International Journal of Advanced Science and Technology
Vol. 29, No. 7s, (2020), pp. 2610-2619

Application of Random Forest For Robust Prediction Of Social Media Comments: A Case Approach

[1]Thangaraja Arumugam, [2]Vignesh Karthik & [3]Ameena Babu

[1]Assistant Professor, Business School, Vellore Institute of Technology, Chennai,
[2]Assistant Professor, SCMS School of Technology & Management, Kochi, Kerala,
[3]Assistant Professor, Amity Global Business School, Kochi, Kerala

Abstract

The measure of information that gets added to the system builds step by step and it is a gold mine of analysts who need to comprehend the complexities of client conduct and client commitment. Right now, we examine one such issue where we make a stride towards understanding the profoundly unique conduct of clients towards Social media platform posts. The objective is to anticipate what number of comments a client created present is normal on get in the given arrangement of hours. We have to show the client comments design over a lot of factors which are given and get to the correct number of comments for each post with least blunder conceivable. The assessment has revealed that a noteworthy piece of the comment volume of a post is directed by the features of that post's Social media platform page and is respectably arbitrary to inherent features of the post. Overall, this examination would assist the associations with understanding the clients conduct on posting remarks in social media platform in different days and different timings just as the factors affecting their remarking design. With these data, they can foresee the perceivability of their notice. To maintain a strategic distance from an inappropriate planning for causing commercial with the goal that cost to can be spared. Greatest reach can be accomplished.

Key words: *Random forest, Social Media marketing, predictive model.*

Aim of the study

Based on the problem statement, the objectives of the study to understand data in order to predict the patterns, insights and information and build a model to predict the volume of comments a post is expected to receive in the given set of hours.

1. Introduction

For both independent companies and enormous partnerships, social media is assuming a key job in brand building and customer correspondence. Social media platform is the social networking site important for firms to make themselves genuine for customers. Just to place things in setting, the promoting income of Social media platform in the United States in 2018 confronts 14.89 billion US dollars. The publicizing income outside the United States boils down to 18.95 billion US dollars. Most recent research reports have shown that client produced content on Social media platform drives higher commitment than promotions. The measure of information that gets added to the system builds step by step and it is a gold mine of analysts who need to comprehend the complexities of client conduct and client commitment. Right now, we examine one such issue where we make a stride towards understanding the profoundly unique conduct of clients towards Social media platform posts.

The objective is to anticipate what number of comments a client created present is normal on get in the given arrangement of hours. We have to show the client comments design over a lot of factors which are given and get to the correct number of comments for each post with least blunder conceivable.

2. Problem Statement

In light of the dataset gave the objective is to anticipate what number of comments client created presents are normal on get in the given arrangement of hours. We have to demonstrate the user comments generated over a lot of factors which are given and get to the correct number of comments for each post with least mistake conceivable and determine business bits of knowledge for powerful advertising procedures through Social media platform Posts.

3. Data Preparation and analytics

3.1. Data Description

For this study a the social media platform comments and variables have been collected form an open source platform (Kaggle). The dataset used is 'Social media platform Comment Volume Prediction' has Total 32749 observations and 43 variables in which 42 are Independent variables (40 — Numerical Variables and 2 — Categorical Variables) and 1 Dependent Variable (Numerical Variable)

3.2. Summary of the data:

In order to study the data better, we performed a preliminary variable reduction in the beginning itself. At this stage, we reduced the variable on the following criteria:

- Redundant Variables
- Business relevance
- Correlated Variables
- Target Variable

