Facebook Comment Prediction

Table of Contents

Introduction to the business problem

- In this project, we will identify the dynamic behavior of users towards a post shared on Facebook, we will try to understand how a user comments based on a post or how many comments a post can receive in a certain period.
- We will do this to predict what type of posts receives the highest number of comments or reach. So that the advertisements can be promoted in those posts.
- The User Generated Content (UGC) can be classified as below in Table 1

		Psychological Motivation for Engaging in UGC creation			
		Rational		Emotional	
		Knowledge Sharing	Advocacy	Social Connection	Self-Expression
Platform Base	Group	Wikis (e.g., Wikipedia)	Issue-Centric Communities (e.g., Rachel ray Sucks Community)	Multiplayer online games (e.g., PUBG)	Virtual Presences (e.g., SecondLife)
	Individual	Blogs by experts (e.g., askanexpertblog.com)	Consumers reviews (e.g., Epinions)	Social Networking sites (e.g., Facebook)	Consumer creative inventions (e.g., jumpcut)

Table 1. Typology of UGC Classification and Exemplars

Data Insights

- The data has 43 columns and 32,760 rows
- The data has been collected daily for each posted content
- The columns CC1-CC5 are the count of comments for the post on different timelines like 24 hours after the content is posted
- Column **ID** is a continuous variable, and it is of no importance to us for the prediction
- Out of **32,760** records, only **634** records have values other than **24** in the **"H Local"** column.
- Column **Post Promotion Status** has only values 0 in it and it won't help us in predicting.

Exploratory Data Analysis

- Most of the Independent Variables are **Left Skewed**, most variables have value 0

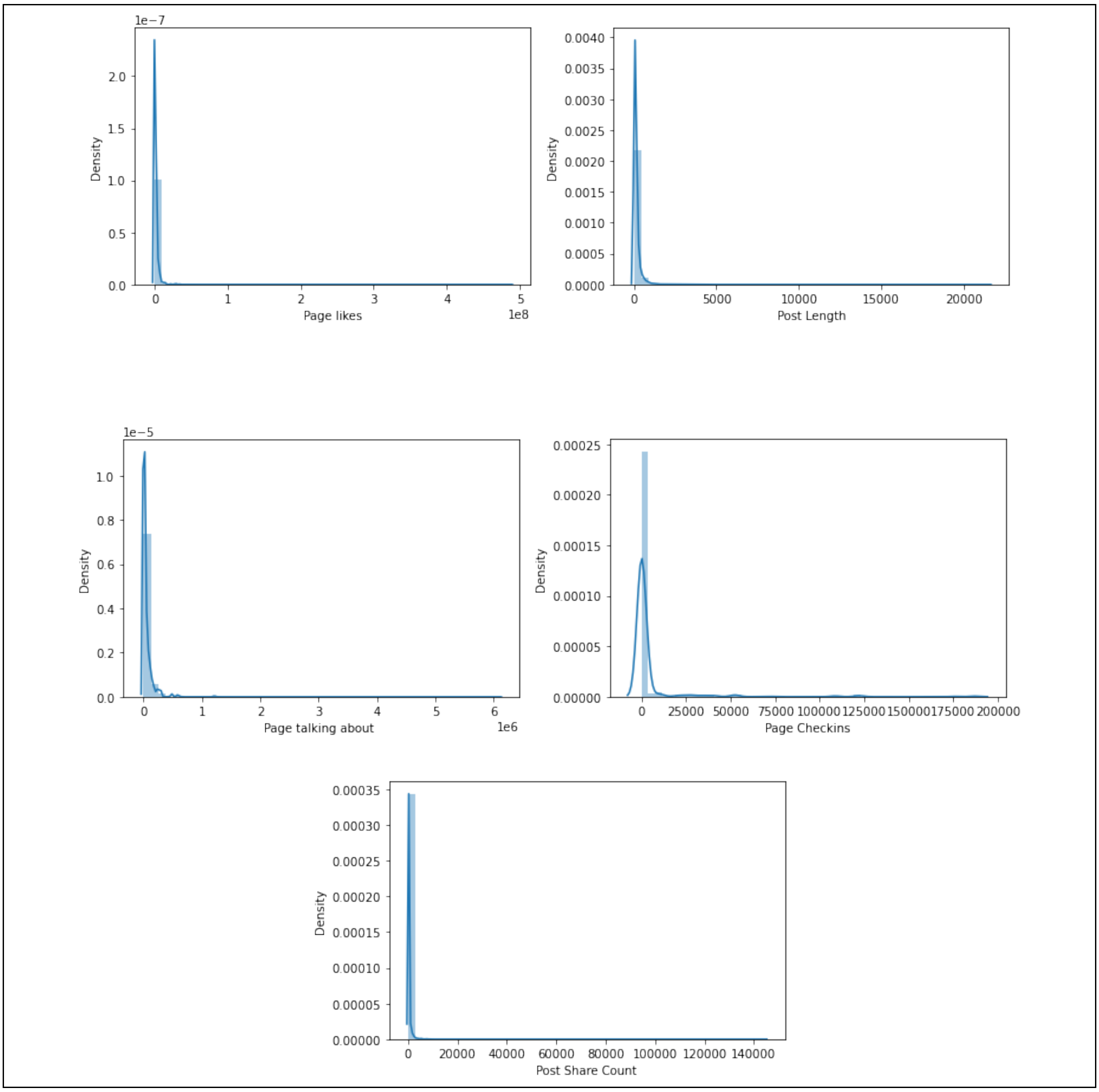


Figure 1: Distribution of Independent Variables

- Column Post Length can be converted into a categorical variable as length less than 100 can be termed as Short Post, post length greater than 100 and less than 500 can be termed as Medium Post and above 500 can be termed as Long post and create a new column postSize.

