Facebook Comments Volume Prediction

Post Graduate Program in Business Analytics and Business Intelligence

Capstone Project

GREAT LAKES
INSTITUTE OF MANAGEMENT
Global Mindset - Indian Roots

Submitted by :

Neha Narendra Jasani

Batch : PGP BABI JAN'19

Table of Contents

Introduction 4
Problem Statement 5
Project Objective : 5
Objective and scope of the project: 6
Scope 6
Data Dictionary 6
Data Report 8
Descriptive statistic inference of the variables. 10
Exploratory data analysis 11
Plotting the independent variables. 13
Distribution of Page Likes. 13
Distribution of Page Check ins. 13
Distribution of page talking about 14
Distribution of Page Category 14
Distribution of all the features given in the dataset. 15
Distribution of CC1 to CC5 18
Distribution of Post length 21
Correlation Heat Map 24
Missing Values in the dataset: 25
Removal of unwanted variables. 25
Additional Insights : 25
Model Building and Interpretation: 26
Preparing the data. 26
Model building: 29
Random Forest model: 29
Linear Regression 30
SVM 30
Extreme Gradient boosting 30
Model Tuning 31
a. Step wise regression of LM model 31
b. Performing cross validation 31
Insights from the analysis 32
Recommendations 32
Appendix 32

Introduction :

Social media platforms are considered as one of the most important source for data. On a day to day basis, these platforms are being updated with massive amount of data. One such platform which serves the best source for data is Facebook. Facebook is widely used by individuals as well as corporates. Most of the data that is used for analysis includes the likes, shares and comment activities of the users. However, ‘comments’ are considered to be the most important part of study for the past decade. Comments are of the important measures of popularity towards any post. Hence the data collected via comments has been used majorly for Marketing and Advertising purposes. Facebook has extensively helped the brands to create a significant relationship with their customers online. It has assisted the companies to receive positive response and create more traffic.

For both small business and large corporations Facebook has played a vital role in customer satisfaction and brand building. The advertising revenue of Facebook in the United Stated in 2018 stands up to 14.89 billion US $. The advertising revenue outside the United Stated comes down to 18.95 billion US$. Latest research reports have indicated that the user generated content on Facebook drives higher engagement than ads. Hence the data generated by these comments are extensively used for effective marketing strategies and to create meaningful \ customised advertisement for the users.

Problem Statement

For both small businesses and large corporations, social media is playing a key role in brand building and customer communication. Facebook is the social networking site relevant for firms to make themselves real for customers. Just to put things in context, the advertising revenue of Facebook in the United States in 2018 stands up to 14.89 billion US dollars. The advertising revenue outside the United States comes down to 18.95 billion US dollars. Latest research reports have indicated that user generated content on facebook drives higher engagement than ads. The amount of data that gets added to the network increases day by day and it is a gold mine of researchers who want to understand the intricacies of user behaviour and user engagement. In this Hackathon, we discuss one such problem where we take a step towards understanding the highly dynamic behaviour of users towards Facebook posts.

The goal is to predict how many comments a user generated posts is expected to receive in the given set of hours. We need to model the user comments pattern over a set of variables which are provided and get to the right number of comments for each post with minimum error possible.

Project Objective :

The purpose of this project is to predict how many comments a user-generated post is expected to receive in the given set of hours. We need to model the user comments pattern over a set of variables which are provided and get to the right number of comments for each post with minimum error possible and finally derive meaningful insights for effective marketing strategies.

Objective and scope of the project:

Below are the objectives of the project:

1. To understand if the data can provide us with any form of patterns, provide any insights and give any relevant information to address the problem.
2. Build different models to predict the number of comments in each set of hours.

Scope of the project:

1. Validation and interpretation of the models build.
2. Building various model and checking the accuracy.
3. Interpreting the best model
4. Providing business recommendations

Data Dictionary

Variable name	**Description**	**Feature type**
Page Popularity/likes	Defines the popularity or support for the source	Page feature
Page Checkins	Describes how many individuals so far visited this place. This feature is only associated with the places eg:some institution, place, theater etc.	Page feature
Page talking about	Defines the daily interest of individuals towards source. The people who actually come back to the page, after liking the page. This include activities such as	Page feature

	comments, likes to a post, shares, etc by visitors to the page.	
Page Category	Defines the category of the source eg: place, institution, brand etc	Page feature
Feature 5 – Feature 29	These features are aggregated by page, by calculating min, max, average, median and standard deviation of essential features.	Derived features
CC1	The total number of comments before selected base date/time	Essential feature
CC2	The number of comments in last 24 hours, relative to base date/time.	Essential feature
CC3	The number of comments in last 48 to last 24 hours relative to base date/time.	Essential feature
CC4	The number of comments in the first 24 hours after the publication of post but before base date/time	Essential feature
CC5	The difference between CC2 and CC3	Essential feature
Base time	Selected time in order to simulate the scenario	Other feature
Post length	Character count in the post	Other feature
Post Share Count	This features counts the no of shares of the post, that how many peoples had shared this post on to their timeline.	Other feature
Post Promotion Status	To reach more people with posts in News Feed, individual promote their post and this features tells that whether the post is promoted(1) or not(0).	Other feature

H Local	This describes the H hrs., for which we have the target variable/ comments received.	Other feature
Post published weekday	This represents the day (Sunday...Saturday) on which the post was published.	Day of the week (Categorical)
Base Date Time weekday	This represents the day(Sunday...Saturday) on selected base Date/Time.	Day of the week (Categorical)
Comments	The no of comments in next H hrs.(H represents H Local).	Target Variable

Data Report

The dataset used for the project is the training dataset of Facebook comment volume prediction. There are a total of 43 variables in the dataset and 32759 observations. Out of the 43 variables, 42 variables are independent and only 1 dependent variable. Also, there are just 2 categorical variables in the entire data set, and all other are numeric variables.

Below are the feature grouping of the variables of the dataset.

a. Page features : The main 4 variables that are included as page feature are the page check ins , page popularity or likes , page talking about and page category. The page popularity / likes variable defines the popularity or the support for the source. The check in variable talks about how many individuals have visited the page so far. The feature is only associated with the place for example institution, place or theatre. The next feature explains the daily interest of individuals towards the source , the people who actually come back to the page, after liking the page. This include activities such as comments, likes to a post, shares, etc by visitors to the page. The page category defines the category of the source place, institution, brand etc.

b. Essential features:

There are 5 features in the dataset that are noted to be the essential feature. These features provides data on the number of comments that are posted in different time intervals.

CC1 : The total number of comments before selected base date/time. CC2 : The number of comments in last 24 hours, relative to base date/time.

CC3 : The number of comments in last 48 to last 24 hours relative to base date/time. CC4 : The number of comments in the first 24 hours after the publication of post but before base date/time CC5 : The difference between CC2 and CC3

c. Other features: The base time explains the selected time in order to simulate the scenario. The post length provides character count in the post. Post Share Count counts the no of shares of the post, that how many peoples had shared this post on to their timeline. H Local describes the H hrs., for which we have the target variable/ comments received.

d. Target variable: Comments is the target variable for the given dataset. The variable gives the number of comments a post has received in H hours.