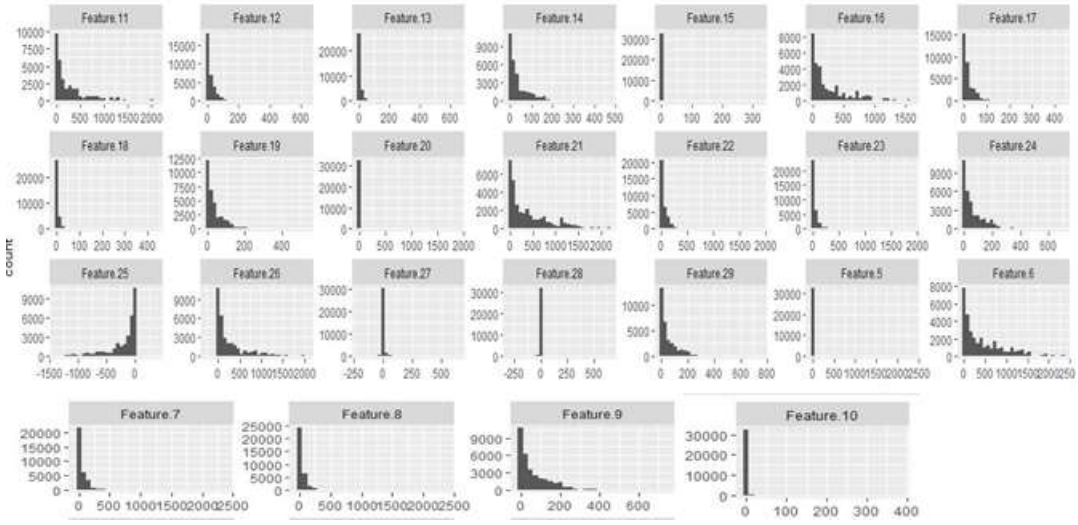

Figure 1: Histogram of the variables

Feature 5 to Feature 29 are aggregated by page, by calculating min, max, average, median and standard deviation of essential features. In our whole dataset only variables for which negative values present are for Feature 25, Feature 26, Feature 27, Feature 28 & Feature 29. None of the above-mentioned features is normally distributed and highly skewed to the right.

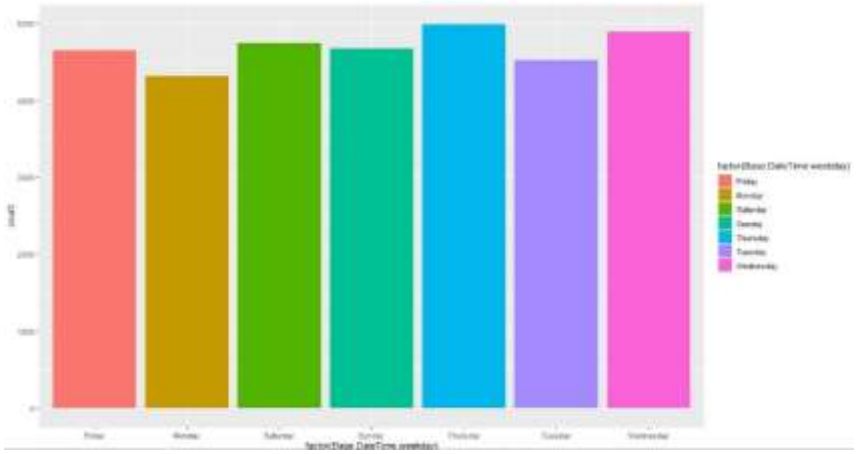

Figure 2. Distribution of Base Date and Time weekday

From the figure 3, It is understood that the characteristics of length of the post. With the count and mean mentioned, we can clearly understand how the data is distributed.

From the above graph, we can understand that frequency of post increases on daily basis and it reaches its maximum point at Wednesday and then it declines gradually. Now we must understand how the comments are coming for these posts when compared to base time.