- The categorical variable Page Category has more density between 0 and 40

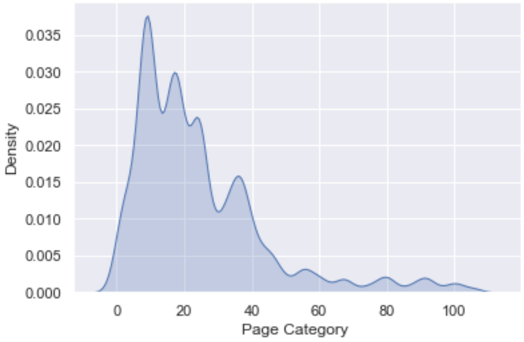


Figure 2: Page category Density

- Post posted on Wednesday has received the greatest number of comments, which can be visualized in below figure3

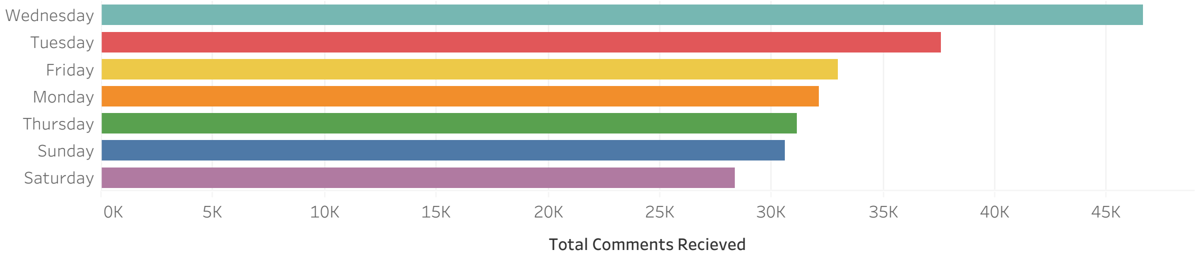
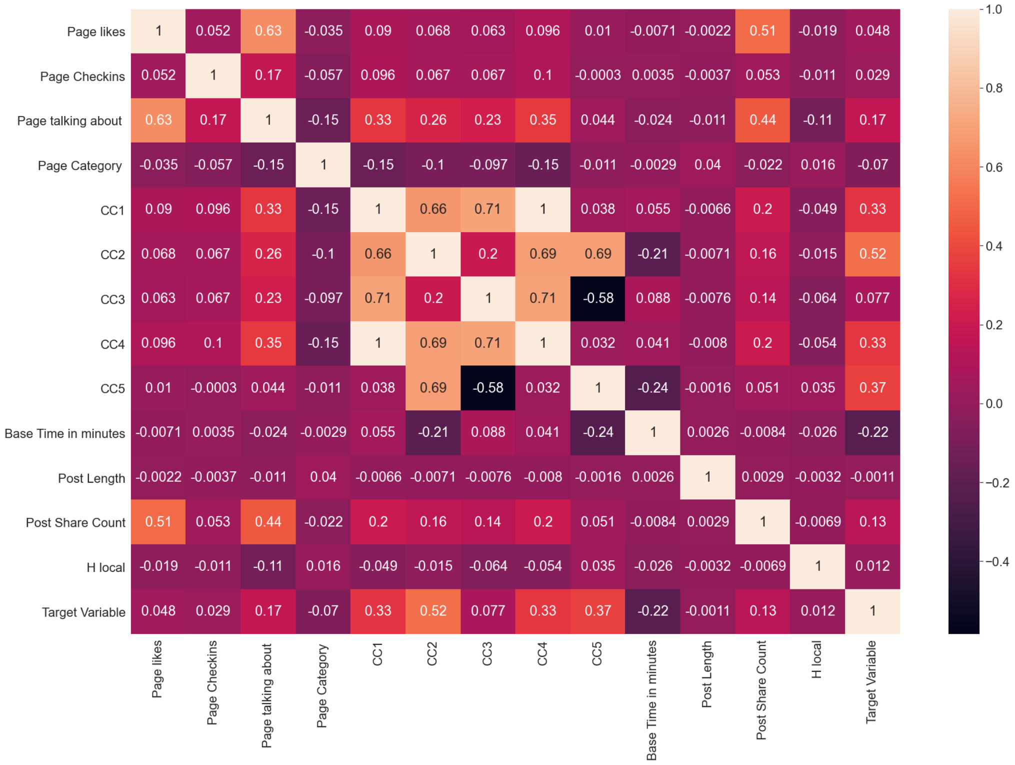


Figure 3 Weekday vs Comment Received

- Correlation on different page features can be understood from the below figure 4



	Page likes	Page Checkins	Page talking about	Page Category	CC1	CC2	CC3	CC4	CC5	Base Time in minutes	Post Length	Post Share Count	H local	Target Variable
Page likes	1	0.052	0.63	-0.035	0.09	0.068	0.063	0.096	0.01	-0.0071	-0.0022	0.51	-0.019	0.048
Page Checkins	0.052	1	0.17	-0.057	0.096	0.067	0.067	0.1	-0.0003	0.0035	-0.0037	0.053	-0.011	0.029
Page talking about	0.63	0.17	1	-0.15	0.33	0.26	0.23	0.35	0.044	-0.024	-0.011	0.44	-0.11	0.17
Page Category	-0.035	-0.057	-0.15	1	-0.15	-0.1	-0.097	-0.15	-0.011	-0.0029	0.04	-0.022	0.016	-0.07
CC1	0.09	0.096	0.33	-0.15	1	0.66	0.71	1	0.038	0.055	-0.0066	0.2	-0.049	0.33
CC2	0.068	0.067	0.26	-0.1	0.66	1	0.2	0.69	0.69	-0.21	-0.0071	0.16	-0.015	0.52
CC3	0.063	0.067	0.23	-0.097	0.71	0.2	1	0.71	-0.58	0.088	-0.0076	0.14	-0.064	0.077
CC4	0.096	0.1	0.35	-0.15	1	0.69	0.71	1	0.032	0.041	-0.008	0.2	-0.054	0.33
CC5	0.01	-0.0003	0.044	-0.011	0.038	0.69	-0.58	0.032	1	-0.24	-0.0016	0.051	0.035	0.37
Base Time in minutes	-0.0071	0.0035	-0.024	-0.0029	0.055	-0.21	0.088	0.041	-0.24	1	0.0026	-0.0084	-0.026	-0.22
Post Length	-0.0022	-0.0037	-0.011	0.04	-0.0066	-0.0071	-0.0076	-0.008	-0.0016	0.0026	1	0.0029	-0.0032	-0.0011
Post Share Count	0.51	0.053	0.44	-0.022	0.2	0.16	0.14	0.2	0.051	-0.0084	0.0029	1	-0.0069	0.13
H local	-0.019	-0.011	-0.11	0.016	-0.049	-0.015	-0.064	-0.054	0.035	-0.026	-0.0032	-0.0069	1	0.012
Target Variable	0.048	0.029	0.17	-0.07	0.33	0.52	0.077	0.33	0.37	-0.22	-0.0011	0.13	0.012	1

Figure 4: Page Features Correlation

- From figure 4 we can interrupt that Variable CC1 and CC4 are highly correlated
- Also, we can see there is an inverse correlation between CC3 and CC5
- And CC1 is closely related to CC2 and CC3. CC2 is closely related to CC4 and CC5. CC3 is closely related to CC4.
- To Support our correlation identification, refer to the below figure 5 how data are distributed.

