1/10179*(sum(testing$shares - prediction)^2))

rmse

Output:

```
Coefficients:
                                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                              7.723e+00  1.854e-01  41.660  < 2e-16 ***
Pop_Entertainment$n_tokens_content                      -3.992e-04  1.171e-04  -3.409 0.000655 ***
Pop_Entertainment$num_imgs                               3.232e-02  6.211e-03   5.204 2.00e-07 ***
Pop_Entertainment$n_tokens_content.2                     1.380e-07  2.572e-08   5.367 8.25e-08 ***
as.factor(Pop_Entertainment$Title.S.N.)Sentiment         6.839e-02  6.030e-02   1.134 0.256784
as.factor(Pop_Entertainment$Text.Sentiment.S.N.)Sentiment 1.862e-01  1.869e-01   0.996 0.319219
Pop_Entertainment$num_imgs.2                            -2.948e-04  1.250e-04  -2.358 0.018396 *
Pop_Entertainment$num_videos                             3.568e-02  1.446e-02   2.468 0.013619 *
Pop_Entertainment$num_videos.2                          -1.078e-03  5.488e-04  -1.964 0.049625 *
as.factor(Pop_Entertainment$is_weekend)1                 2.252e-01  8.164e-02   2.758 0.005830 **
Pop_Entertainment$n_tokens_content:Pop_Entertainment$num_imgs -5.811e-06  3.093e-06  -1.879 0.060327 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 18602.97)

    Null deviance: 40624081  on 7056  degrees of freedom
Residual deviance: 39598474  on 7046  degrees of freedom
AIC: NA



Deviance Residuals:
    Min       1Q   Median       3Q      Max
-164.55   -33.63   -24.26    -7.77  1457.88

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        7.520e+00  1.581e-01  47.550  < 2e-16 ***
Pop_World$n_tokens_content                        -4.788e-04  1.192e-04  -4.016 5.98e-05 ***
Pop_World$num_imgs                                 5.384e-02  8.689e-03   6.196 6.05e-10 ***
Pop_World$n_tokens_content.2                       1.083e-07  2.670e-08   4.056 5.04e-05 ***
as.factor(Pop_World$Title.S.N.)Sentiment           1.206e-01  5.445e-02   2.215 0.026821 *
as.factor(Pop_World$Text.Sentiment.S.N.)Sentiment  1.845e-01  1.587e-01   1.163 0.244949
Pop_World$num_imgs.2                              -2.795e-04  1.118e-04  -2.499 0.012460 *
Pop_World$num_videos                               7.566e-02  2.161e-02   3.501 0.000466 ***
Pop_World$num_videos.2                            -1.210e-03  7.437e-04  -1.627 0.103696
as.factor(Pop_World$is_weekend)1                   1.783e-01  7.596e-02   2.347 0.018966 *
Pop_World$n_tokens_content:Pop_World$num_imgs     -1.130e-05  5.566e-06  -2.030 0.042388 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 14200.61)


Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                          7.734e+00  8.721e-01   8.869  < 2e-16 ***
Pop_Business$n_tokens_content                        7.380e-05  2.073e-04   0.356   0.7218
Pop_Business$num_imgs                                5.180e-02  3.034e-02   1.708   0.0878 .
Pop_Business$n_tokens_content.2                      3.860e-08  4.869e-08   0.793   0.4279
as.factor(Pop_Business$Title.S.N.)Sentiment          -1.037e-01  1.061e-01  -0.978   0.3283
as.factor(Pop_Business$Text.Sentiment.S.N.)Sentiment  9.082e-02  8.767e-01   0.104   0.9175
Pop_Business$num_imgs.2                             -3.198e-04  1.076e-03  -0.297   0.7664
Pop_Business$num_videos                              1.242e-01  1.861e-02   6.674 2.70e-11 ***
Pop_Business$num_videos.2                           -1.762e-03  4.288e-04  -4.110 4.01e-05 ***
as.factor(Pop_Business$is_weekend)1                  2.485e-01  1.667e-01   1.491   0.1360
Pop_Business$n_tokens_content:Pop_Business$num_imgs -1.699e-05  1.905e-05  -0.892   0.3727
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 52895.67)
```

```
Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        8.519e+00  3.835e-01  22.214  < 2e-16 ***
Pop_Lifestyle$n_tokens_content                     2.872e-04  1.491e-04   1.927   0.0542 .