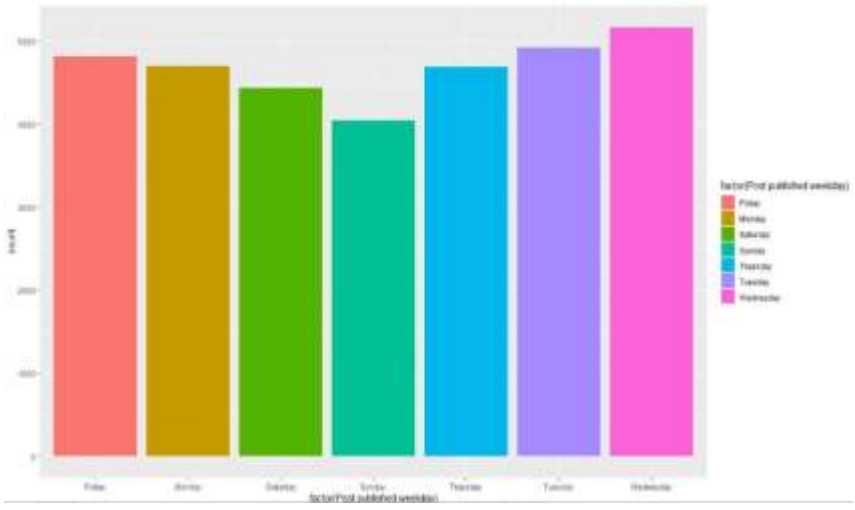

Figure 3. Distribution of Post published weekday

Missing values are present for Feature 27, Feature 29, Feature 25, Feature 20, Feature 22, Feature 18, Feature 10, Feature 13, Feature 7, Feature 15, Page Category, CC4, CC1, CC5, Page Likes, Page talking about, Page Check-ins variables. Post Promotion Status variable has only '0' entries

14.8% missing value found in the dataset.For Page Check-ins, Feature 5, Feature 10, Feature 15, Feature 20, Feature 28, CC2, CC3, Target Variable variables have more than 50% of values are '0'

In the Dataset, Feature 27, Feature 29, Feature 25, Feature 20, Feature 22, Feature 18, Feature 10, Feature 13, Feature 7, Feature 15, Page Category, CC4, CC1, CC5, Page Likes, Page talking about, Page Check-ins variables have missing values.

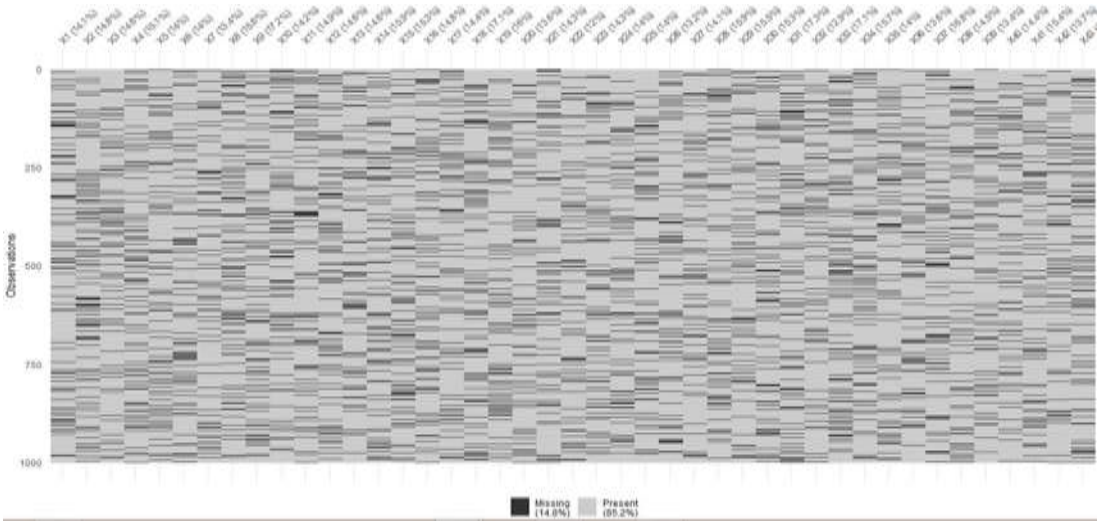

Figure 4: Missing value pattern

Table.1: Model Comparison after FeatureSelection:

Models	Dataset	After Feature Selection
		RMSE
Multiple Linear Regression	Train	6.756
	Test	6.681
SVM	Train	7.698
	Test	7.409
Decision Tree	Train	6.253
	Test	6.002
Neural Network	Train	10.537
	Test	10.314
Random Forest	Train	2.60
	Test	5.42
Extreme Gradient Boosting	Train	3.25
	Test	4.36
Bagging	Train	4.61
	Test	5.21

After the Model building with all the features selected by wrapper methods and Embedded methods, we ran a VIF check and noticed 5 variables are having very high VIF values. So, we performed a Dimension Reduction by doing PCA.