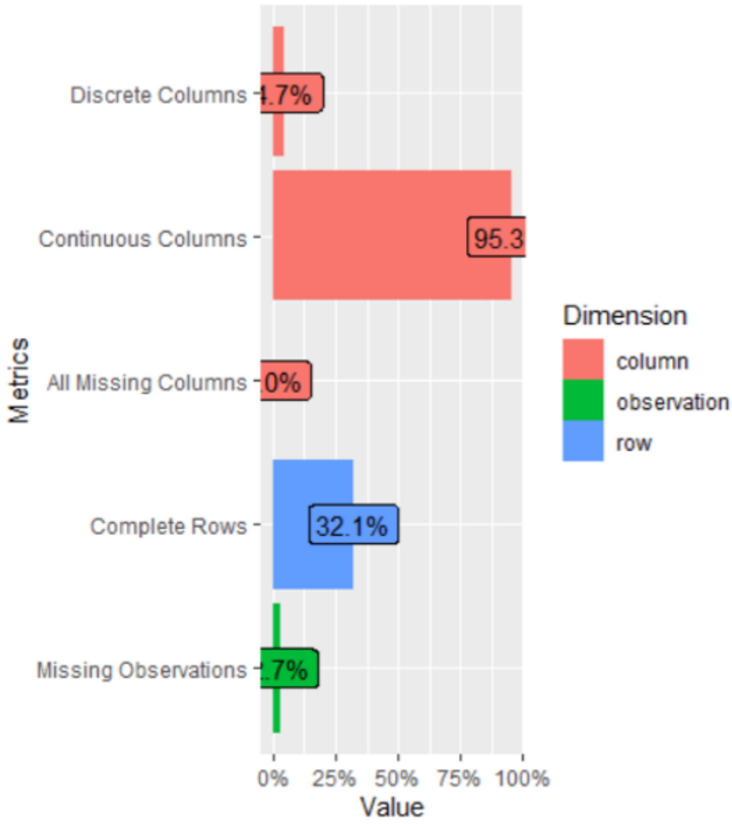
Descriptive statistic inference of the variables.

```
       ID          Page.likes          Page.Checkins      Page.talking.about Page.Category      Feature.5          Feature.6        Feature.7
Min.   :129525   Min.   :       36   Min.   :     0   Min.   :      0   Min.   :  1.00   Min.   :   0.000   Min.   :   0   Min.   :   0.000
1st Qu.:137715   1st Qu.:    35879   1st Qu.:     0   1st Qu.:    698   1st Qu.:  9.00   1st Qu.:   0.000   1st Qu.:  45   1st Qu.:   5.318
Median :145904   Median :   287698   Median :     0   Median :   6802   Median : 18.00   Median :   0.000   Median : 241   Median :  23.374
Mean   :145904   Mean   :  1346069   Mean   :  4645   Mean   :  44913   Mean   : 24.31   Mean   :   1.541   Mean   : 443   Mean   :  55.650
3rd Qu.:154094   3rd Qu.:  1204214   3rd Qu.:    99   3rd Qu.:  50264   3rd Qu.: 32.00   3rd Qu.:   0.000   3rd Qu.: 717   3rd Qu.:  71.829
Max.   :162283   Max.   :486972297   Max.   :186370   Max.   :6089942   Max.   :106.00   Max.   :2341.000   Max.   :2341   Max.   :2341.000
                 NA's   :3208        NA's   :3255     NA's   :3255      NA's   :3024                                        NA's   :1679
  Feature.8        Feature.9        Feature.10          Feature.11        Feature.12         Feature.13       Feature.14         Feature.15
Min.   :   0.0   Min.   :  0.00   Min.   :  0.0000   Min.   :   0.0   Min.   :  0.000   Min.   :  0.00   Min.   :  0.000   Min.   :  0.0000
1st Qu.:   2.0   1st Qu.:  7.88   1st Qu.:  0.0000   1st Qu.:  26.0   1st Qu.:  1.902   1st Qu.:  0.00   1st Qu.:  4.109   1st Qu.:  0.0000
Median :  12.0   Median : 35.07   Median :  0.0000   Median : 118.0   Median :  8.438   Median :  2.00   Median : 17.383   Median :  0.0000
Mean   :  35.6   Mean   : 67.45   Mean   :  0.1811   Mean   : 285.3   Mean   : 22.122   Mean   :  7.49   Mean   : 40.446   Mean   :  0.0286
3rd Qu.:  42.0   3rd Qu.:101.73   3rd Qu.:  0.0000   3rd Qu.: 403.0   3rd Qu.: 29.006   3rd Qu.:  8.00   3rd Qu.: 60.760   3rd Qu.:  0.0000
Max.   :2341.0   Max.   :731.39   Max.   :381.0000   Max.   :2079.0   Max.   :639.000   Max.   :649.00   Max.   :469.539   Max.   :324.0000
                                  NA's   :1632                                          NA's   :1643                       NA's   :1692
  Feature.16      Feature.17        Feature.18        Feature.19        Feature.20          Feature.21       Feature.22          Feature.23
Min.   :   0   Min.   :  0.000   Min.   :  0.00   Min.   :  0.000   Min.   :   0.000   Min.   :   0.0   Min.   :   0.000   Min.   :   0.00
1st Qu.:  26   1st Qu.:  2.027   1st Qu.:  0.00   1st Qu.:  4.095   1st Qu.:   0.000   1st Qu.:  41.0   1st Qu.:   4.945   1st Qu.:   2.00
Median : 116   Median :  8.584   Median :  1.00   Median : 18.640   Median :   0.000   Median : 224.0   Median :  21.859   Median :  12.00
Mean   : 268   Mean   : 19.661   Mean   :  4.94   Mean   : 38.689   Mean   :   1.459   Mean   : 415.2   Mean   :  52.486   Mean   :  33.99
3rd Qu.: 381   3rd Qu.: 24.843   3rd Qu.:  5.00   3rd Qu.: 54.634   3rd Qu.:   0.000   3rd Qu.: 670.0   3rd Qu.:  67.914   3rd Qu.:  40.00
Max.   :1605   Max.   :437.684   Max.   :433.00   Max.   :533.639   Max.   :1897.000   Max.   :2184.0   Max.   :1897.000   Max.   :1897.00
                                 NA's   :1605                       NA's   :1600                         NA's   :1601
  Feature.24        Feature.25         Feature.26         Feature.27           Feature.28          Feature.29          CC1
Min.   :  0.000   Min.   :-1366.0   Min.   :-204.0   Min.   :-210.5000   Min.   :-288.000   Min.   :  0.000   Min.   :   0.00
1st Qu.:  7.528   1st Qu.: -310.0   1st Qu.:  23.0   1st Qu.:  -0.4832   1st Qu.:  -2.000   1st Qu.:  5.991   1st Qu.:   2.00
Median : 32.369   Median :  -92.0   Median : 109.0   Median :   0.2738   Median :   0.000   Median : 25.547   Median :  11.00
Mean   : 63.144   Mean   : -219.8   Mean   : 275.6   Mean   :   2.4752   Mean   :  -2.113   Mean   : 55.801   Mean   :  55.45
3rd Qu.: 95.880   3rd Qu.:  -21.0   3rd Qu.: 379.0   3rd Qu.:   2.9747   3rd Qu.:   0.000   3rd Qu.: 81.209   3rd Qu.:  45.00
Max.   :703.144   Max.   :  381.0   Max.   :2079.0   Max.   : 639.0000   Max.   : 649.000   Max.   :749.710   Max.   :2341.00
                  NA's   :1600                       NA's   :1598                            NA's   :1600      NA's   :3199
     CC2             CC3               CC4                CC5               Base.Time       Post.Length       Post.Share.Count
Min.   :  0.00   Min.   :  0.00   Min.   :   0.00   Min.   :-1366.000   Min.   : 0.00   Min.   :   0.0   Min.   :     1.0
1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:   2.00   1st Qu.:   -6.000   1st Qu.:17.00   1st Qu.:  38.0   1st Qu.:     2.0
 Post.Promotion.Status    H.local       Post.published.weekday Base.DateTime.weekday Target.Variable
 Min.   :0              Min.   : 1.00   Length:32759           Length:32759          Min.   :   0.000
 1st Qu.:0              1st Qu.:24.00   Class :character       Class :character      1st Qu.:   0.000
 Median :0              Median :24.00   Mode  :character       Mode  :character      Median :   0.000
 Mean   :0              Mean   :23.77                                                Mean   :   7.304
 3rd Qu.:0              3rd Qu.:24.00                                                3rd Qu.:   3.000
 Max.   :0              Max.   :24.00                                                Max.   :1305.000
```

• With the above output we can infer that there are missing values present in Page Category , Page likes, Page Check ins, Page talking about, CC1- CC5 , Feature 27, Feature 29, Feature 25, Feature 20, Feature 22, Feature 18, Feature 10, Feature 13, Feature 7, Feature 15.

