Facebook Comment Volume Prediction

ABSTRACT

In the past decade, tremendous work and progress have been made towards online networking. Huge amount of data has been generated and hence pushed the necessity for the pragmatic examination of such data. In today's digital marketing era, social networking services have been playing a key role in brand building and customer communication for any small businesses as well as large corporations. The amount of data that has been uploaded to these social networking services is increasing rapidly. This paper demonstrates a preliminary work to exhibit the proficiency of machine learning model to predict the comment volume a post will receive in next H hours at a randomly selected base time.

INTRODUCTION

In the third quarter of 2012, the number of active Facebook users surpassed one billion, making it the first social network ever to do so (*Source* - Statista on Oct'19). With 2.41 billion monthly active users as on second quarter of 2019, Facebook is the biggest social network worldwide. Every 60 seconds 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded on Facebook *(Source*: The Social Skinny on Oct'19*)*. The advertising revenue of Facebook in the United States in 2018 stands up to $14.89 billion USD while $18.95 billion USD in all other countries combined. Latest research reports have indicated that user generated content on Facebook drives higher engagement than advertisement. The huge amount of data that gets added to the network can help us to understand the intricacies of user behavior and user engagement.

This paper is focused on leading Social Networking Application service Facebook, especially 'Facebook Pages', for automatic analysis of trends and patterns of users. Feature Selection has been concentrated more. The analysis is based on the comment volume prediction (CVP) that a page is expected to be received in next H hours. The goal is to predict how many comments a user generated post is expected to receive in the given set of hours.

Dataset Understanding

The crawled data has been extracted from Facebook. With respect to the specifics of the dataset, there is information for 32,759 Facebook pages, each of which contains 43 attributes out of which one is Target variable: number of comments received after publishing the post in H hours (H Local variable).

Features

The dataset includes a large number of features including but not limited to –

- Number of likes on the page
- Number of individuals who have visited the place (if the page corresponds to an institution, place etc.)
- Daily interest of page (measured by other comments, like, posts, shares in a given day)
- Page Category (Place, Brand, Institution etc.)
- Total number of comments before a given Date/Time
- The number of comments in the preceding 24 hours, and in preceding 48 to 24 hours
- Number of characters in the post and number of post shares
- Whether or not the post has been promoted and the day of the week on which the post was made.

1) ***Page Features / Page Likes*** – It is a feature that defines users support for specific comments, pictures, wall posts, statuses or pages. *Page Category* – This defines the category of source of document e.g. Local business or place, brand or product, company or institution, artist, brand, entertainment, community etc. *Page Check-in's* – It is an act of showing presence at place, and under the category of place/institution pages only. *Page Talking about* – This is the actual count of users that were engaged and interacting with the Facebook page.

2) ***Essential Features*** **–** This includes the pattern of comment on the post in various time intervals w.r.t the randomly selected base date/time.

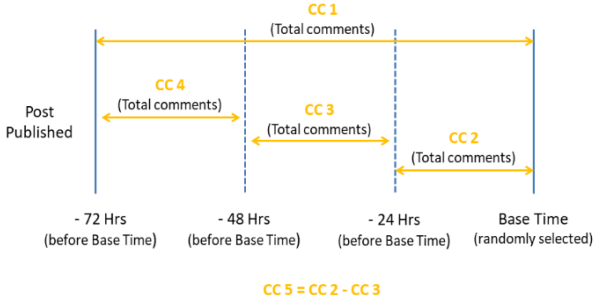


Fig 1: Essential Feature Details

CC1: Total comment count before selected base date/time. C*C2:* Comment count in last 24 hours w.r.t selected base date/time. C*C3:* Comment count in last 48 hours to last 24 hours w.r.t base date/time. C*C4:* 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. Furthermore, these features were aggregated by source and essential features were derived by calculating min, max, average, median and standard deviation of above 5 features.

3) ***Weekday Features*** – These are used to represent the day on which the post was published and the day on selected base date/time.

4) ***Other Features*** **–** This includes some document related features like length of posts, Post Share count, Post promotion Status (0,1), Post published date/time.

Initial Analysis

Univariate Analysis **-** Univariate analysis is the simplest form of analyzing data. It doesn't deal with causes or relationships (unlike regression) and its major purpose is to describe; it takes data, summarizes that data and finds patterns in the data. Here the page category has been considered as a Factor Variable. The Target variable is quantitative – discrete data type.

- Most of the comments ranged between 0-100 for target variables with many outliers. (Fig 2 and 3)

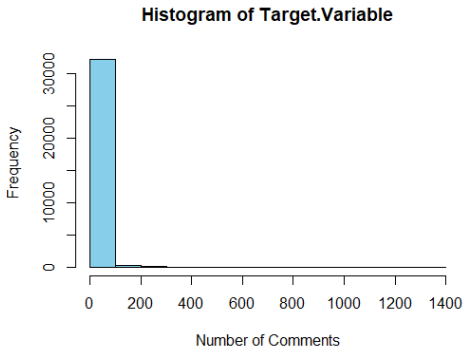
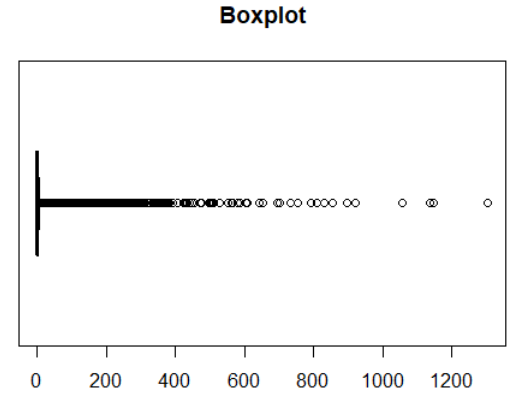


Fig 2




```
  Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
 0.000    0.000    0.000    7.304    3.000 1305.000        1
```

Fig 3

- The likes ranged between 36 and 48.69 Crores with many outliers. (Fig 4 and 5)

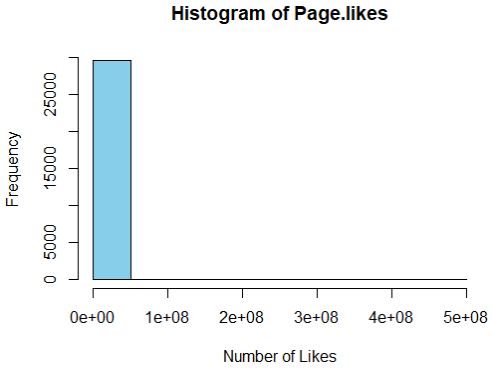
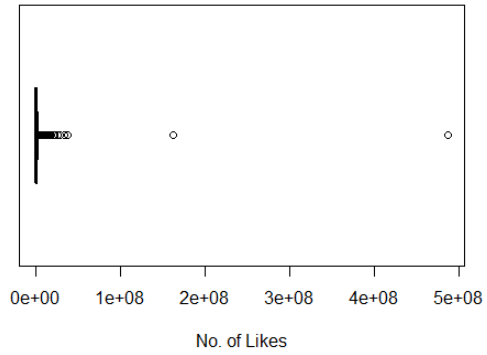