Table 2: Variance Inflation Factor for understanding variables

Variable Name	VIF
page likes	1.598809319
feature_9	239.7377926
feature_12	34.74918909
feature_13	10.90980991
feature_18	5.06507191
feature_23	15.4041934
feature_24	252.4623331
feature_27	2.238790971
feature_28	1.199715508
cc1	8.739252957
cc2	3.737948044
cc3	4.204233512
cc4	8.718787229
cc5	4.912841131
base time	1.351266551
post share_count	1.026245992

In the wake of running the PCA we found that lone the primary part has eigen value more prominent than 1 so we have chosen just a single segment after the Dimension decrease which covers around 91% information for the 5 factors.

4. Random forest for predicting comments on the given time

Random forest is a technique which is used to model the data with large number of trees and to identify the best decision tree for the model which best fit for the data. Here in this study, in order to predict the number of comments in the given timing. The Target variable is classified in to three slaps as high, medium and low level of comments.

The data set has been split into two categories as training data and test data which have been assigned with 70% and 30% of data respectively.

Table 3: Initial Model output& Confusion matrix

Type of random forest	: classification
Number of trees	: 300
No. of variables tried at each split	: 8
OOB estimate of error rate	: 5.35%

Confusion Matrix	1	2	3	Class error
1	1139	15	6	0.018103
2	48	164	1	0.230047
3	7	3	112	0.081967

From the Table, The random forest model is a classification method as the target variable is factor as such. The number of trees is 300. 8 variables tried at each split which shows that mtry is 8. That denotes the number of variables considered for each split. The out of Bag(OOB) error is 5.35% which is good and indicates the 96.65% accuracy. The confusion matrix shows that there is less error in the class one (0.018103), the error is comparatively high in the class two (0.230047) for the predictions.

Table.4: Confusion matrix and prediction: Train data

	Reference		
Prediction	1	2	3
1	1076	1	0
2	1	259	0
3	0	0	158
Model Accuracy : 0.9987 , 95% CI : (0.9952, 0.9998), No Information Rate : 0.7759 , P-Value [Acc> NIR] : < 2.2e-16 , Kappa : 0.9964			
Statistics by class	**Class: 1**	**Class: 2**	**Class: 3**
Sensitivity	0.9991	0.9953	1
Specificity	0.997	0.9992	1
PosPred Value	0.9991	0.9953	1
NegPred Value	0.997	0.9992	1
Prevalence	0.7759	0.1425	0.08161
Detection Rate	0.7753	0.1418	0.08161
Detection Prevalence	0.7759	0.1425	0.08161
Balanced Accuracy	0.9981	0.9973	1

Table 4 depicts the result of the model one prediction. The confusion matrix compares the actual value and the model prediction. The diagonal values denotes the prediction levels. Class 1 has got 1076 actual values and that has been predicted by the model as well. The same comparison for class 2 and class 3 but class 2 has got one wrong prediction. The overall model accuracy is 0.998 which is known as a good prediction model. Sensitivity and specificity shows the level of prediction accuracy as it needs to be predicted. Generally, sensitivity signifies the true positive prediction and specificity signifies the truenegativity of the model. Here the accuracy level (.998) which is deviated from the initial model. This is identified from the out of bag error value (5.35) from Table 3. Which indicates that 94% accuracy was there in initial model.