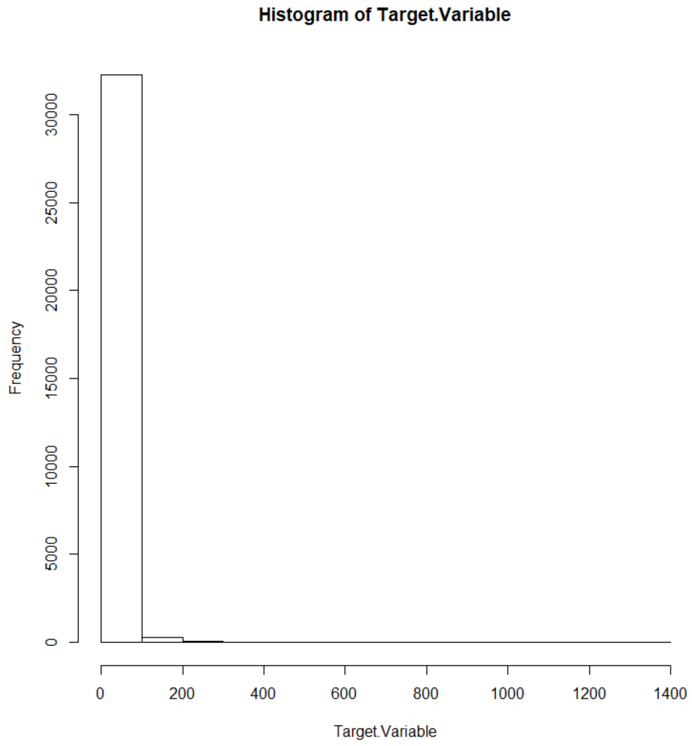
• There are few variables of which the maximum value is very high compared to the 3rd quartile. For example, Post share count , post length , values of CC1- CC5 etc. Hence there might be outliers in those variables.

Exploratory data analysis



a. Univariate analysis of the variables – Since there are multiple independent variables creating histograms and box plot of the variables and drawing inferences in order to understand the outliers.

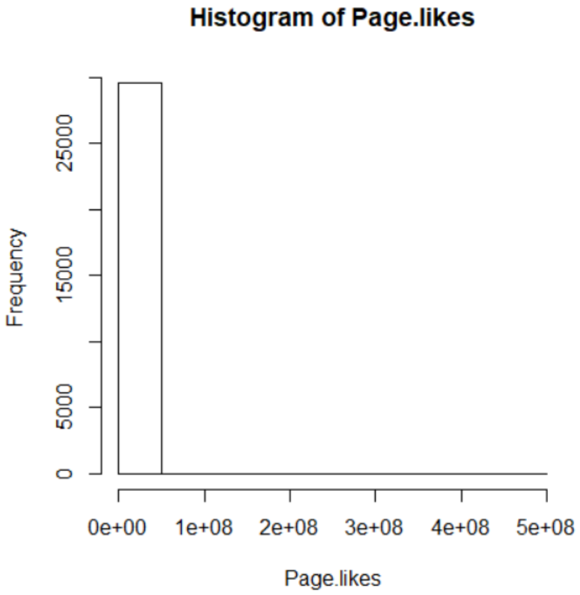
1. Histogram of Target Variable (Dependent Variable)



The target variable is right skewed and only 3 values are to the right. There are outliers in the dataset. Most of the values are between 0-300.

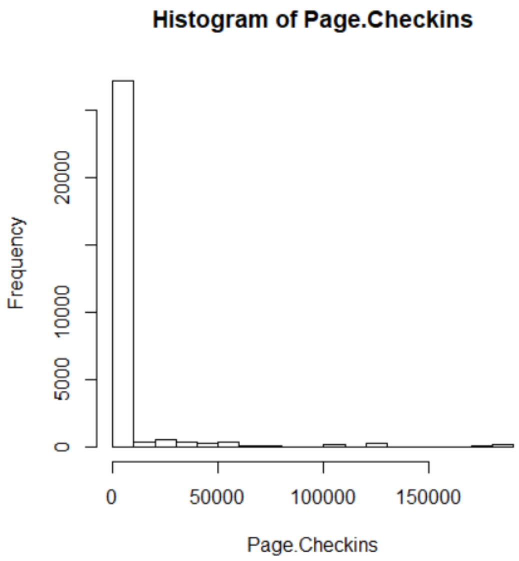
Plotting the independent variables.

Distribution of Page Likes.



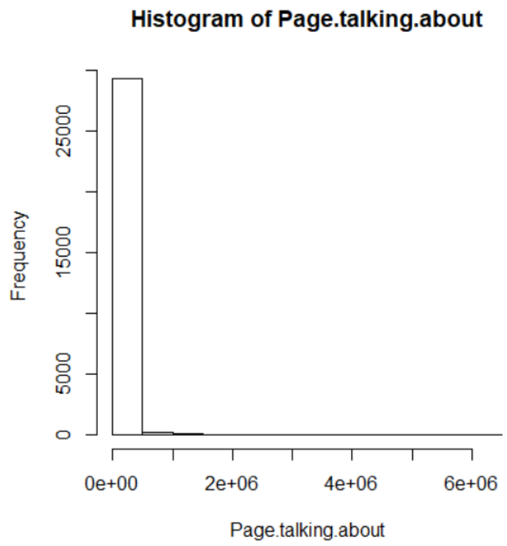
The data is right skewed with some outliers.

Distribution of Page Check ins.



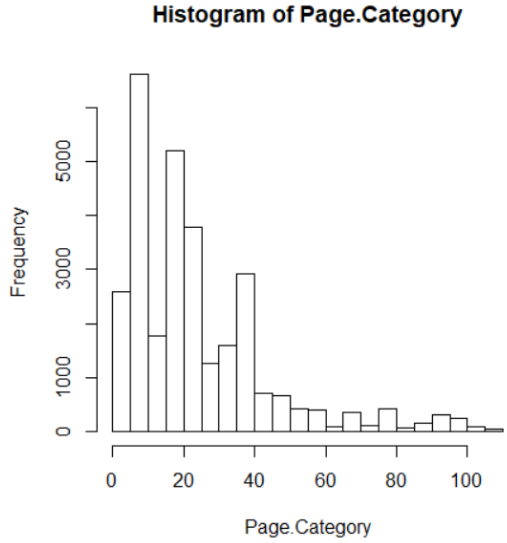
The graph clearly shows that most of the values are zero, which means majority of the users are not using this feature , they do not check in on Facebook when they are visiting new places. The data is right skewed with possible outliers.