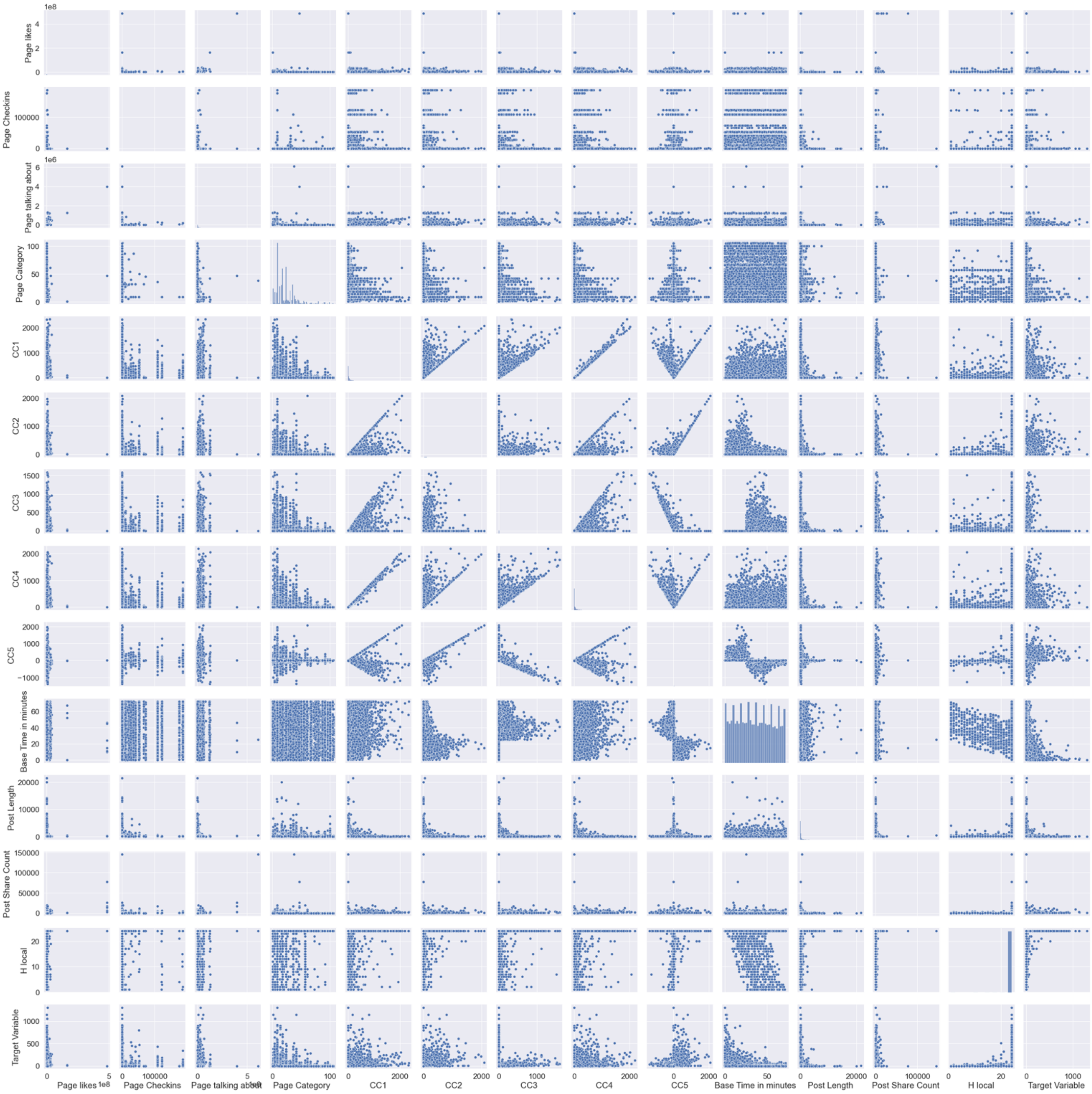


Figure 5 Distribution of all Page Features

- The relationship between post length and number of comments received are detailed in below figure 6

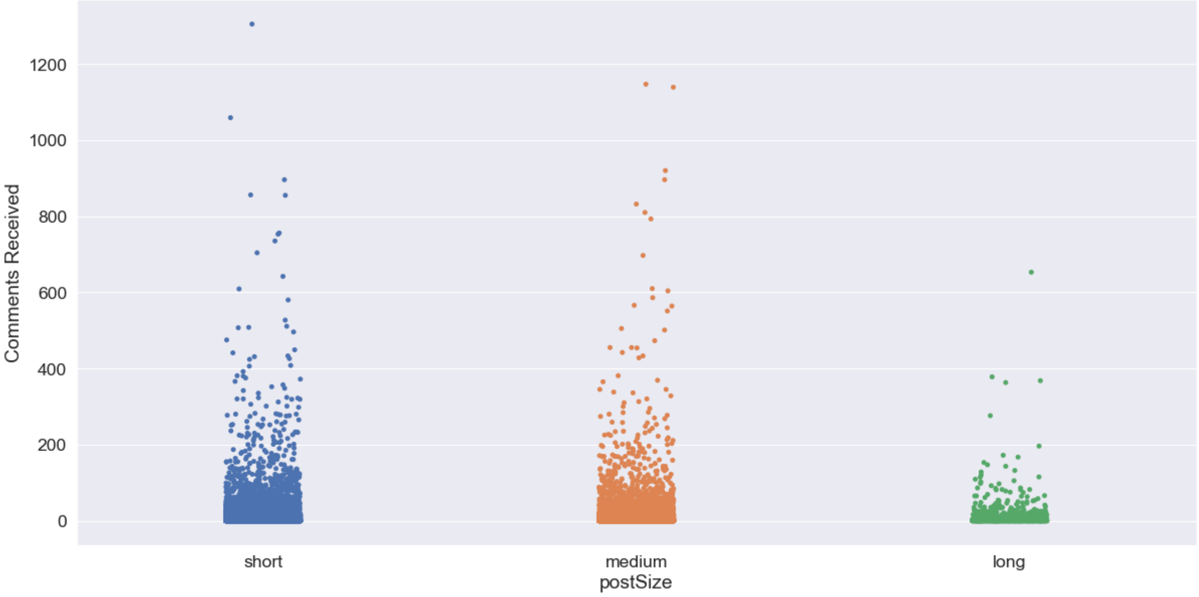


Figure 6 Post Length vs Comments Received

- We can see comments are high for a post with less than 500 characters
- We can also see in figure 7 the content posted at 24 hours received more comments