Fig 4




```
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
     36     35879    287698   1346069   1204214 486972297      3209
```

Fig 5

- The check-ins ranged between 0 and 10000 with many outliers. (Fig 6 and 7)

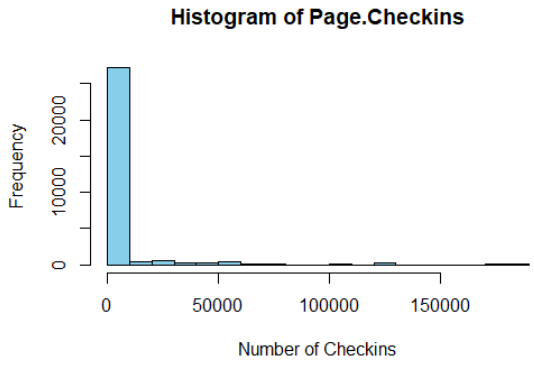
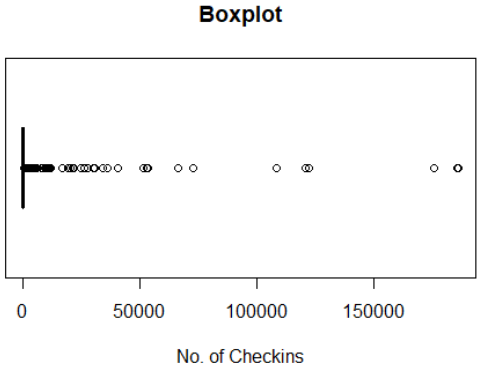


Fig 6




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0       0       0    4645      99  186370    3256
```

Fig 7

- Most of the revisits ranged between 0 and 500000 with many outliers. (Fig 8 and 9)

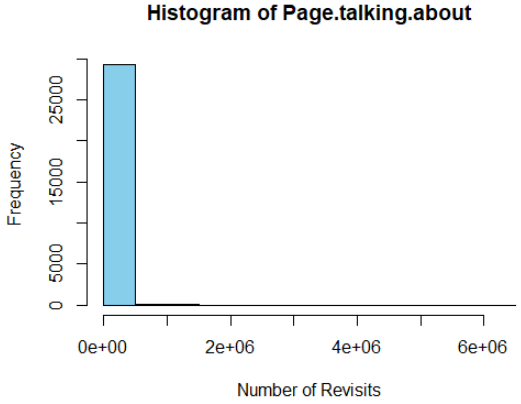
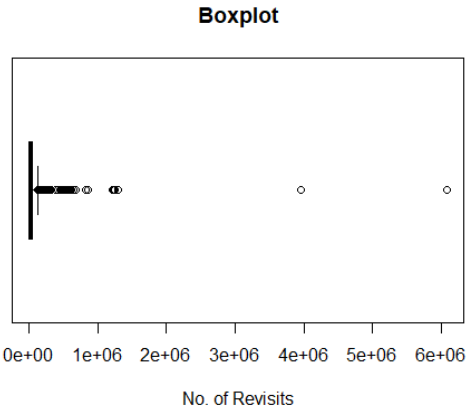


Fig 8




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0     698    6802   44913   50264 6089942    3256
```

Fig 9

- Most of the post shares are between 1 and 10000 with many outliers. (Fig 10 and 11)

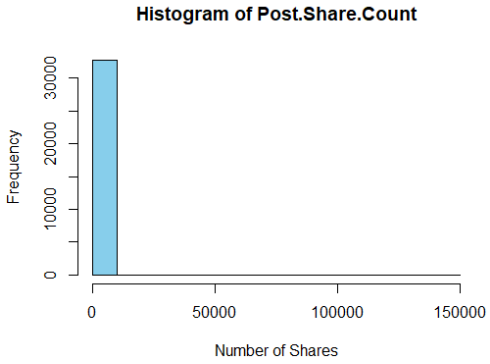
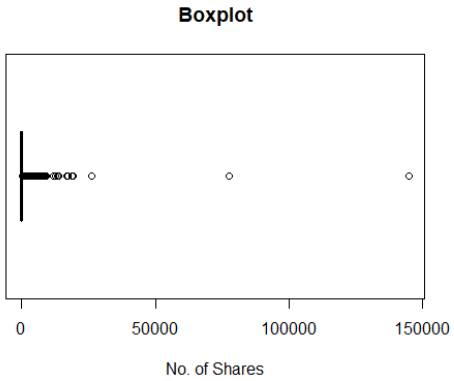


Fig 10




```
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
 1.0      2.0     13.0    120.1     61.0 144860.0        1
```

Fig 11

- Base time ranged between 0 and 72 hours with both Mean and Median as 35 Hours. No outliers

were observed. The character length of the posts ranged between 0 and 1000 with many outliers.

- Most of the time H local is 24 hours with some outliers.

- Page category 9 has maximum number of pages (5434 pages). (Fig 12)

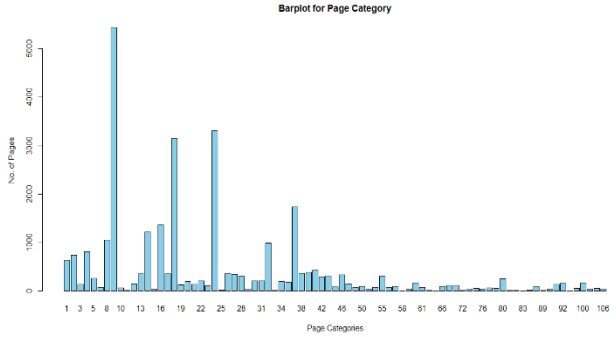


Fig 12

- Maximum number of pages has published their posts on Wednesday and minimum on Sunday. (Fig 13)

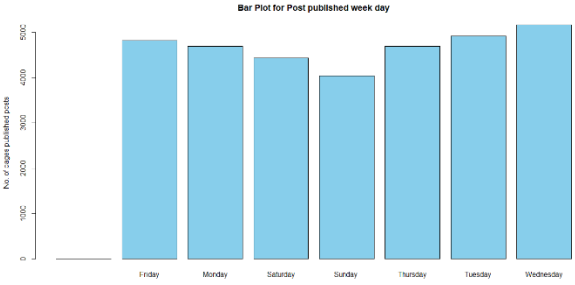


Fig 13

- For maximum number of pages selected base date time week day is Thursday with not much different from other week days.

Bivariate Analysis **-** Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, for the purpose of determining the empirical relationship between them.

Scatter plot is used to check the relationship between independent variables w.r.t Target variable.