Distribution of page talking about



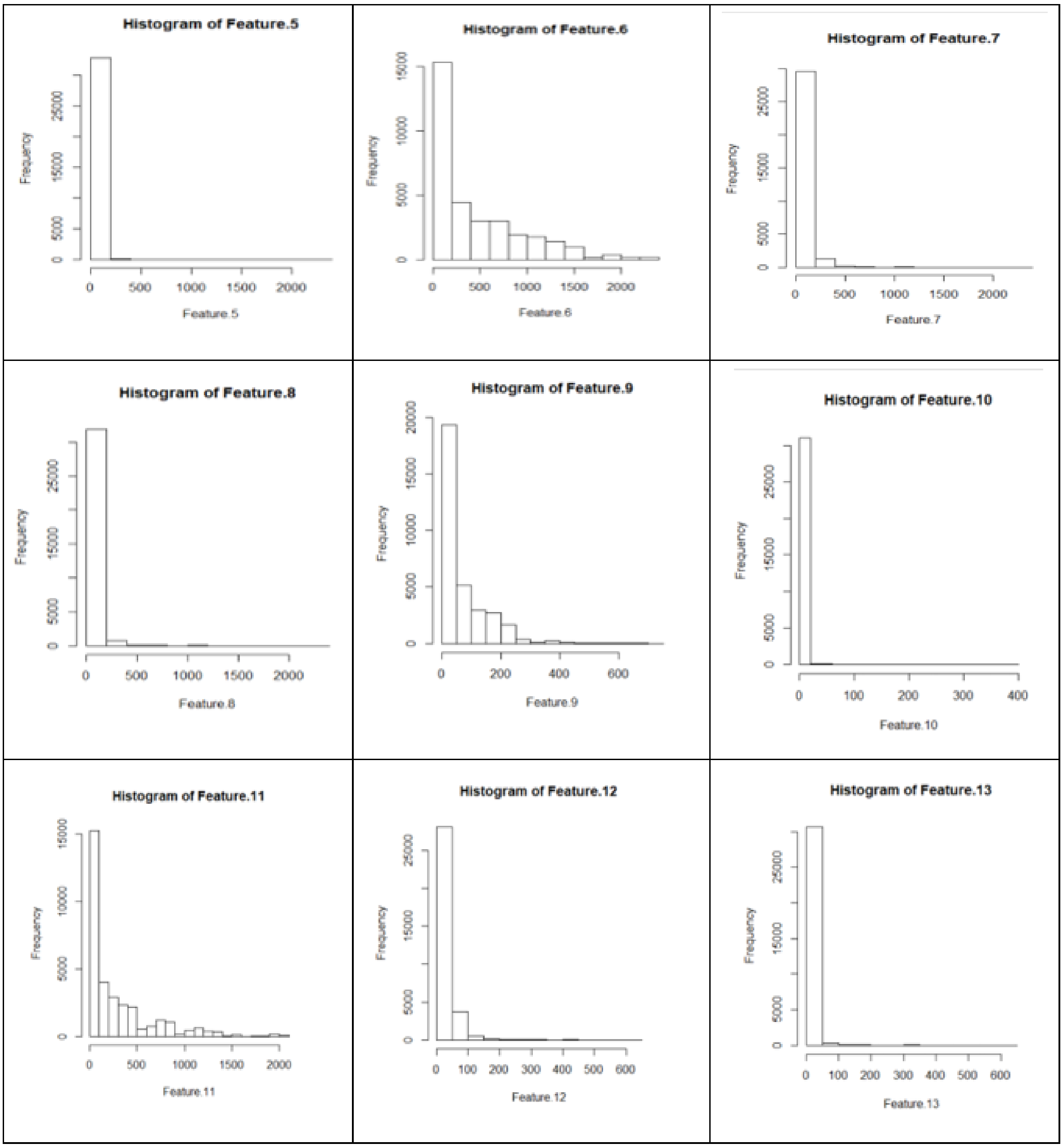
The data is right skewed with possible outliers.

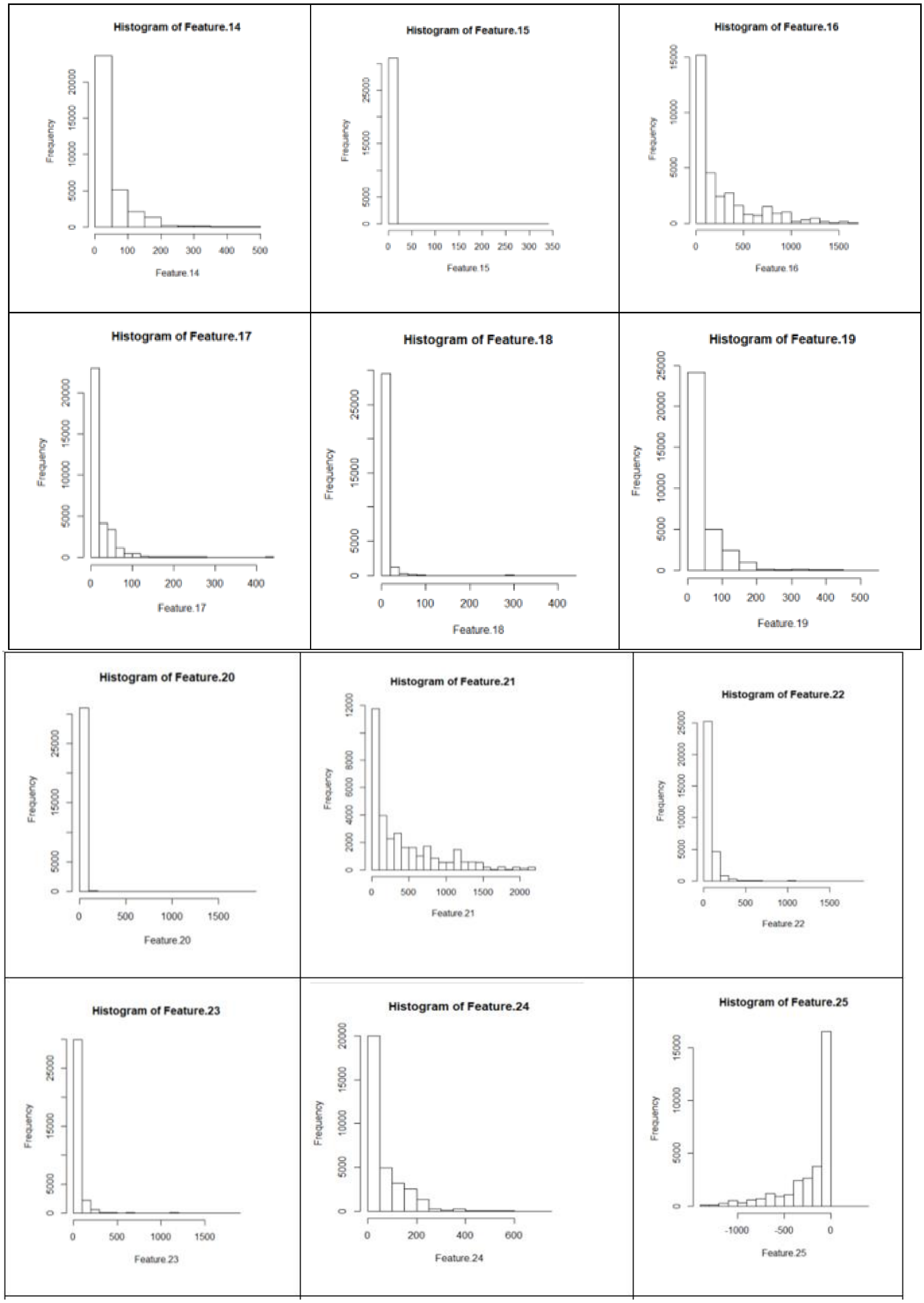
Distribution of Page Category



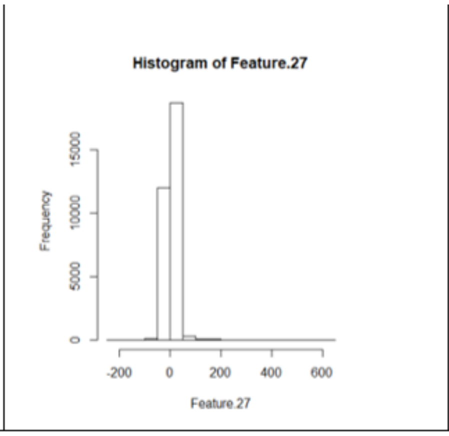
Page 0-20 have the maximum records in the dataset.