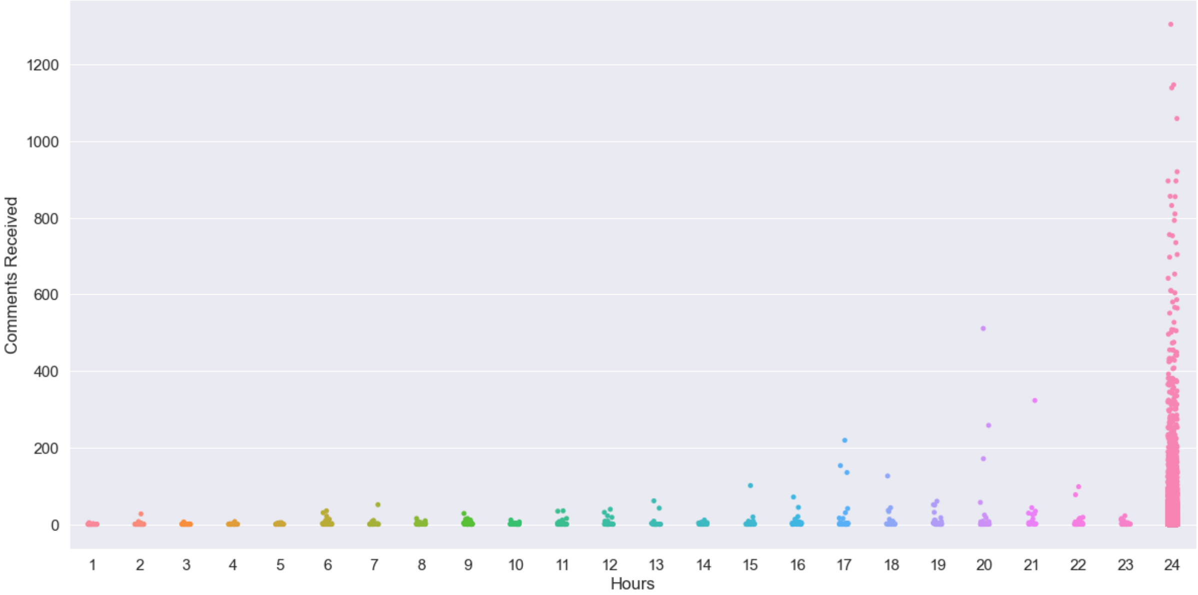
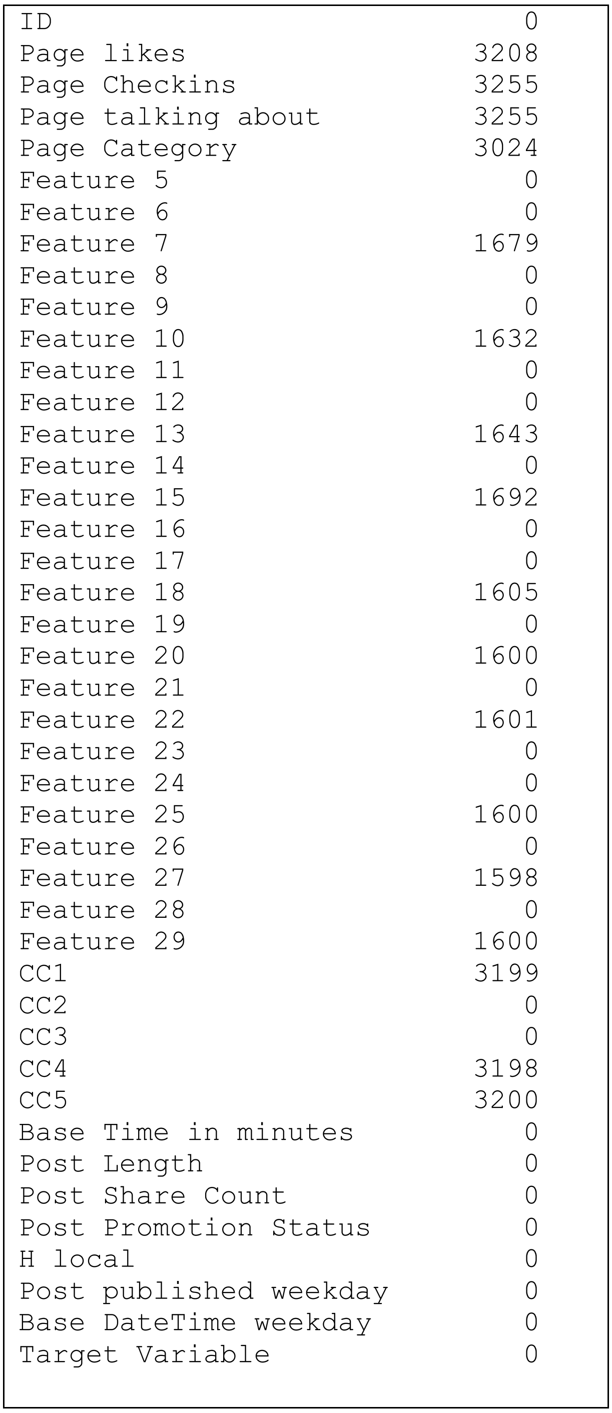


Figure 7 Hours vs Comments Received

Data Pre-processing

- The variable ID is a continuous variable and we do not have any use with it so we can remove it.
- The variable Post promotion Status has only one value 0, so we can remove this column as well.
- There are a lot of values missing in most of the variables

- We will find mean and impute these missing values as in figure 8



```
ID                               0
Page likes                    3208
Page Checkins                 3255
Page talking about            3255
Page Category                 3024
Feature 5                        0
Feature 6                        0
Feature 7                     1679
Feature 8                        0
Feature 9                        0
Feature 10                    1632
Feature 11                       0
Feature 12                       0
Feature 13                    1643
Feature 14                       0
Feature 15                    1692
Feature 16                       0
Feature 17                       0
Feature 18                    1605
Feature 19                       0
Feature 20                    1600
Feature 21                       0
Feature 22                    1601
Feature 23                       0
Feature 24                       0
Feature 25                    1600
Feature 26                       0
Feature 27                    1598
Feature 28                       0
Feature 29                    1600
CC1                           3199
CC2                              0
CC3                              0
CC4                           3198
CC5                           3200
Base Time in minutes             0
Post Length                      0
Post Share Count                 0
Post Promotion Status            0
H local                          0
Post published weekday           0
Base DateTime weekday            0
Target Variable                  0
```

Figure 8 Missing Values

- There are few Outliers in our data as seen in figure 9. We will find the interquartile range and bring them into the accepted range.