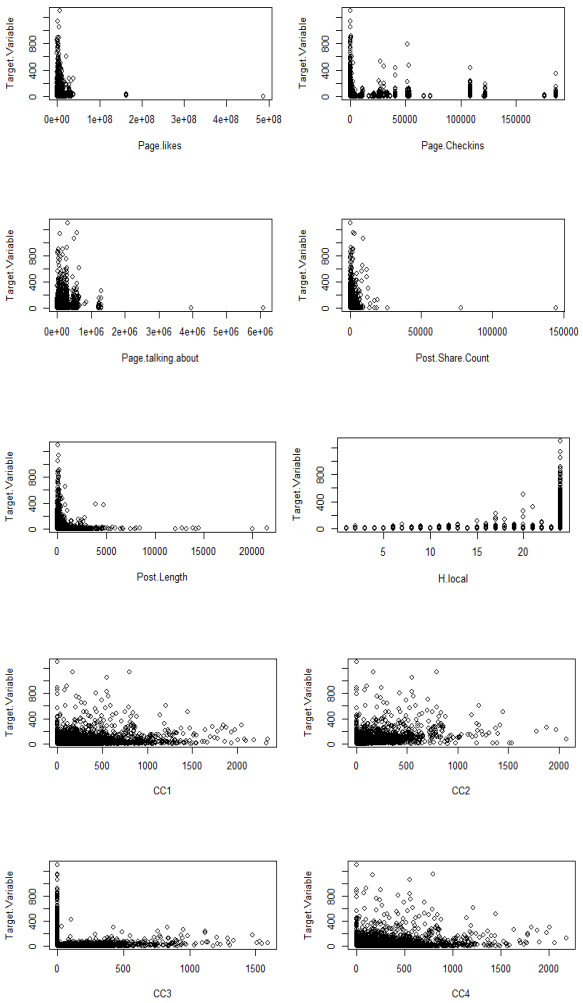


Fig 14

It has been observed that Target variable do not have any linear relationship with predictor variables. (Fig 14)

As data is non-Parametric, instead of using Pearson method, Spearman correlation method is used between Dependent and other Independent variables. (Fig 15)

Fig 15

Used Spearman rank correlation with the hypothesis at 95% confidence level:

- Null hypothesis - Spearman correlation coefficient, ρ ("rho"), is 0
- Alternate hypothesis – Spearman correlation coefficient, ρ ("rho"), is not equal to 0

This suggests that there is high positive correlation between Target variable and CC2. (Table 1.1)

Spearman Correlation (non-Parametric measure)				
Dependent Variable	**Independent Variable**	**Correlation**	**Positive / Negative**	**Statistical significance**
Target variable	Page Popularity/likes	0.3754545	Positive	Significant
Target variable	Page Check-ins	0.02445163	Positive	Significant
Target variable	Page talking about	0.4422047	Positive	Significant
Target variable	CC1	0.5326407	Positive	Significant
Target variable	CC2	0.7205032	Positive	Significant
Target variable	CC3	-0.001446182	Negative	Not Significant
Target variable	CC4	0.5399369	Positive	Significant
Target variable	CC5	0.2861564	Positive	Significant
Target variable	Base time	-0.4594163	Negative	Significant
Target variable	Post length	0.03714025	Positive	Significant
Target variable	Post Share Count	0.4656733	Positive	Significant
Target variable	H Local	-0.00361372	Negative	Not Significant

Table 1.1: Correlation test

Independent variables have high correlation with other independent variables. (Table 1.2)

Spearman Correlation (non-Parametric measure)				
Independent Variable	**Independent Variable**	**Correlation**	**Positive / Negative**	**Statistical significance**
Page Popularity/likes	Page talking about	0.81891372	Positive	Significant
Page talking about	CC1	0.65915288	Positive	Significant
Page talking about	CC4	0.65904373	Positive	Significant
CC1	CC2	0.73792109	Positive	Significant
CC1	CC4	0.99801049	Positive	Significant
CC1	Post Share Count	0.64702368	Positive	Significant
CC2	CC4	0.73935081	Positive	Significant
CC3	CC5	-0.841344315	Negative	Significant
CC4	Post Share Count	0.64446479	Positive	Significant

Table 1.2: Correlation test

Missing Value Treatment

Three variables were removed as they were not contributing any new or useful information in the dataset:

1. ID – unique page id's
2. CC5 – derived from other variables, CC2 & CC3
3. Post Promotion Status – all values were 0

Missing value calculation was performed post removing three variables which were not intended to be used for analysis.

Total missing values in 16 variables were 38136.

Page Popularity/likes -	3209
Page Checkins -	3256
Page talking about -	3256
Page Category -	3025
Feature 7 -	1680
Feature 10 -	1163
Feature 13 -	1644
Feature 15 -	1693
Feature 18 -	1606
Feature 20 -	1601
Feature 22 -	1602
Feature 25 -	1601
Feature 27 -	1599
Feature 29 -	1601
CC1 -	3200
CC4 -	3199

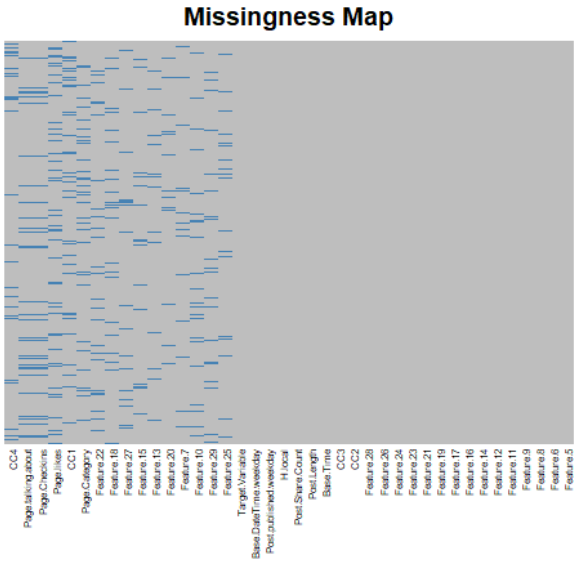


Fig 16: Missing Value Map

Missing value map as shown in Fig 16 showcased that wherever Page talking about value was missing, Page Checkins value was also missing in complete dataset.

Manual missing value imputation: CC1 and CC4 missing values with Base time 0 to 24 hours were imputed with the logic of CC1 = CC2 = CC4 and CC1 missing values with Base time between 25 to 48

hours is imputed with the logic of CC1 = CC2 + CC3. All other missing values were imputed using k-nearest neighbors algorithm (k-NN) with number of neighbors selected as 3 (k=3).

Noise and Outlier Treatment

The noises in the data were the extreme values found in 3 variables namely Page likes, Page talking about and Post length. Total of 13 rows were removed, wherever Page likes was more than 4 crore, Page talking about was more than 30 lakhs and Post length was more than 20 thousand characters.

For outlier treatment Tukey's rule was used which says that the outliers are values more than 1.5 times the interquartile range from the quartiles either below Q1 - 1.5IQR, or above Q3 + 1.5IQR (IQR: Inter Quartile Range). Tukey's method is not dependent on distribution of data and it ignores the mean and standard deviation, which are influenced by the extreme values (outliers).