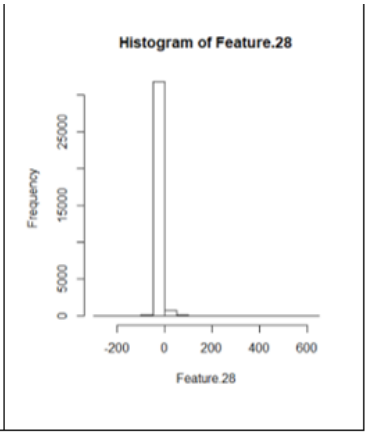
Distribution of all the features given in the dataset.

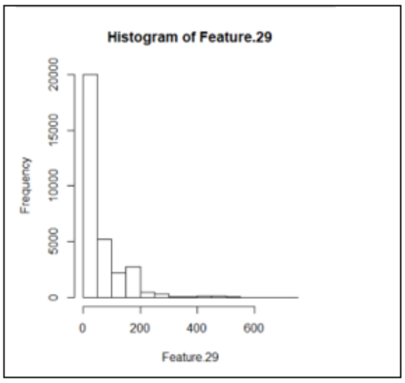




Histogram of Feature.14
Frequency
Feature.14
Histogram of Feature.15
Frequency
Feature.15
Histogram of Feature.16
Frequency
Feature.16
Histogram of Feature.17
Frequency
Feature.17
Histogram of Feature.18
Frequency
Feature.18
Histogram of Feature.19
Frequency
Feature.19
Histogram of Feature.20
Frequency
Feature.20
Histogram of Feature.21
Frequency
Feature.21
Histogram of Feature.22
Frequency
Feature.22
Histogram of Feature.23
Frequency
Feature.23
Histogram of Feature.24
Frequency
Feature.24
Histogram of Feature.25
Frequency
Feature.25



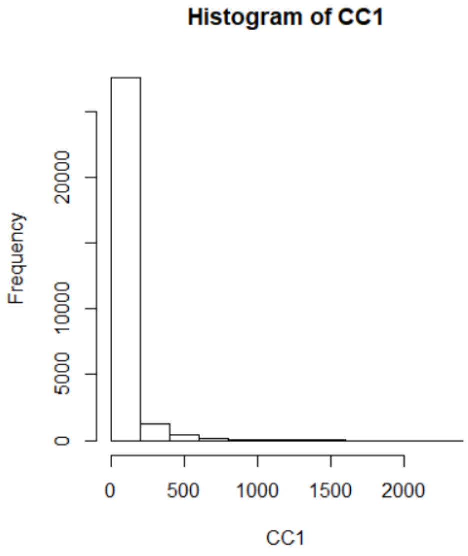




All the Features from 5 to feature 29 are aggregated by page, by calculating min, max, average median, and standard deviation of essential features. All these features are rightly skewed.

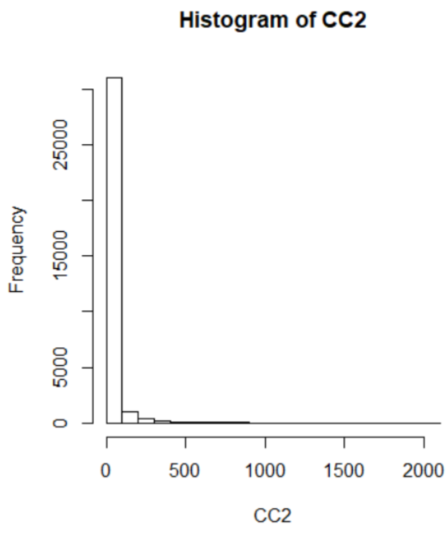
Distribution of CC1 to CC5

- CC1



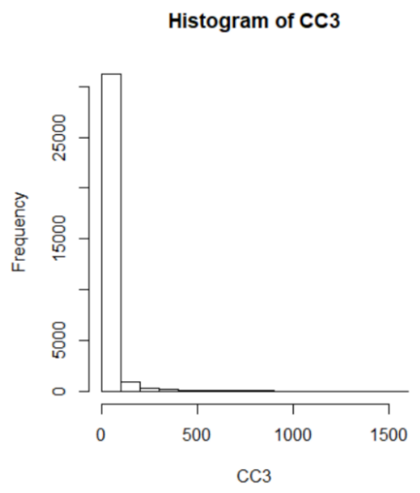
Most of the values are between 0 to 500.

- CC2



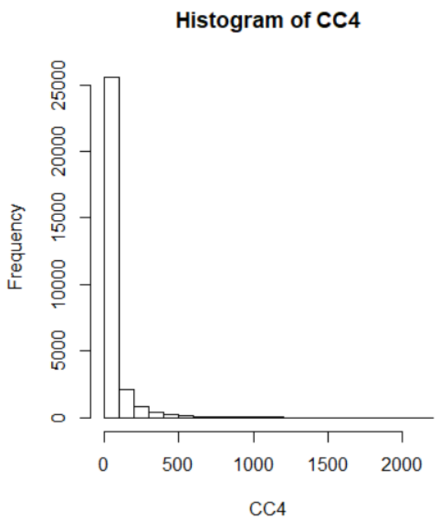
The records commented between the last 24 hours relative to base date/time is between 0 to 550. Data right skewed.

- CC3



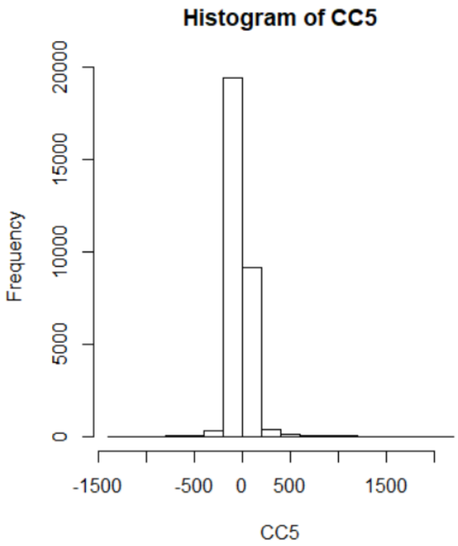
The data sows the number of comments in last 48 to 24 hours relative to base time. Most of the comments are 0 for this distribution as well

- CC4



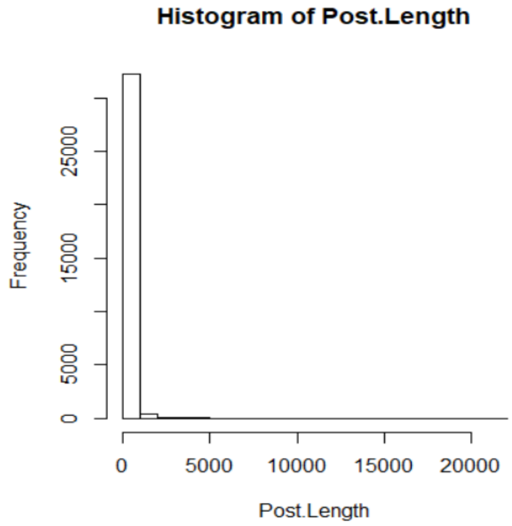
The number of comments in the first 24 hours after the publication of post but before base date/time.