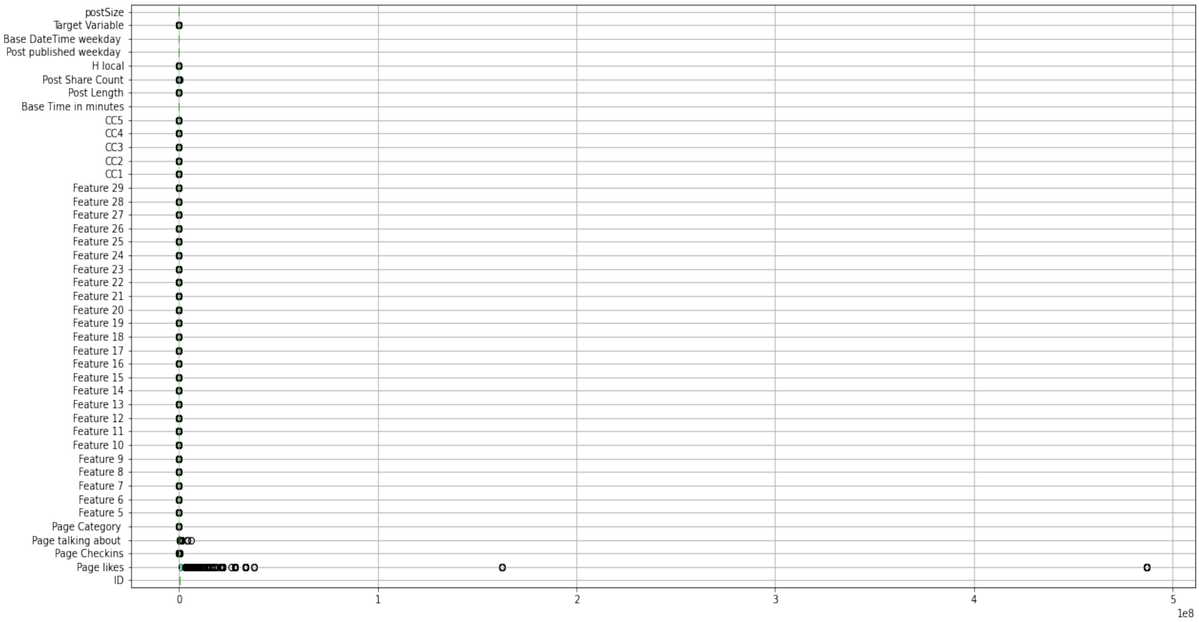


Figure 9 Outlier on Variables

- We can see even our Target variable has some outliers we will not treat it.
- “Page likes” variable has most outliers followed by “Page talking about”
- We, Will, transform the “Post Length” variable into a categorical variable by putting them in bins of post length less than 100 as Short Post, greater than 100, and less than 500 as Medium post and length greater than 500 as long post.

Model Building

Linear Regression

- We started with Linear Regression model as our Target Variable was a continuous variable. As Linear Regression easy and best way to start for regression based problems. We will try to find the best fit line by applying this model.
- We split the data for training and testing in the ratio of 75:25 respectively.
- We applied linear regression and identified the coefficient of each independent variable. The below *figure 10* denotes the coefficients of each model.
- The coefficients denote the positive and negative impact on prediction
- If the coefficient lies within the positive value means proportional to the prediction else inversely proportional for the prediction
- From the *figure10* we can see **Feature 8** has high positive coefficient which is the most important variable in the Linear Regression model prediction
- Followed by **length of the post, Feature 28, Feature 14**.
- The accuracy score of Linear Regression Model on **Train data is 19%.**
- The accuracy score of Linear Regression Model on **Test data is 17%.**
- The Intercept of the model is **5.94.**

- As we can see model is under performed, model scores are very low, we will proceed with other models.

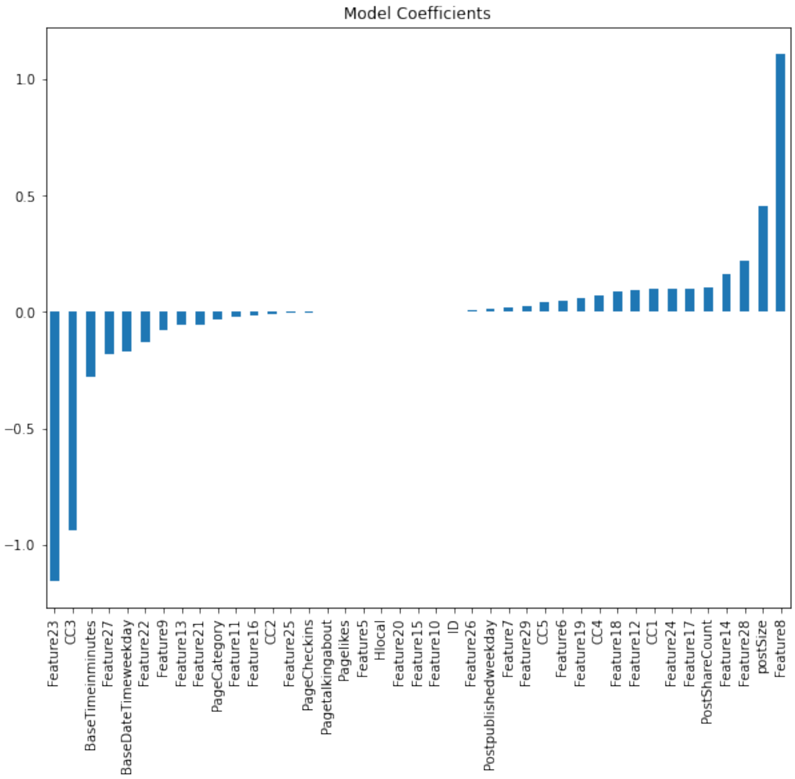


Figure 10 Coefficients of Independent Variable Linear Regression

Lasso Regression

- It is a popular type of regularized linear regression that includes an L1 penalty. This has the effect of shrinking the coefficients for those input variables that do not contribute much to the prediction task. This penalty allows some coefficient values to go to the value of zero, allowing input variables to be effectively removed from the model, providing a type of automatic feature selection.
- The Accuracy on **Train data** is **18%**
- The Accuracy on **Train data** is **15%**
- From *figure 11* we can see **CC5** have a positive impact on Lasso Prediction
- Followed by **Feature12, CC2, CC1**
- We can see from *figure 11* most of the variables have coefficient near to 0 or in negative values