Outlier Analysis of Page likes (Fig 17 & 18):
Outliers identified: 4047 from 32747 observations
Proportion (%) of outliers: 12.36%
Mean of the outliers: 6493157.03
Mean without removing outliers: 1220447.03
Mean if we remove outliers: 476939.81

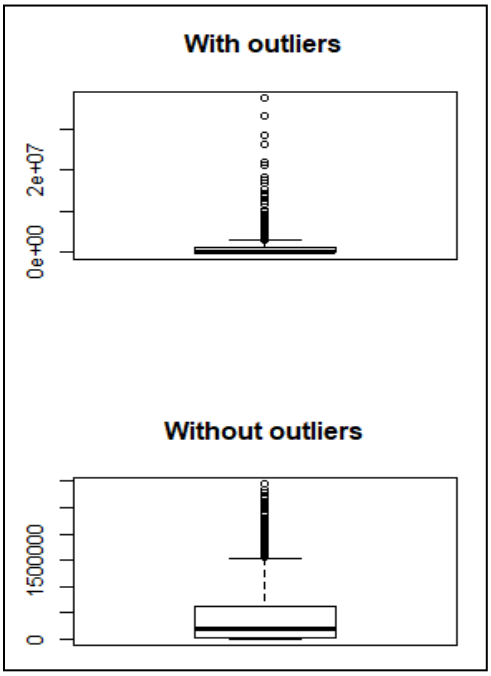


Fig 17: Boxplot for distribution of Page likes (With & Without outlier)

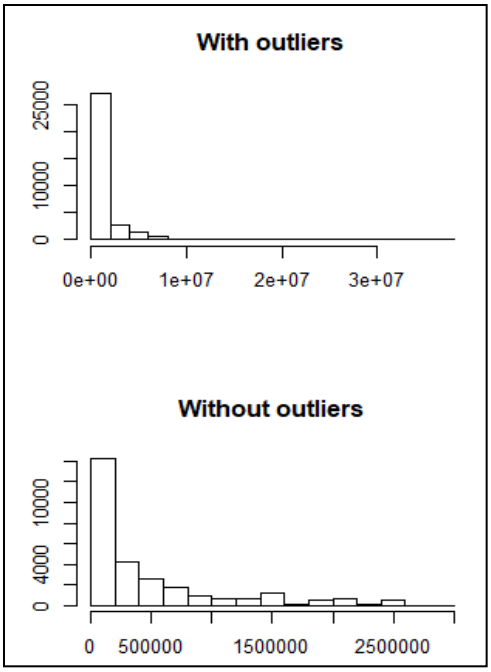


Fig 18: Histogram for distribution of Page likes (With & Without outlier)

Outlier Analysis of Page Checkins (Fig 19 & 20):
Outliers identified: 6419 from 32747 observations
Proportion (%) of outliers: 19.60%
Mean of the outliers: 23319.24
Mean without removing outliers: 4582.02
Mean if we remove outliers: 13.72

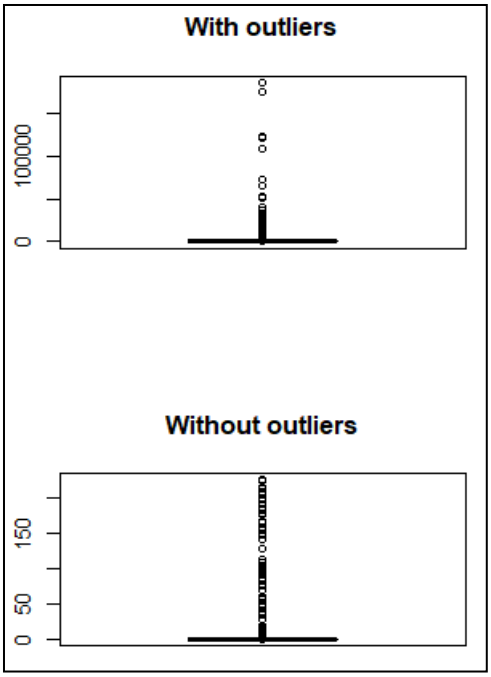


Fig 19: Boxplot for distribution of Page Checkins (With & Without outlier)

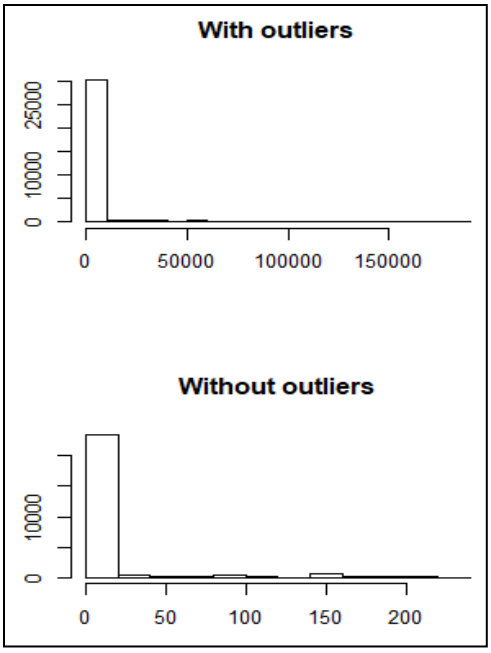


Fig 20: Histogram for distribution of Page Checkins (With & Without outlier)

Outlier Analysis of Page talking about (Fig 21 & 22):

Outliers identified: 2917 from 32747 observations
Proportion (%) of outliers: 8.91%
Mean of the outliers: 268130.91
Mean without removing outliers: 43975.97
Mean if we remove outliers: 22056.43

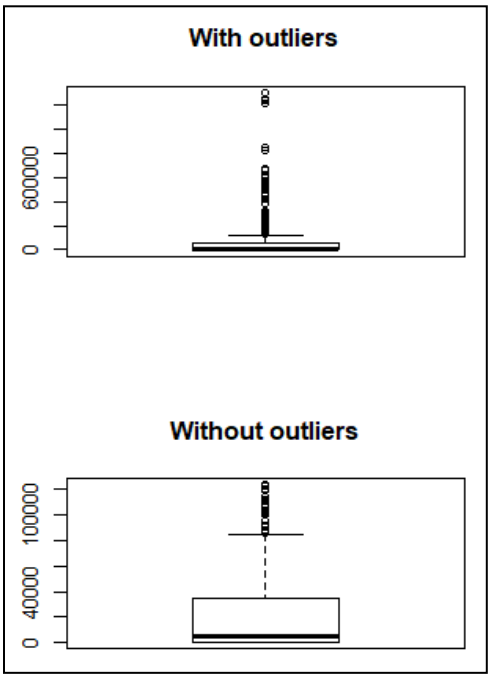


Fig 21: Boxplot for distribution of Page talking about (With & Without outlier)