Pop_Lifestyle$num_imgs                             2.167e-02  1.098e-02   1.974   0.0485 *
Pop_Lifestyle$n_tokens_content.2                  -9.722e-10  2.490e-08  -0.039   0.9689
as.factor(Pop_Lifestyle$Title.S.N.)Sentiment      -8.850e-02  9.666e-02  -0.916   0.3600
as.factor(Pop_Lifestyle$Text.Sentiment.S.N.)Sentiment -6.264e-01  3.853e-01  -1.626   0.1042
Pop_Lifestyle$num_imgs.2                          -7.613e-05  3.856e-04  -0.197   0.8435
Pop_Lifestyle$num_videos                           1.426e-01  3.236e-02   4.407  1.1e-05 ***
Pop_Lifestyle$num_videos.2                        -2.937e-03  1.310e-03  -2.243   0.0250 *
as.factor(Pop_Lifestyle$is_weekend)1              -3.314e-04  1.245e-01  -0.003   0.9979
Pop_Lifestyle$n_tokens_content:Pop_Lifestyle$num_imgs -3.374e-06  4.805e-06  -0.702   0.4826
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 17563.62)
```

```
Coefficients:
                                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                                         7.885e+00  4.315e-01  18.275  < 2e-16 ***
Pop_Social_Media$n_tokens_content                   5.614e-04  1.415e-04   3.967  7.5e-05 ***
Pop_Social_Media$num_imgs                           9.798e-03  1.070e-02   0.916   0.3599
Pop_Social_Media$n_tokens_content.2                -1.014e-07  5.506e-08  -1.841   0.0657 .
as.factor(Pop_Social_Media$Title.S.N.)Sentiment      9.380e-02  6.027e-02   1.556   0.1197
as.factor(Pop_Social_Media$Text.Sentiment.S.N.)Sentiment -1.950e-03  4.339e-01  -0.004   0.9964
Pop_Social_Media$num_imgs.2                        -1.018e-04  3.802e-04  -0.268   0.7888
Pop_Social_Media$num_videos                         1.106e-02  1.297e-02   0.853   0.3938
Pop_Social_Media$num_videos.2                      -1.777e-04  3.069e-04  -0.579   0.5626
as.factor(Pop_Social_Media$is_weekend)1             3.874e-02  8.545e-02   0.453   0.6504
Pop_Social_Media$n_tokens_content:Pop_Social_Media$num_imgs -1.466e-05  6.857e-06  -2.139   0.0326 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 7557.548)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-190.01   -37.95   -24.67     1.46  2081.58

Coefficients:
                                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                                         8.176e+00  1.509e+01   0.542    0.588
Pop_Technology$n_tokens_content                     5.537e-04  4.002e-03   0.138    0.890
Pop_Technology$num_imgs                             8.696e-03  3.481e-01   0.025    0.980
Pop_Technology$n_tokens_content.2                  -4.885e-09  1.132e-06  -0.004    0.997
as.factor(Pop_Technology$Title.S.N.)Sentiment       4.137e-02  1.856e+00   0.022    0.982
as.factor(Pop_Technology$Text.Sentiment.S.N.)Sentiment -5.383e-01  1.517e+01  -0.035    0.972
Pop_Technology$num_imgs.2                           3.449e-04  9.305e-03   0.037    0.970
Pop_Technology$num_videos                           1.388e-01  1.255e+00   0.111    0.912
Pop_Technology$num_videos.2                        -5.922e-03  1.140e-01  -0.052    0.959
as.factor(Pop_Technology$is_weekend)1               1.943e-01  2.574e+00   0.076    0.940
Pop_Technology$n_tokens_content:Pop_Technology$num_imgs -2.599e-05  2.113e-04  -0.123    0.902

(Dispersion parameter for quasipoisson family taken to be 19226411)
```