Facebook Comment Prediction

### Model Building

* We have split the data in the ration **75:25**, for train and test data respectively
* We started with the Linear Regression model and identified the coefficient of each independent variable.
* And we have ended with the below figure 1

A picture containing histogram

Description automatically generated

Figure 1 Coefficient of Independent Variables

* The Linear Regression model score for Train data is **0.196**
* The Linear Regression model score for Test data is **0.171**
* The RMSE Score for Train data is **31.09**
* The RMSE score for Test data is **32.81**
* The Intercept for the model is **5.94**
* The Model score is very low we will move forward with other models

### Lasso Regression

* To reduce the model complexity we used Lasso Regression Technique
* The model score for Train data is 0.182
* The model score for Test data is 0.157
* The RMSE score for Train Data is 996.22
* The RMS score for Test Data is 1100.90
* The model Coefficients are put in below figure 2

Text

Description automatically generated with medium confidence

Figure 2 Lasso Regression Coefficients

* We can see most variables have a coefficient of 0 or in negative
* We can see the density of our predictions in below figure 3

Shape

Description automatically generated

Figure 3 Predicted Density

### Ridge Regression

* To improve much more we started with Ridge regression
* Using GridsearchCV we identifier the hyperparameters for Ridge regression
* The alpha we identified is **0.001**
* The Model score for Train Data is 0.20
* The Model score for Test Data is 0.16
* The RMSE score for Train Data is 967.90
* The RMS score for Test Data is 1085.63
* The model Coefficients are put in below figure 4

Chart

Description automatically generated

Figure 4 Ridge Model Coefficients

* We can see most variables have a coefficient 0 or in negative
* We can see the density of our predictions in below figure 5

Shape

Description automatically generated with medium confidence

Figure 5 Ridge Predicted Density

* We can see the density is more towards 0.

### Elastic Net Regression

* Elastic net is a regularized regression method that linearly combines the L₁ and L₂ penalties of the lasso and ridge methods
* Using GridsearchCV we identifier the hyperparameters for Elastic Net regression
* The alpha we identified is **0.01**
* The Model score for Train Data is 0.20
* The Model score for Test Data is 0.16
* The RMSE score for Train Data is 968.12
* The RMS score for Test Data is 1084.90
* The model Coefficients are put in below figure 6

Chart

Description automatically generated

Figure 6 Elastic Net Regression Coefficients

* We can see most variables have a coefficient of 0 or in negative
* We can see the density of our predictions in below figure 7

Shape

Description automatically generated

Figure 7 Elastic Net Prediction Density

### Interpretation from the Models

* We can see the score didn’t improve much in all the models
* We have any how able to find the Best Params which can be used they are **“Feature8”, “postSize”, “Feature28”, “PostShareCount”, “Feature17”, “Feature24”, “CC1”, “Feature12”, “Feature18”, “CC4”, “Feature19”, “Feature6”, “CC5”, “Feature29”, “Feature7”, “Postpublishedweekday”**
* The Prediction density is more towards 0 as most of our data contains 0 in the **Target Variable** column

Note**: *To Avoid this we can change the Target Variable into Categorical variable by dividing the values Like 0-10 comments as LOW, 10-100 as MEDIUM, and above 100 as GOOD. So that we can convert the Regression problem into Classification Problem and we can able to predict Better by applying SMOTE to balance the data.***

### **Model Tuning and business implication**

* We can further tune our model by using **XG Boosting.**
* 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 learning rate as 0.01
* The Model score for Train Data is 0.61
* The Model score for Test Data is 0.55
* The RMS score for Test Data is 1067.73
* The Model score have increased drastically after implementing XGBoost

### Business Implication

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

Graphical user interface, table

Description automatically generated

Figure 8 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 wat 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 more number of comments.

### Other Approach

* By Converting the Regression problem to classification problem and applying Logistic Regression
* We Used Logistic Regression with following hyper parameters
  + Solver: “newton-cg”
  + Max Iteration: “100,000”
  + Number of Jobs: “2”
* We were able to increase the model score
* The model score for train data is 0.9895
* The model score for test data is 0.9875