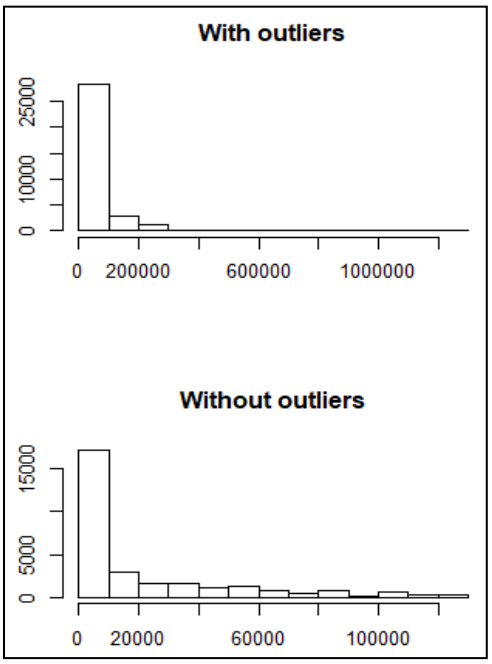


Fig 22: Histogram for distribution of Page talking about (With & Without outlier)

Outlier Analysis of Post Share Count (Fig 23 & 24):

Outliers identified: 4368 from 32747 observations
Proportion (%) of outliers: 13.34%
Mean of the outliers: 683.66
Mean without removing outliers: 111.39
Mean if we remove outliers: 23.31

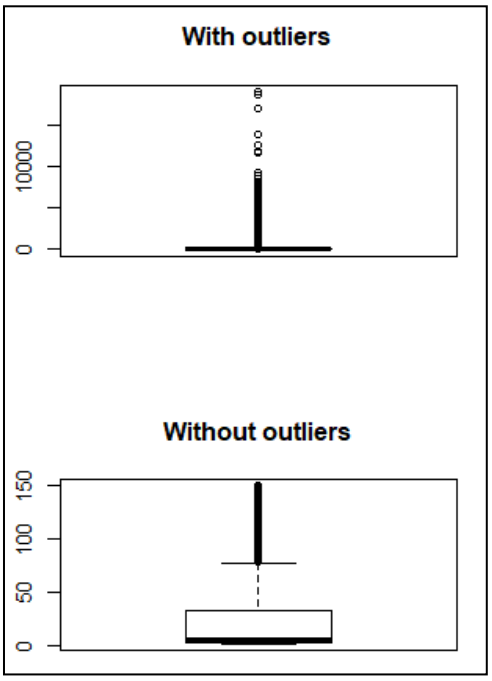


Fig 23: Boxplot for distribution of Post Share

Count (With & Without outlier)

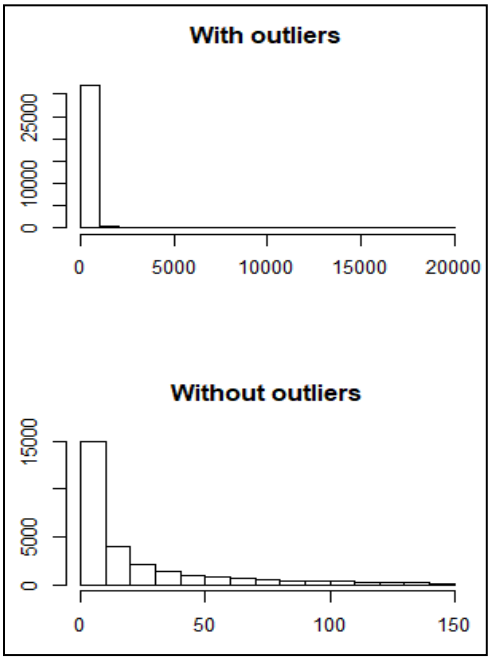


Fig 24: Histogram for distribution of Post Share Count (With & Without outlier)

Outlier Analysis of Post length (Fig 25 & 26):
Outliers identified: 2606 from 32747 observations
Proportion (%) of outliers: 7.96%
Mean of the outliers: 872.77
Mean without removing outliers: 162.99
Mean if we remove outliers: 101.63

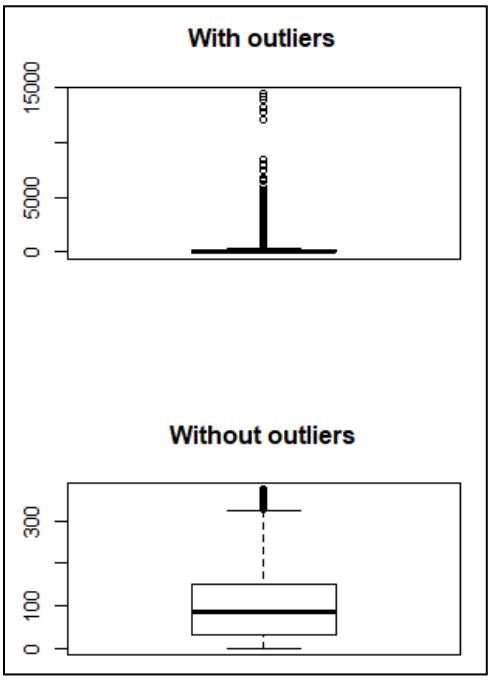


Fig 25: Boxplot for distribution of Post length

(With & Without outlier)

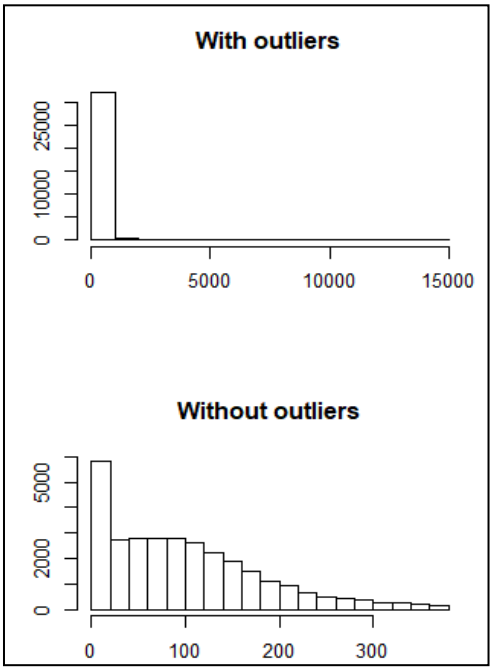


Fig 26: Histogram for distribution of Post length (With & Without outlier)

All the outliers were removed out of 32747 rows with out noise. 40% of rows were lost after outlier treatme nt and 19416 rows were retained.

Dimension Reduction

(For data without Outliers)

Factor Analysis was used to extract the factors from Feature 5 to Feature 29. Used Kaiser's rule which says to retain factors whose eigenvalues are greater than 1, Scree plot suggested to extract 4 factors (as shown in Fig 27).