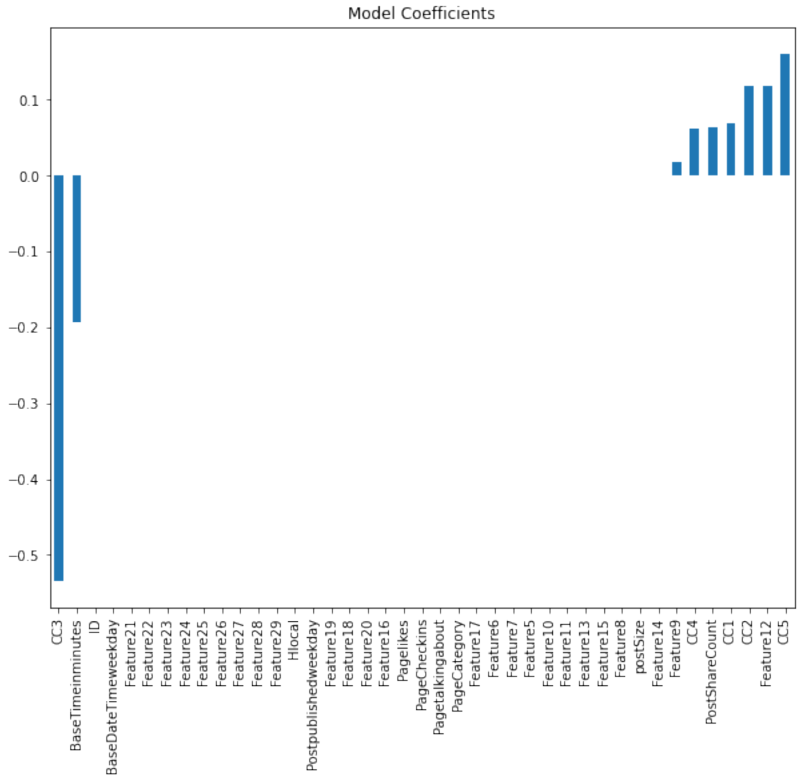


Figure 11 Coefficient of Lasso Model

- We can see our density of our prediction in *figure 12*, its more towards 0

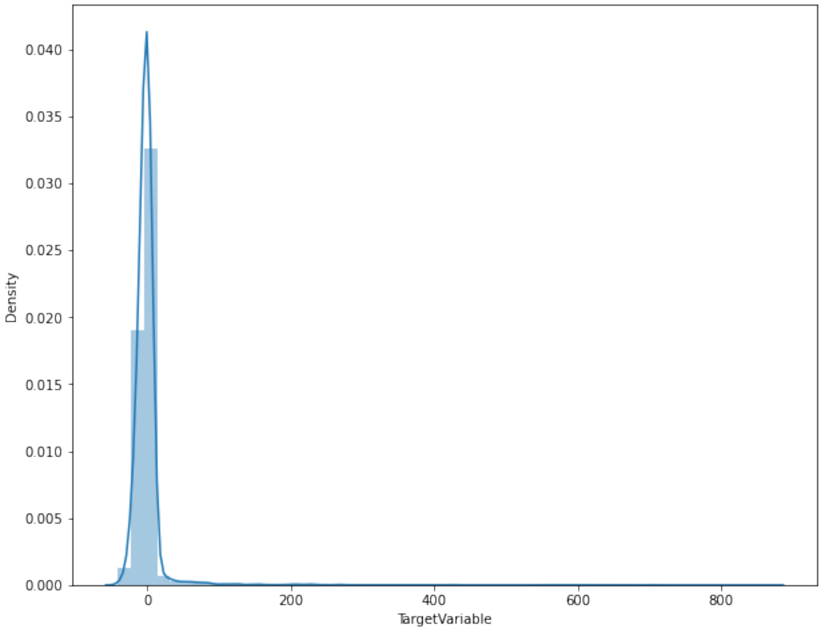


Figure 12 Lasso Prediction Density

Ridge Regression

- Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).
- Using GridsearchCV we identifier the hyperparameters for Ridge regression
- The alpha we identified is **0.001**
- The Model score for **Train Data** is **20%**

- The Model score for **Test Data** is **16%**
- The model Coefficients are put in below *figure 13*

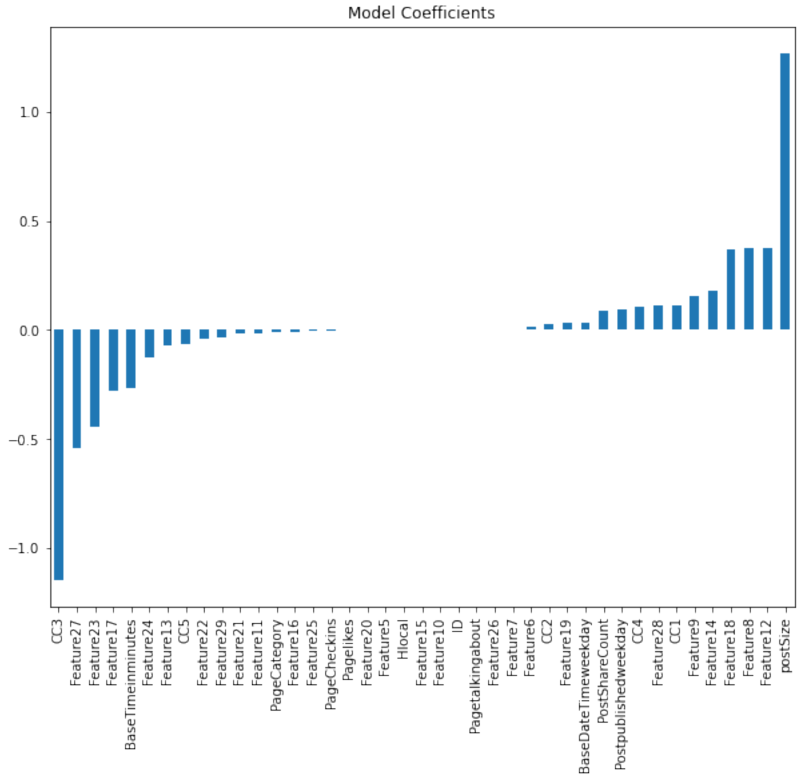


Figure 13 Ridge model Coefficients

- **Post Length** plays a significant role on predict with Ridge Regression followed by **Feature 12 and Feature 8**
- We can see most variables have a coefficient 0 or in negative
- We can see the density of our predictions in below figure