- CC5



This variable shows the difference between the CC2 and CC3 variable. The plot displays the leptokurtic distribution.

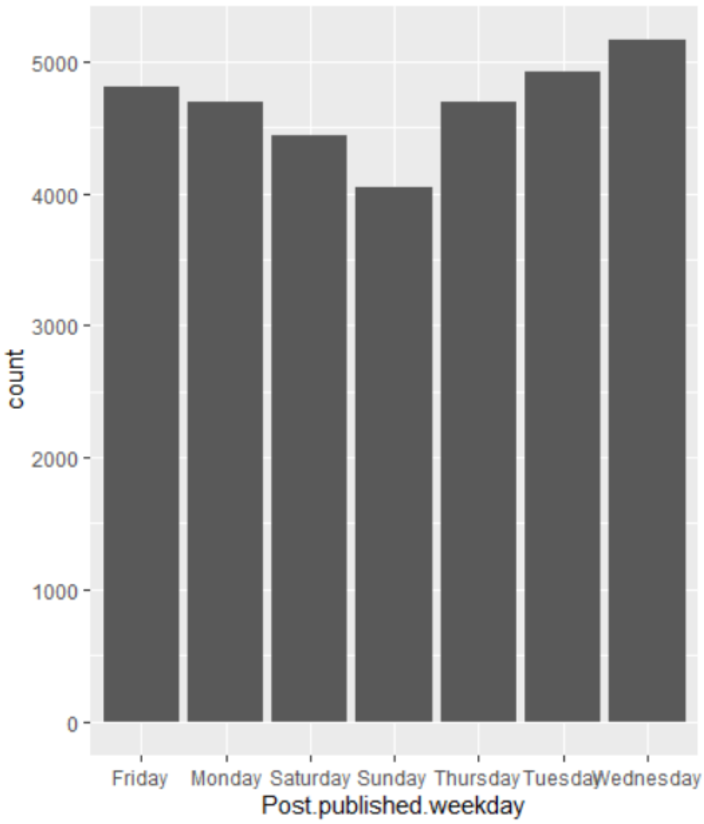
Distribution of Post length



Data is rightly skewed.

1.2Categorical variables.

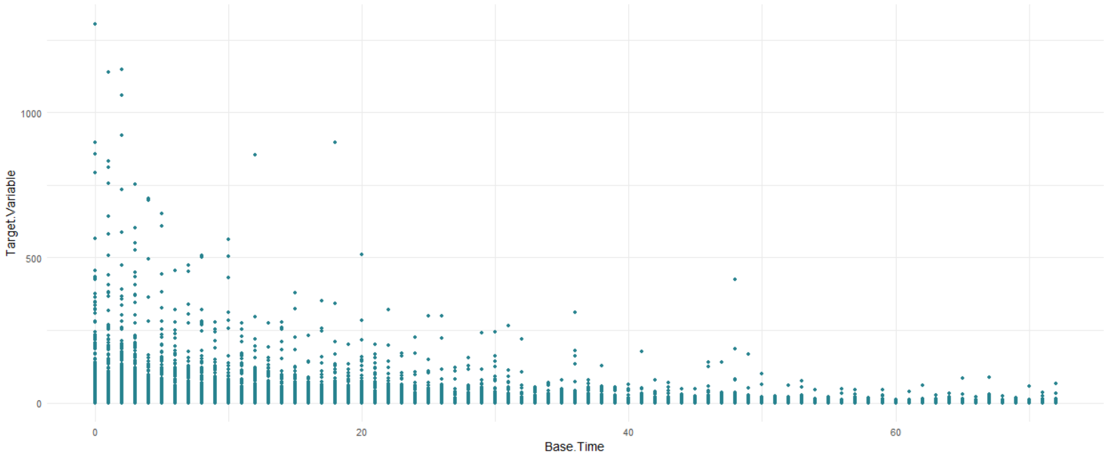
Post Published Weekday



Maximum number of post are made on Wednesday. Least number of post are made on Sunday.

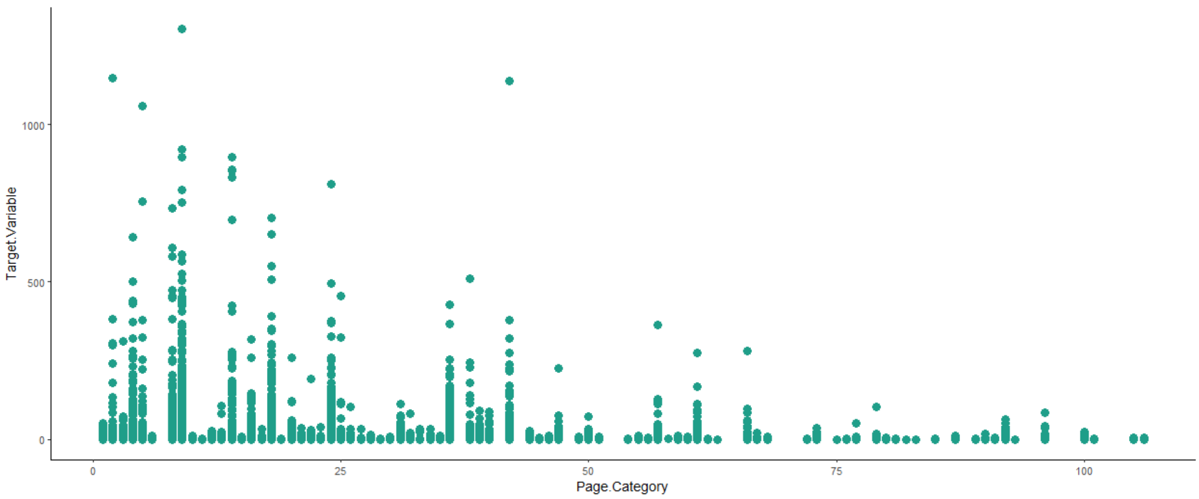
2. Exploring the relationship of target variable with other independent variables.

   a. Relationship between Target Variable and Base time.