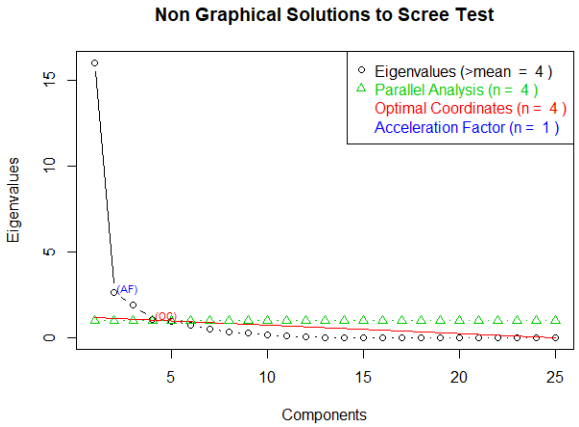


Fig 27: Scree Plot

Out of 29 variables, 4 factors were extracted using maximum likelihood factor analysis with varimax rotation, they were explaining 83% of variance.

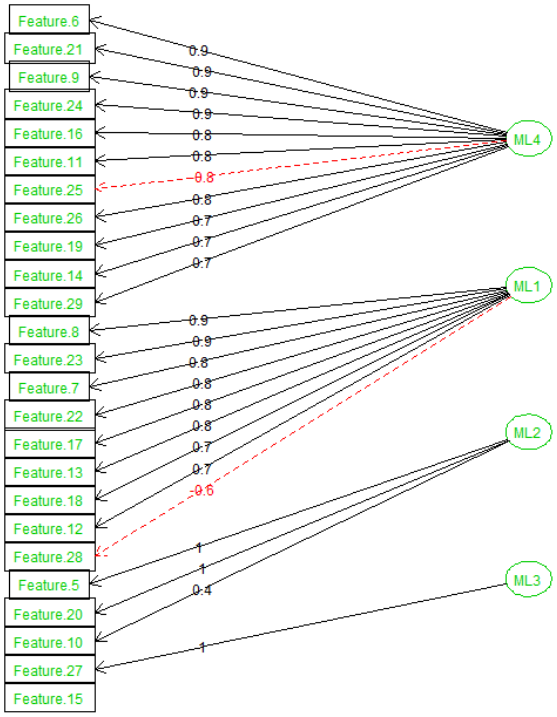


Fig 28: Factor Extraction Diagram

All the 4 Factors (as shown in Fig 28 and Table 2.1) were named as mentioned below:

ML1	ML2	ML3	ML4
Avg.Median	Min	Min.Avg	Max.SD
Feature 7	Feature 5	Feature 15	Feature 6
Feature 8	Feature 10	Feature 27	Feature 9
Feature 12	Feature 20		Feature 11
Feature 13			Feature 14
Feature 17			Feature 16
Feature 18			Feature 19
Feature 22			Feature 21
Feature 23			Feature 24
Feature 28			Feature 25
			Feature 26
			Feature 29

Table 2.1: Extracted Factors

Finding Important Variables

(For data without Outliers)

Extracted 4 factors were used and Linear Regression was done to find the important variables affecting the Target variable. With R square value of 33%, it suggested below 8 variables are important:

1. Max.SD
2. Avg.Median
3. CC2
4. CC3
5. Base time
6. Post length
7. Post Share Count
8. H Local

Dimension Reduction

(For data with Outliers)

Factor Analysis was used to extract the factors from Feature 5 to Feature 29. Used Kaiser's rule which says to retain factors whose eigenvalues are greater than 1, Scree plot suggested to extract 4 factors (as shown in Fig 29).

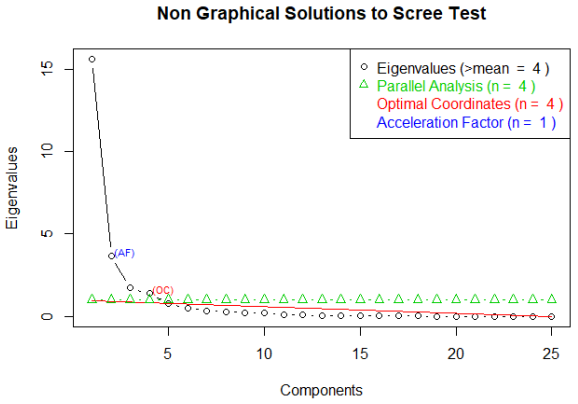


Fig 29: Scree Plot

Out of 29 variables, 4 factors were extracted using maximum likelihood factor analysis with varimax rotation, they were explaining 85% of variance.

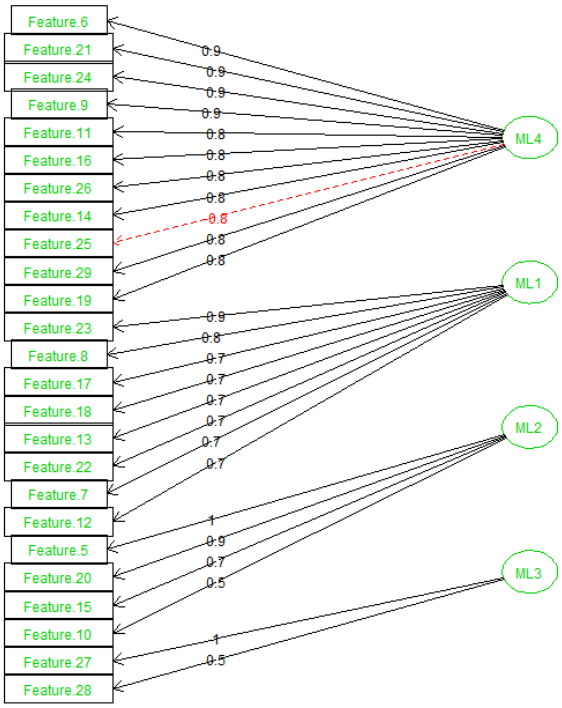


Fig 30: Factor Extraction Diagram

All the 4 Factors (as shown in Fig 30 and Table 2.2) were named as mentioned below:

ML1	ML2	ML3	ML4
Avg.Median1	Min	Avg.Median2	Max.SD
Feature 7	Feature 5	Feature 27	Feature 6
Feature 8	Feature 10	Feature 28	Feature 9
Feature 12	Feature 15		Feature 11
Feature 13	Feature 20		Feature 14
Feature 17			Feature 16
Feature 18			Feature 19
Feature 22			Feature 21
Feature 23			Feature 24
			Feature 25
			Feature 26
			Feature 29

Table 2.2: Extracted Factors

Finding Important Variables
(For data with Outliers)

Extracted 4 factors were used and Linear Regression was done to find the important variables affecting the Target variable. With R square value of 32%, it suggested below 11 variables are important:

1. Max SD
2. Avg.Median1
3. Min
4. Page likes
5. Page Checkins
6. CC2
7. CC3
8. CC4
9. Base time
10. Post Share Count
11. H Local

Modeling Process and Comparisons

After missing value imputation and outlier treatment, Feature 5 to Feature 29 were replaced by 4 extracted factors using Factor Analysis. 2 data sets were created for comparison and creating model: one with outlier and the other one without outlier.

For both data sets, train and test sets were divided in 70:30 ratio using important variables suggested by Linear Regression.