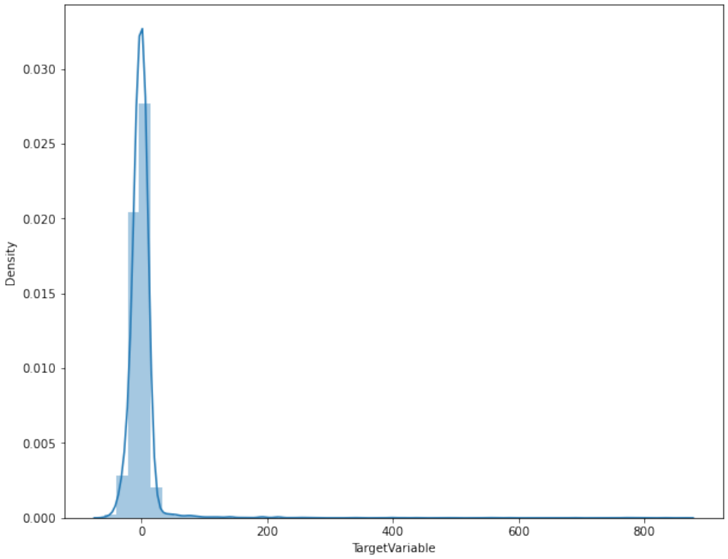


Figure 14 Ridge Regression Predicted Density

- We can our predictions with ridge also mostly towards 0

XGBoost

- To improve the model performance we can tune it further with boosting techniques.
- One of the boosting technique is XGboosting
- **XGBoost** is an optimized distributed gradient boosting library designed to be highly **efficient**, **flexible** and **portable**. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.
- We have used XGboost with hyper parameter **learning rate** as **0.01**.
- The accuracy of the model on **Train data** is **61%**
- The accuracy of the model on **Test Data** is **55%**
- Model performed better with XGBoosting but still we don't have greater accuracy

Problems With Data

- The data looks imbalanced as the Target variable is more skewed to the left as seen in figure 10

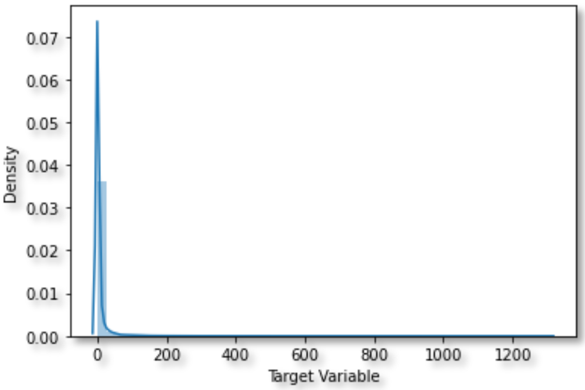


Figure 15 Skewness on Target Variable

- Even the data is dense more towards the bottom as seen in figure 11

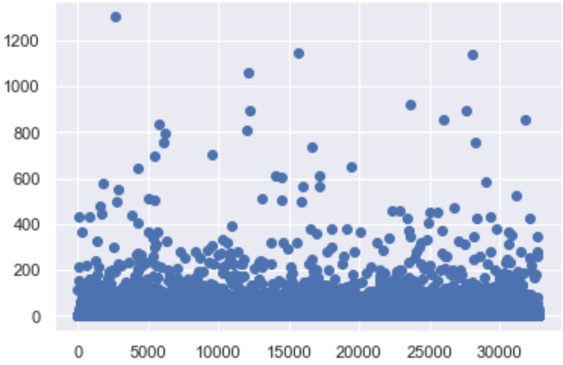


Figure 16 Scattering of Target Variable

- We can solve the Imbalance in continuous variables using Synthetic Minority Over-Sampling Technique for Regression (SMOTE for Regression).

- Or we can convert the Target Variable into a classification variable as 0-10 comments are **low comments,** 10-100 are **medium comments** and above 100 **high comments**
- By doing so we can solve the imbalance easily with SMOTE and our model can perform better

Logistic Regression

- Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression
- With the help of Grid Search we were able to identify the appropriate Hyper parameters for the Logistic Regression Model.
- They are
  - **Solver**: "newton-cg"
  - **Max Iteration**: "100,000"
  - **Number of Jobs**: "2"
- This improved the model accuracy
- We were able to achieve 98% accuracy on train and test data

Model validations

- Linear Regression
  - The model is validated with R^2 Test and Root Mean Squared Error (RMSE)
  - The value for **R^2Test** on Train data is **0.19**
  - The value of **RMSE** on Train Data is **31.09**
  - The value for **R^2Test** on Test data is **0.17**
  - The value of **RMSE** on Test Data is **32.81**

- Lasso Regression
  - The model is validated with R^2 Test and Root Mean Squared Error (RMSE)
  - The value for **R^2Test** on Train data is **0.18**
  - The value of **RMSE** on Train Data is **996.22**
  - The value for **R^2Test** on Test data is **0.15**
  - The value of **RMSE** on Test Data is **1100.90**

- Ridge Regression
  - The model is validated with R^2 Test and Root Mean Squared Error (RMSE)
  - The value for **R^2Test** on Train data is **0.20**
  - The value of **RMSE** on Train Data is **967.90**
  - The value for **R^2Test** on Test data is **0.16**
  - The value of **RMSE** on Test Data is **1085.63**

- XGBoost
  - The model is validated with R^2 Test and Root Mean Squared Error (RMSE)
  - The value for **R^2Test** on Train data is **0.61**
  - The value of **RMSE** on Train Data is **1271.71**
  - The value for **R^2Test** on Test data is **0.55**
  - The value of **RMSE** on Test Data is **1358.97**

- Logistic Regression
  - Logistic Regression model is validated with ROC_AUC_Score
  - The ROC_AUC_Score on **Train** data is **0.99**
  - The ROC_AUC_Score on **Test** data is **0.98**

Model to Use

- As we can clearly see Logistic Regression played a significant role in predicting compared with other models
- As the data is imbalanced mostly skewed to left it is very hard for us to predict the exact value.
- So we have converted our Target Variable into a Classification variable so we can bin the output into 3 buckets first bucket contains count of comments between 0-10, second bucket contains count of comments between 10-100, third bucket contains comments received greater than 100.
- We applied Logistic Regression and find the results for train data in the ROC AUC curve below in *figure17*