Table.5:*Prediction & Confusion matrix: Test data*

	Reference		
Prediction	1	2	3
1	479	17	3
2	14	61	2
3	2	4	49
Accuracy : 0.9334, 95% CI : (0.9111, 0.9516), No Information Rate : 0.7845, P-Value [Acc> NIR] : <2e-16, Kappa : 0.8132**,** Mcnemar's Test P-Value : 0.7633			
Statistics by class	**Class 1**	**Class 2**	**Class 3**
Sensitivity	0.9677	0.7439	0.90741
Specificity	0.8529	0.97086	0.9896
PosPred Value	0.9599	0.79221	0.89091
NegPred Value	0.8788	0.96209	0.99132
Prevalence	0.7845	0.12995	0.08558
Detection Rate	0.7591	0.09667	0.07765

Detection Prevalence	0.7908	0.12203	0.08716
Balanced Accuracy	0.9103	0.85738	0.9485

The test data prediction accuracy is slightly lesser than the train data. There are several misclassification error but still the model accuracy is good with 93% accuracy. The sensitivity for class1(.9677) is good comparatively with the other classes class 2 (.74) & class 3 (.907). The next step to measure the level of error in model prediction.

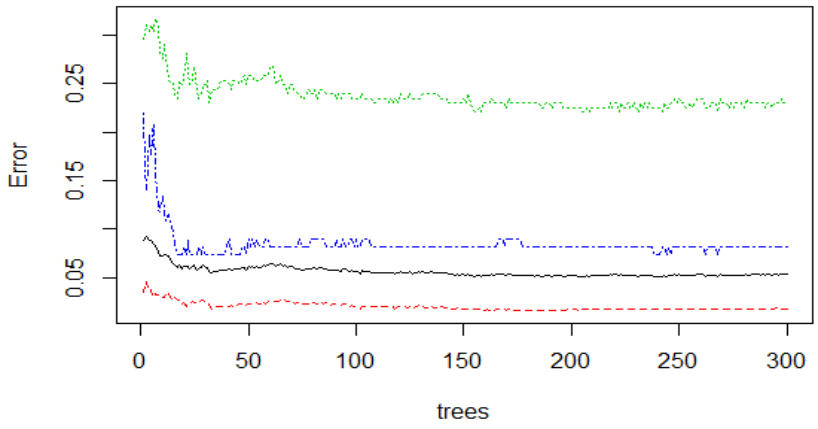

Figure.5: Error rate of random forest model

The figure 5 shows the error rate of OOB error. Initially the error rates drops and at the point of 300 it becomes constant. So optimum level of decision tress would be 300.

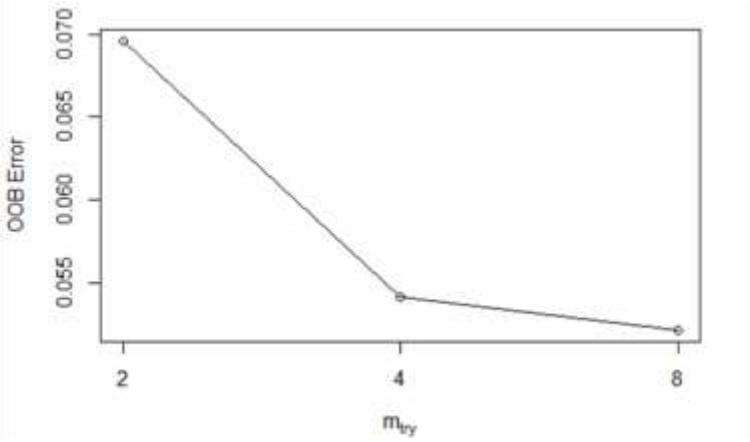

Figure.6: Tuning Random forest

Figure 6 denotes the tuning process of error level of model. Initially the error levels are high with .07 and it has become less at the mtry 8. So optimum model with less error would be 8.From OOB error and mtry values, the classification of 300 tress and the mtry of 8 has fetched a better accuracy for the model.

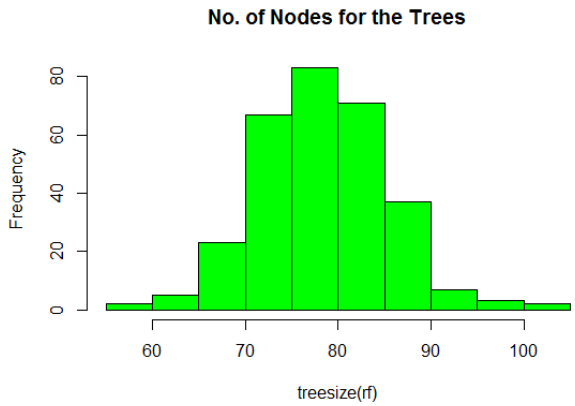

Figure.7: Number of nodes for the trees

From the figure 7, It is revealed that the number of nodes are 80 for 80 tree size. The node ranges from 60 to 100 for the trees. That indicates minimum of 60 branches in the tree and maximum of 100. The highest number of nodes for maximum number of trees are 80.