Random Forest
(For dataset with outliers)

Below 11 variables were used to build Random Forest model with different number of trees and suggested number of variables by 'tuneRF' function in each tree, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) was calculated post prediction to evaluate the model:

12. Max SD
13. Avg.Median1
14. Min
15. Page likes
16. Page Checkins
17. CC2
18. CC3
19. CC4
20. Base time
21. Post Share Count
22. H Local

Se no.	No. of Trees	No. of Variables in each tree	Dataset	RMSE	MAE	Variance explained	Most Important Variables
1	501	3	Training dataset	11.78	2.1	61.72%	Base Time
			Testing - dataset	20.16	3.9		CC2
2	601	3	Training dataset	11.76	2.1	61.95%	Base Time
			Testing - dataset	20.26	3.9		CC2
3	401	4	Training dataset	11.25	1.93	61.75%	Base Time
			Testing - dataset	20.38	3.9		CC2

Table 3: Random Forest results (data with outlier)

For the dataset with outlier 61.72 % of variance is explained by the Random Forest model with 501 trees and 3 numbers of variables in each tree, RMSE was 20.16 and MAE was 3.9.

61.95 % of variance is explained by the Random

Forest model with 601 trees and 3 numbers of variables in each tree, RMSE was 20.26 and MAE was 3.9.

61.75 % of variance is explained by the Random Forest model with 401 trees and 4 numbers of variables in each tree, RMSE was 20.38 and MAE was 3.9.

All models performed better in training dataset and most important variable suggested by the models was Base time followed by CC2.

Random Forest

(For dataset without outliers)

Below 8 variables were used to build Random Forest model with different number of trees and suggested number of variables by 'tuneRF' function in each tree, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) was calculated post prediction to evaluate the model:

9. Max.SD
10. Avg.Median
11. CC2
12. CC3
13. Base time
14. Post length
15. Post Share Count
16. H Local

Se no.	No. of Trees	No. of Variables in each tree	Dataset	RMSE	MAE	Variance explained	Most Important Variables
1	501	4	Training dataset	4.53	0.95	63.16%	CC2
			Testing - dataset	11.53	2.14		Base Time
2	601	3	Training dataset	4.77	0.99	63.05%	CC2
			Testing - dataset	11.63	2.14		Base Time
3	401	4	Training dataset	4.53	0.96	63.70%	CC2
			Testing - dataset	11.47	2.13		Base Time

Table 4: Random Forest results (data without outlier)

For the dataset without outlier 63.16 % of variance is explained by the Random Forest model with 501 trees and 4 numbers of variables in each tree, RMSE was 11.53 and MAE was 2.14.

63.05 % of variance is explained by the Random Forest model with 601 trees and 3 numbers of variables in each tree, RMSE was 11.63 and MAE was 2.14.

63.7 % of variance is explained by the Random Forest model with 401 trees and 4 numbers of variables in each tree, RMSE was 11.47 and MAE was 2.13.

All models performed better in training dataset and most important variable suggested by the models was CC2 followed by Base time as shown in Fig 31.

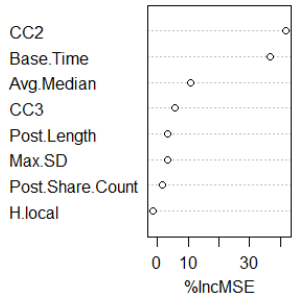


Fig 31: Random Forest Variable Importance (data without outlier, no. of trees – 401, variable in each tree - 4)

Among all the models of Random Forest best model was, for the dataset without outlier explaining 63.26 % of variance with 401 trees and 4 numbers of variables in each tree, RMSE was 11.4 and MAE was 2.13. CC2 was the most important variable for the number of comments prediction (Target Variable).

Artificial Neural Network

(For dataset with outliers)

Below 11 variables were used to build Artificial Neural Network model with 1 and 2 number of neurons in 1 layer, RMSE, MAE and R square was calculated post prediction to evaluate the model:

1. Max SD
2. Avg.Median1
3. Min
4. Page likes
5. Page Checkins
6. CC2
7. CC3
8. CC4
9. Base time
10. Post Share Count
11. H Local

Se no.	No. of Layers	No. of Neurons	Dataset	RMSE	MAE	R square
1	1	1	Training dataset	26.3	6.9	48.00%
			Testing - dataset	23.8	6.6	
2	1	2	Training dataset	23.1	5.6	52.00%
			Testing - dataset	20.2	5.6	

Table 5: Neural Network results (data with outlier)

For the dataset with outlier 48% was the R square value by ANN model with 1 layer and 1 neuron, RMSE was 23.8 and MAE was 6.6. Model performed

badly in training data set as RMSE and MAE was higher than testing dataset.

52% was the R square value by ANN model with 1 layer and 2 neurons, RMSE was 20.2 and MAE was 5.6. Model performed badly in training data set as RMSE was higher and MAE was same compare to testing dataset as shown in Table 5.

Artificial Neural Network

(For dataset without outliers)

Below 8 variables were used to build Artificial Neural Network model with 1 and 2 number of neurons in 1 layer, Feedforward neural network type was used with hyperbolic tangent function as activation function, RMSE, MAE and R square was calculated post prediction to evaluate the model:

1. Max.SD
2. Avg.Median
3. CC2
4. CC3
5. Base time
6. Post length
7. Post Share Count
8. H Local

Se no.	No. of Layers	No. of Neurons	Dataset	RMSE	MAE	R square
1	1	1	Training dataset	11.2	3.24	50.20%
			Testing - dataset	12.52	3.24	
2	1	2	Training dataset	9.9	2.4	60.20%
			Testing - dataset	11.2	2.41	

Table 6: Neural Network results (data without outlier)

For the dataset without outlier 50.2% was the R square value by ANN model with 1 layer and 1 neuron, RMSE was 12.52 and MAE was 3.24. Model performed slightly better in training data set as RMSE was lower and MAE was same compare to testing dataset.

60% was the R square value by ANN model with 1 layer and 2 neurons, RMSE was 11.2 and MAE was 2.41. Model performed slightly better in training data set as RMSE was lower and MAE was almost same compare to testing dataset.

Ensemble Technique

(Stacked Generalization)

To use stack generalization ensemble technique, modeldata without outlier was divided in to train and test dataset in a ratio of 70:30. Train dataset was then used to build 3 models using Random Forest, Classification and Regression Technique (CART) and K nearest neighbor (Knn).

Post creating 3 models, prediction was done on the train data itself and 3 different predictions were received, new dataset was created and saved as 'newtraindata' with independent variable as 3 predicted data sets and dependent variable as the target variable of train data. (Fig 32)

After training three models using train data, prediction was made using test data and 3 different test predictions were received, new dataset was created and saved as 'newtestdata' with independent variable as 3 different predicted data sets and dependent variable as the target variable of test data.