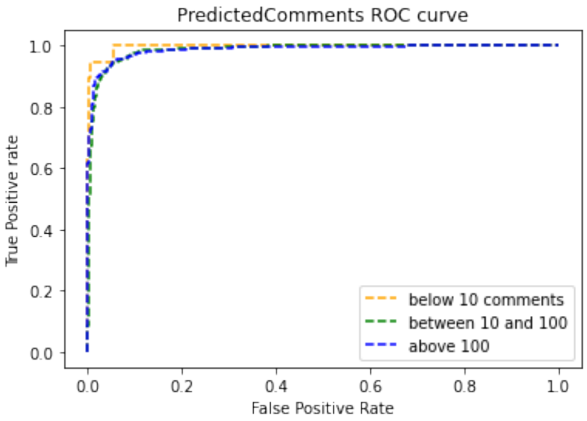


Figure 17 ROC AUC Curve for Train Data

- The results on test data after applying Logistic Regression

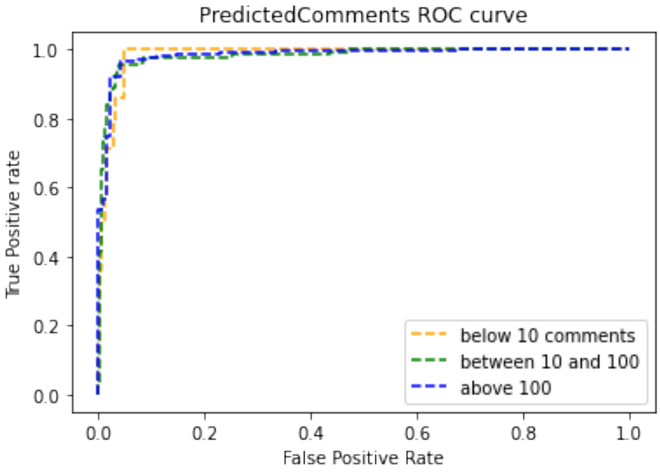


Figure 18 ROC AUC Curve for Test Data

Insights from the Models

- We can use Logistic Regression to predict the no of comments for a particular Post
- The important features for the prediction are displayed in the below figure 8

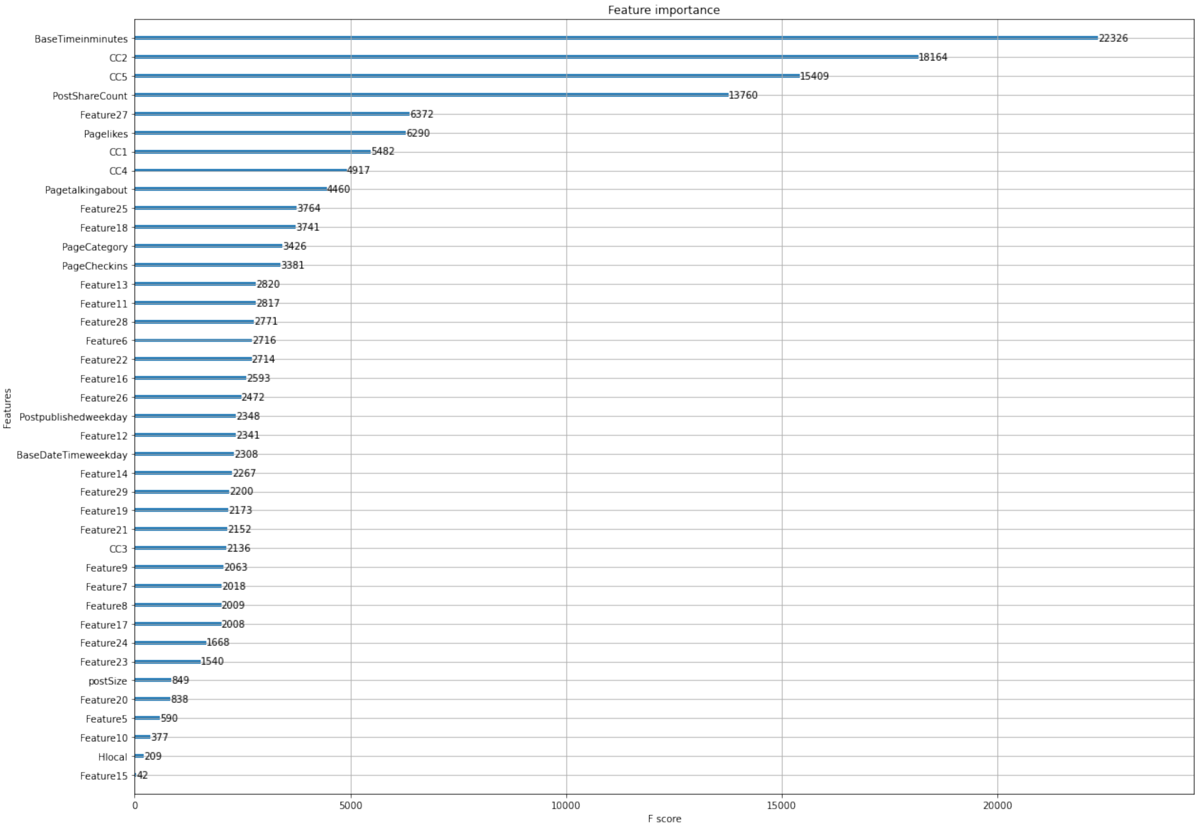


Figure 19 Feature Importance

- The top 5 variables are “Base Time”, “CC2”, “CC5”, “Post Share Count”, “Feature 27”.
- The comments which the post received with in 48 hours determine the no of comments a post can have that is what is derived from CC2 and CC5 variables
- The highest number of times the post is shared there is a high chance of getting more comments
- The time in which the Content is posted also contributes for a greater number of comments.

My Recommendation

To receive a greater number of comments, for user generated content, it is important to focus on 3 essential factors. Duration of the day, Timeline of the Post, and richness and easy understanding of the content within **500 char limits** and shared more than 10 times.

When a highly rich and understandable content (<500 char limit) posted on **Thursday** at **00 hours midnight** and shared more than 10 times with the time limit of **48 hours** can fetch maximum user comments with high confidence level of 98%

Appendix

UGC:
https://www.researchgate.net/profile/Sandeep_Krishnamurthy/publication/274176873_Note_from_Special_Issue_Editors/links/5553be2c08aeaaff3bf19cc5.pdf

SMOTE for Regression: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
https://pypi.org/project/smogn/