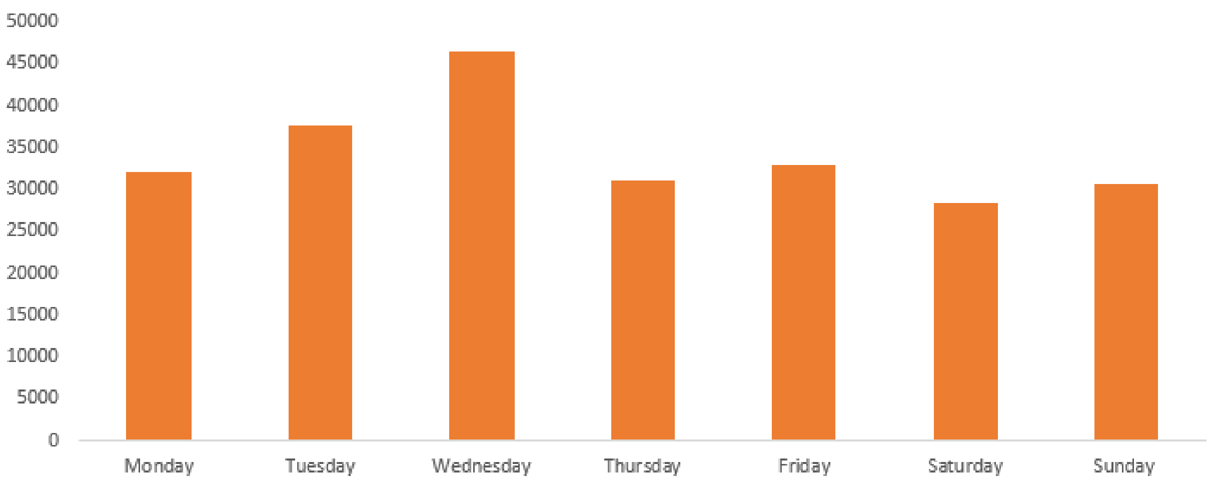


As we can infer from the graph, that most of the comments are received in the first 20 hours of the post, as the number of hour increases the comments gradually decreases.

b. Relationship between target variable and page category.

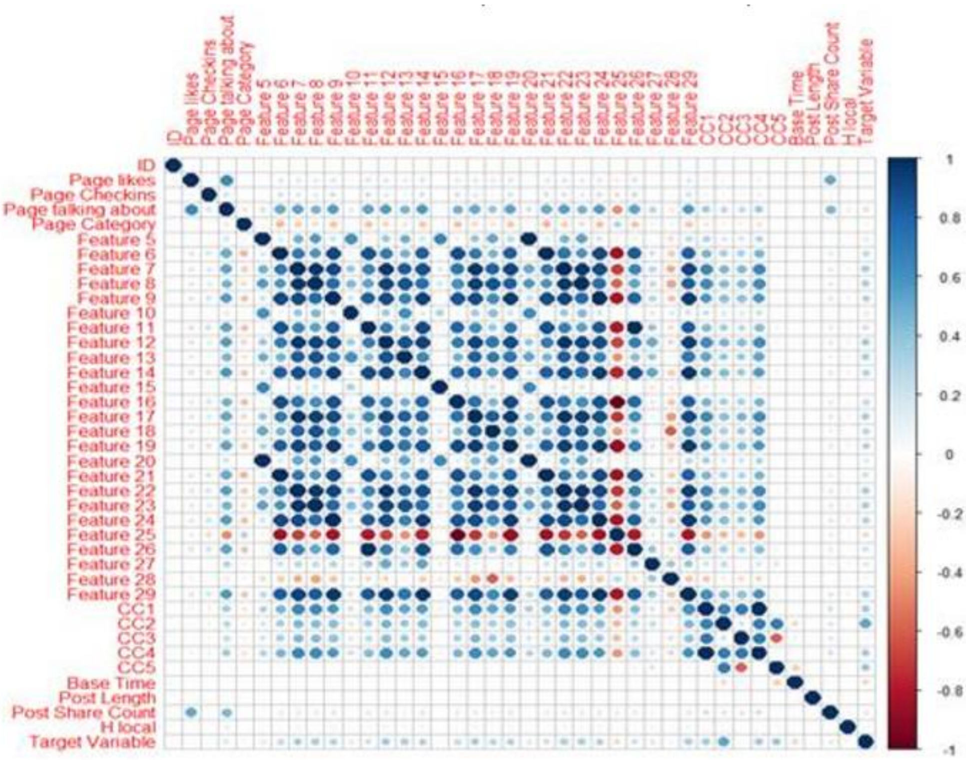


c. Relationship between the Comments and the Base date time weekday.



We can see that when a Post is published on weekday variable is compared with Target Variable, the frequency of post increases daily and it reaches its maximum point on Wednesday and then it declines gradually.

Correlation Heat Map



- Target Variable is highly correlated with CC2 variables.
- Post Share Count variable is highly correlated with Page Likes and Page Talking about.
- CC5 Variable is positively correlated with CC2 variable and negatively correlated variable.
- All the Derived Features (Feature 5 – Feature 29) are highly correlated with each other except for Feature 25 which is negatively correlated with other derived features.

Missing Values in the dataset:

When computed for missing values, the maximum percentage of missing value is 18.60%. Since in the entire dataset, there is no observation which has missing value of more than 18%. Taking 10% as the base, there are 246 observations which have more than 10% of missing value, and removing them from the dataset.

For using the dataset , the new data set , data set 2 , still has few missing observations. Hence have run the model through 'mice' function, and imputed all the necessary missing values.

Removal of unwanted variables.

In the given dataset , ID column is a nominal variable. It is not of any relevance for the analysis/ prediction of user comment volume, hence can be dropped. Also, the column post promotion status does not have any values. Throughout the dataset it has zero value, hence could be dropped as well.

Additional Insights :

All the feature variables, are highly correlated with each other ,except of feature 25 , which is negatively correlated. Page Category 9 has received the highest frequency of comments. Also, when comparing all the CC1- CC5 , the posts have received maximum comments in the initial 24 hours , and then has been gradually decreasing.

Most of the independent variables are right skewed. After performing clustering and PCA , we would get better insights on the variables of more importance.

Also, Regression will provide with the variables which have significance with respect to the target variable.

Model Building and Interpretation:

Preparing the data.

1. Transforming categorical variables to factor variable :

The 2 variables Post.published.weekday and Base.DateTime.weekday are categorical variables. Hence converting them into factor variables.

2. Eliminating the unwanted variables: As stated above, ID column and promotion status does not have any data of relevance to the model, hence removing them.

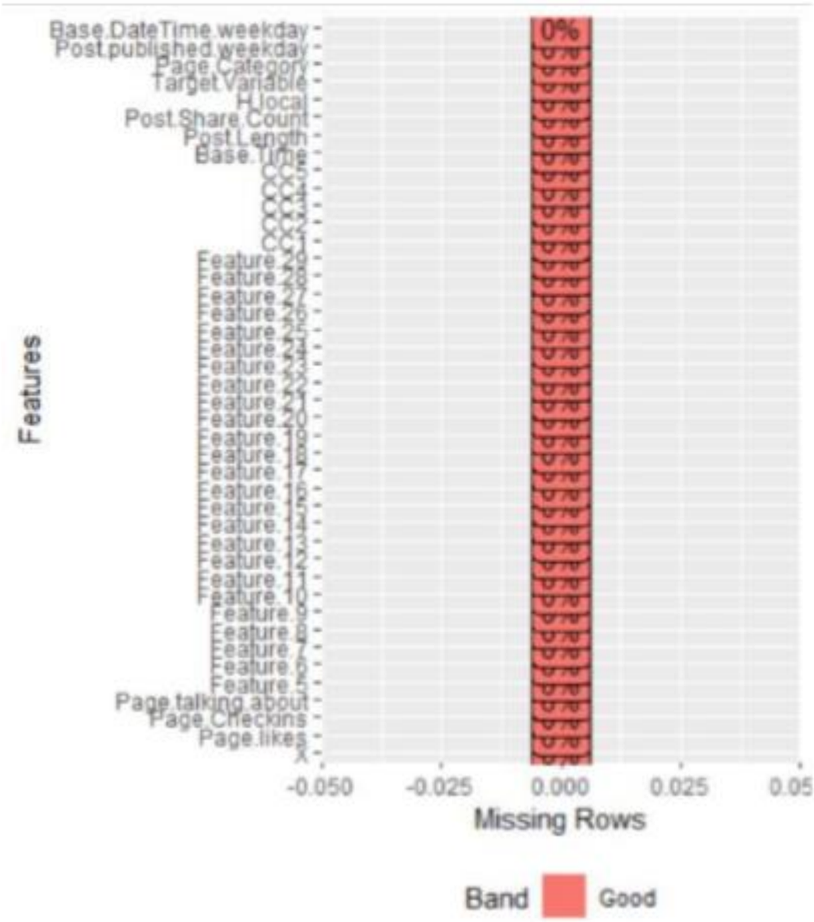
3. Imputation of missing value in the dataset:

There are several variables in the dataset that has missing values.