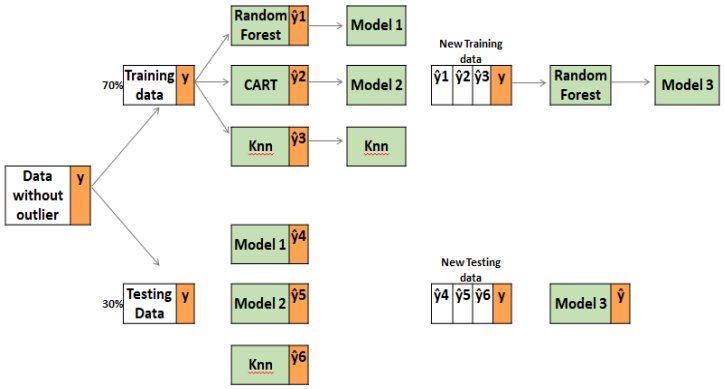


Fig 32: Ensemble technique (Stacked generalization)

Random Forest model with 501 trees and 1 variable in each tree was used to build the model using 'newtraindata' dataset, this model was then used for prediction on 'newtestdata' data set. Model had explained 93.6% of variance with RMSE of 6.4 and MAE of 0.81. Model also performed well in training data set with RMSE of 2.17 and MAE of 0.37. Significant improvement was achieved after doing stacked generalization as shown in Table 7.

Se no.	No. of Trees	No. of Variables in each tree	Dataset (without outlier)	RMSE	MAE	Variance explained	Most Important Variables
1	501	1	Training dataset	2.17	0.37	93.60%	CC2
			Testing - dataset	6.84	0.81		Base time

Table 7: Random Forest results (data without outlier after stacking)

Interpretation of Best Model

<u>Model performance post Stacked generalization</u>

Random Forest model after stacked generalization gave the best result as the RMSE reduced from 11.41

(Table 4) to 6.84 (Table 7) and MAE reduced from 2.83 (Table 4) to 0.81(Table 7). Variance explained was increased from 63.7% (Table 4) to 93.6% (Table 11)

Business Insights

Number of comments on a Facebook post can be used to understand the subject importance of the page and relevance of the post's content. With better prediction of volume magnitude and temporal variation of comments, advertising and marketing strategy can be optimized to maximize return on investment (ROI) of marketing budget.

Number of comments received in first 24 hours was more in weekdays as compare to weekends, it was highest for the post published on Wednesday as shown in Fig 33.

Page category 9 has received maximum number of comments followed by page category18 and 24 in first 24 hours of post published as shown in Fig 34.

It was observed that the comment volume was highest in initial 24 hours after post published and then it decreased between 25 to 48 hours and it further decreased between 49 to 72 hours as shown in Fig 35 and 36.

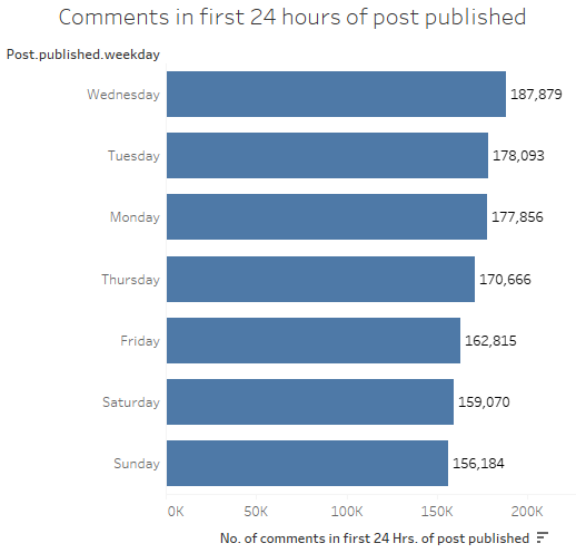


Fig 33: Bar chart of comments in first 24 Hrs. of post published

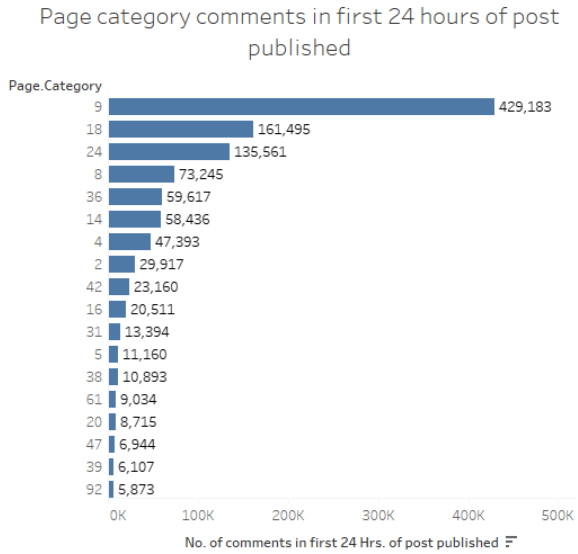


Fig 34: Bar chart of Page category comments in first 24 Hrs. of post published

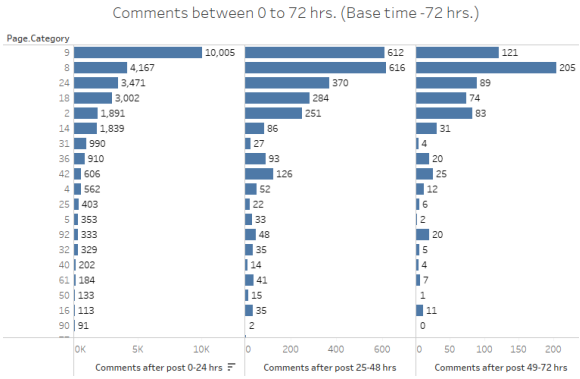


Fig 35: Bar chart of comments in 0-24 hrs. (Base time selected was 72 hrs.)

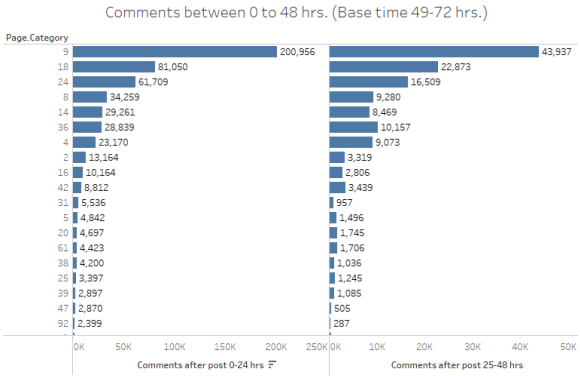


Fig 36: Bar chart of comments in 0-48 hrs. (Base time selected was 49-72 hrs.)

Conclusion and Future Scope

Our analysis has shown that much of the comment volume of a post is determined by the features of that

post's Facebook page and is relatively not related to intrinsic features of the post. The number of comments on the page in the preceding 24 hours of the base time largely predicts the amount of comments a post will receive. Among features that can be controlled by the user, the character length of a post and the number of post shares were the most predictive, but their relative importance is small when compared to other important features.

Random Forest model after stacked generalization gave the best result for comment volume prediction with least possible error, results were shown in Table 7. Our model is producing very good results, although other Ensemble techniques can be used to check the improvement in the performance of the model.