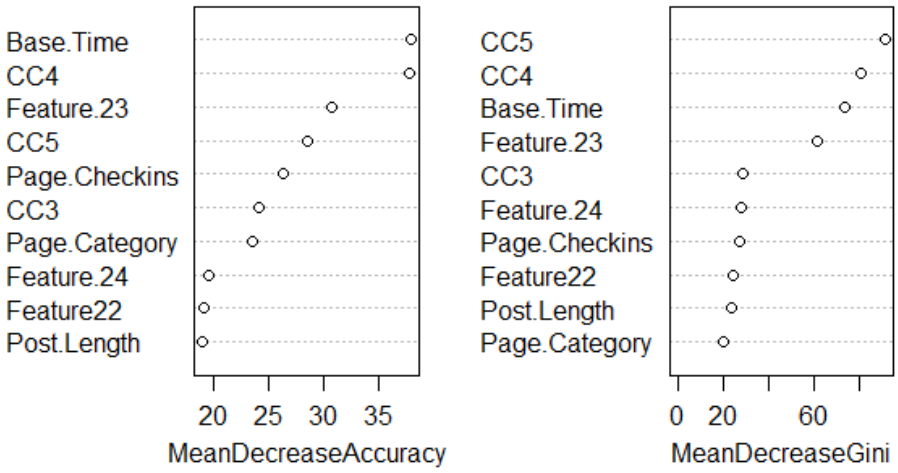

Figure.8: The Graph shows the variable importance in the model

The figure 8 indicates that the graph explains the importance of variables in the random forest model. The mean decrease accuracy denotes that The base.Time. CC4, CC5 are the most importance variables. Removal of these variables would make the model poor. Mean decrese gini measures the purness of the tree with the variables. CC5, CC4, Base.Time are the importance variables since they occupy the maximum pureness of the model.

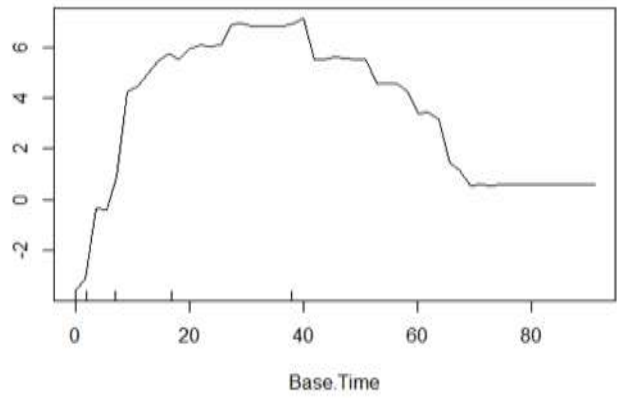

Figure.9: The Partial dependence

From the figure 9, The partial dependence plot provides the marginal effects of a variable on the class probability.In this base time variable, class2 predict moderately and class 3 predict more strongly as per the plot.