• In the given dataset CC1 , CC4 , CC5 , Feature 7 , Feature 10 , Feature 13 , Feature 15 , Feature 18 , Feature 20 , Feature 22 , Feature 25 , Feature 27 , Feature 29 all these variables have missing values.

• Hence using VIM library and K- nearest neighbour imputation function with K = 10 (using the nearest neighbour) in all the above mentioned variables.

• After this treatment there are not any missing values for the dataset.



Missing value Imputation plot

4. Splitting the data set into training and testset.

a. The dataset is split into 70:30 ratio. The Trainset consist of 70% of the data and hence 23033 observations. b. The test set comprises of the rest 30% of the data i.e 9726 observations.

5. Eliminating the outliers. As established in the EDA phase, the data set had multiple outliers, hence capping out the outliers from both the dataset. I am capping the values from both the datasets on the values 0.05th percentile and 95th percentile.

6. Pre-processing the data To scale the data and transform the non-normal dependent in a normal shape, I have scaled the data and used box cox transformation in the early stage of model building. This will pre-process the data.

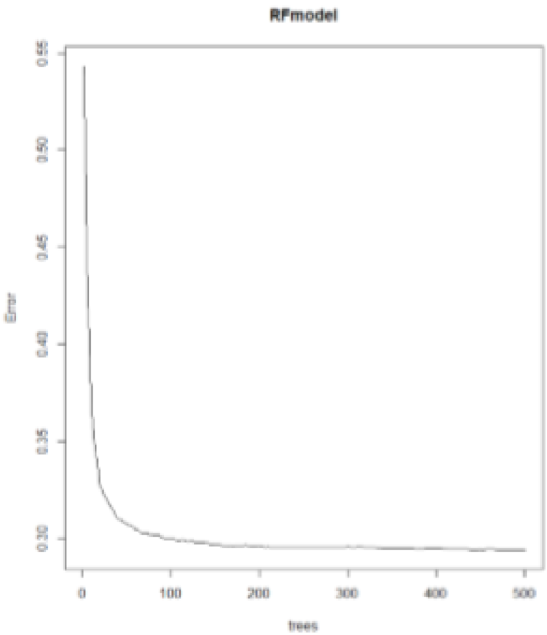
```
> process
Created from 32759 samples and 41 variables

Pre-processing:
  - Box-Cox transformation (3)
  - centered (39)
  - ignored (2)
  - scaled (39)

Lambda estimates for Box-Cox transformation:
0.1, 0.4, -0.1
```

7. Storing the predicted values Since we do not have 2 separate datasets, we had split the data into training and test sets. I have predicted the values of train and test dataset separately and stored them in a separate vector.

Model building:

1. Using Random Forest, linear model, SVM and extreme gradient boosting algorithms to build the model.

Random Forest model:



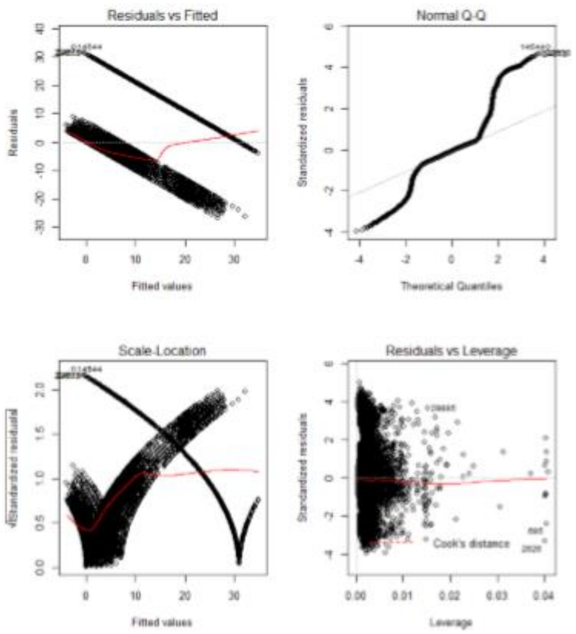

```
Call:
 randomForest(formula = Target.Variable ~ ., data = train1, mtry = 3,      nodesize = 10, ntree = 501, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 501
No. of variables tried at each split: 3

          Mean of squared residuals: 0.2941712
                    % Var explained: 70.58
```

For feature selection we used Random forest model, and found that there were 14 independent variables and 1 dependent variable which will help us predict the dependent variable much better.

Those variables are Page likes, feature 9, feature 12, feature 13, feature 24, feature 27, feature 28, cc1, cc2, cc3, cc4, cc5, base time, post share count.

The random forest model is showing an RMSE score on Test model of 0.36 and MAPE of 0.27. The RMSE is very low of this model and wont be a good fit.

Linear Regression



- As we have established that the variables have a linear co-relationship, we will build a linear model first.
- We can see a lot of variance in the data.
- The RMSE value and mape both are very low, hence the model will need some tuning.
- There is a lot of variance in the data. The RMSE value of this model on test set is 9.03 and mape is 600.27 , which is not a good model. The model will need more tuning to be a better fit.

SVM

The SVM model gives RMSE value of 8.99 and the R2 value as 0.334 when run on a test set. Some model tuning would be required to make it a better fit without overfitting.

Extreme Gradient boosting

After predicting the results on test set of the data I got RMSE value of 6.3.

Model Tuning

a. Step wise regression of LM model

In order understand and choose the best variables that would be a great fit and significant to help us make the best predictions of the on the target variable we'll be be building a regression model.

After performing step wise regression, I selected the model with the least AIC value which had the following variables – CC1 , CC2,CC3,CC4,CC5, page likes, base time, post share count, features 12,23,24,18,9 and feature 13.

b. Performing cross validation

I also performed cross validation on linear model and tried if I get a better RMSE and MAPE value as compare to the previous one. After cross validation I have got RMSE of 0.63 and MAPE value of 1.83. Which means if the actual value of the dependent variable is 10, the predicted value would be 9 or 11, which is a great fir for the model.

```
Linear Regression

32759 samples
   40 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 29483, 29483, 29483, 29483, 29484, 29484, ...
Resampling results:

  RMSE       Rsquared  MAE
  0.6362891  0.595139  0.3894287

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
> mape(test1$Target.Variable , pred)
[1] 1.831751
> |
```

Insights from the analysis

- Post Share Count variable is highly correlated with Page Likes and Page Talking about. So, we can assume that more the people talk about or like the pages then Post Share Count will be higher
- Linear Regression model is the ideal fit since it has provided the best prediction accuracy when compared with other models after tuning.
- Page likes, Base time and Post share count are some of the most important variables.
- This model will help to get the idea of popularity of the topic before its publications

Recommendations

- There are certain variables that are highly correlated, should have been avoided while collecting data.
- Make use of significant variables for solving business problems.
- The business should use the pages with maximum likes variables and post counts for marketing or attracting more comments.
- Implementation of this model for marketing strategy.
- Use the significant variables to attract more traffic.

Appendix

Code for the project has been uploaded with this assignment.