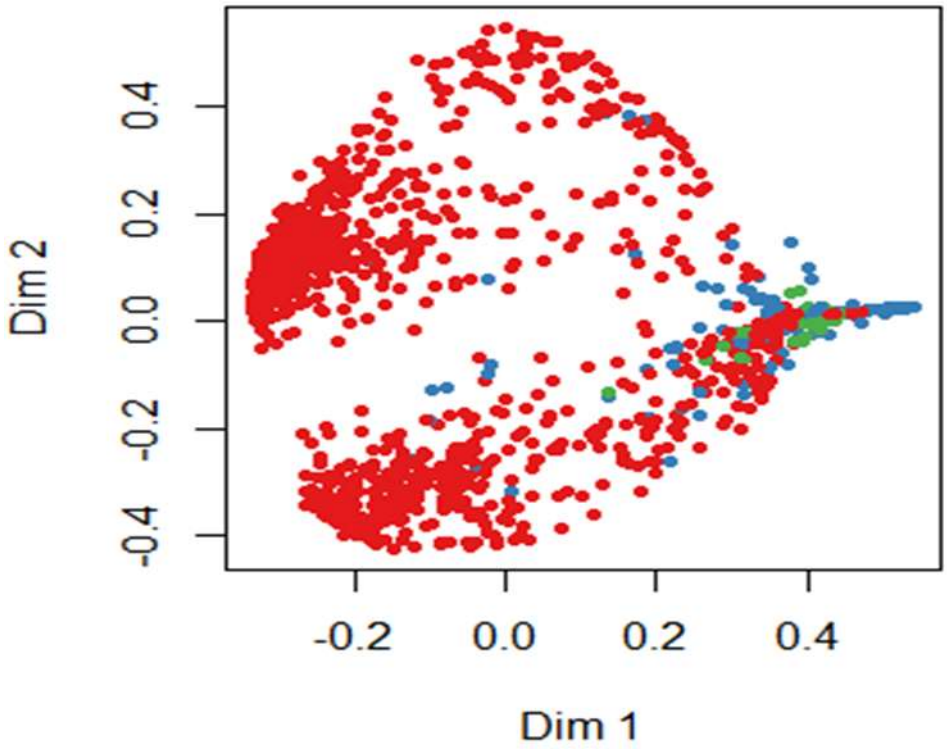

Figure.10: Multidimensional scaling of proximity Matrix

The figure 10 depicts the Multidimensional scaling of proximity Matrix. The three different colors denotes the different classification of the data distribution. The maximum data lies with 1 to 2 that is red colored in the data distribution.

Insights from the Data

Page Category 9 and 24 got most extreme remarks so better focusing on the adjustments in the page 9 category.In the initial 24 hrs. after the Post got distributed more remarks are gotten in the Weekdays, and Wednesday got the most noteworthy number of remarks so Wednesday is the correct decision for new dispatch and suggestions in the online networking.

Focusing on the 16 key factors would be best for foreseeing objective variable. Nonlinear models perform superior to the straight models so receiving the nonlinear systems may get better results. The organizations can use the outcomes for contrasting different models all together with take showcasing choices and speculation choices. Overall, This examination would assist the associations with understanding the clients conduct on posting remarks in social media platform in different days and different timings just as the factors affecting their remarking design. With these data, they can foresee the perceivability of their notice. To maintain a strategic distance from an inappropriate planning for causing commercial with the goal that cost to can be spared. Greatest reach can be accomplished.

5. Conclusion

The assessment has revealed that a noteworthy piece of the comment volume of a post is directed by the features of that post's Social media platform page and is respectably arbitrary to inherent features of the post. In particular, the amount of posts on that page in the past 24 hours and the amount of post offers, all things considered, predicts the proportion of comments a post will get. Among features that can be obliged by the customer, the character length of a post and the day of posting are the most farsighted, anyway their relative importance is little when stood out from the page features. Taking everything into account, future work could be performed to review the effect of hoisting Social media platform presents on check whether such exercises lead to progressively vital comment volume. Such a philosophy would help choose whether Social media platform post progressions are reasonable in extending the introduction of a post.

References:

1. Kamilaris and A. Pitsillides, "Social networking of the smart home," in Personal Indoor and Mobile Radio Communications (PIMRC), 2010 IEEE 21st International Symposium on, Sept 2010, pp. 2632–2637
2. K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on, May 2010, pp. 1–10.
3. T. Reuter, P. Cimiano, L. Drumond, K. Buza, and L. Schmidt- Thieme, "Scalable event-based clustering of social media via record linkage techniques." in ICWSM, 2011.
4. T. Yano and N. A. Smith, "What's worthy of comment? content and comment volume in political blogs." in ICWSM, 2010.
5. https://www.kaggle.com/kiranraje/prediction-social media platform-comment.