STATISTICS FOR DATA SCIENCE
ASSIGNMENT-3

Name:- K.Sai Prasanna Kumar **Roll No:-**S20160010047

Contents

Abstract: -

In this report we discuss about data analysis of Blog Feedback data set. Instances in this dataset contain features extracted from blog posts. There are total **281 attributes** in the dataset. The last attribute is our target attribute. We choose a basetime and select the blog posts before the selected basetime. The task associated with data is to predict how many comments the post will receive in upcoming 24 hours.

Problem Definition: -

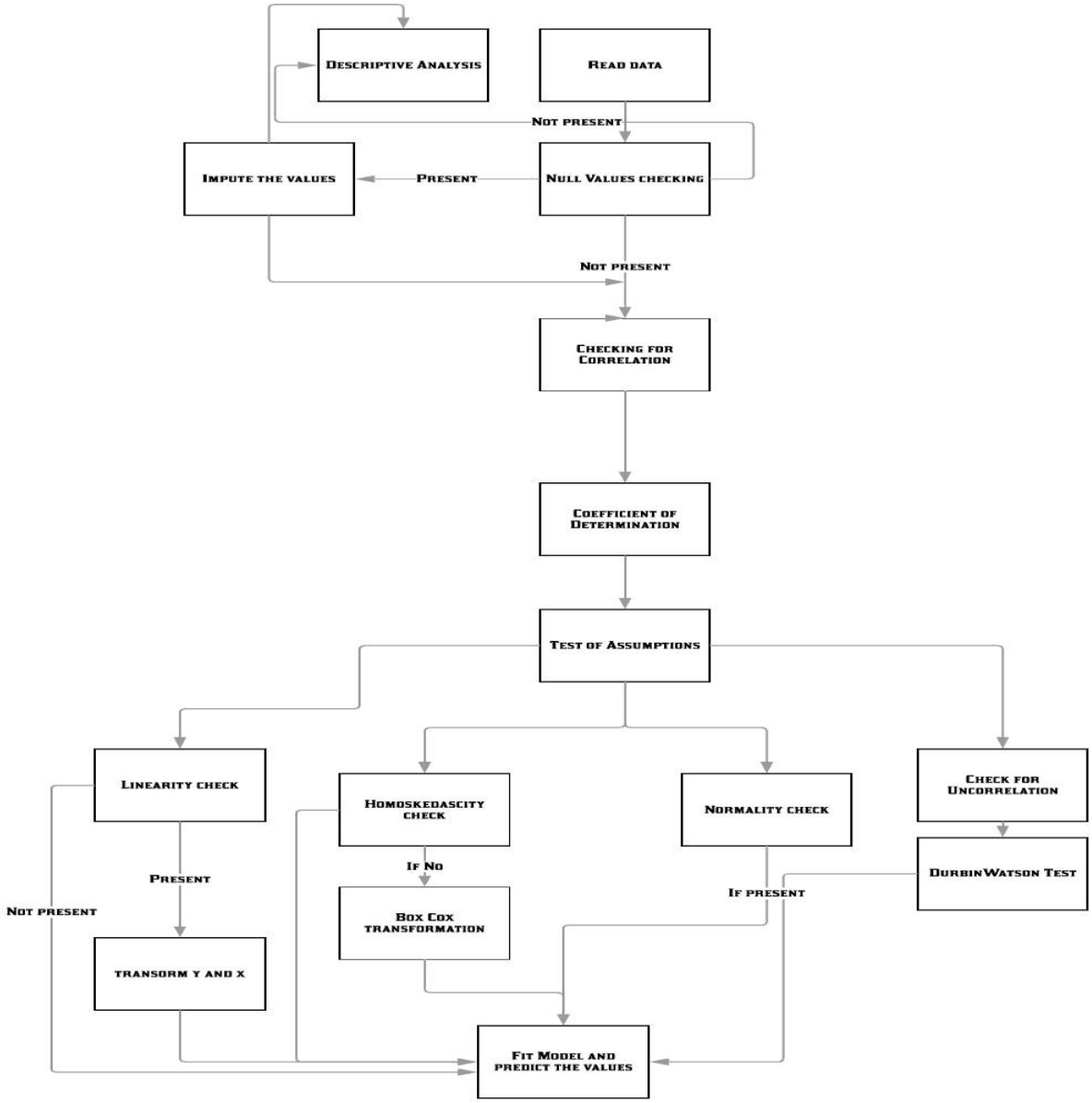
Dataset contain features extracted from blog posts. Each observation in the dataset belongs to one blog post. The prediction task associated with the data is the prediction of the number of comments in upcoming 24 hours.

Dataset Description: -

This data originates from blog posts. The raw HTML-documents of the blog posts were crawled and processed. The prediction task associated with the data is the prediction of the number of comments in the upcoming 24 hours. In order to simulate this situation, we choose a basetime (in the past) and select the blog posts that were published at most 72 hours before the selected base date/time. Then, we calculate all the features of the selected blog posts from the information that was available at the basetime, therefore each instance corresponds to a blog post. The target is the number of comments that the blog post received in the next 24 hours relative to the basetime.

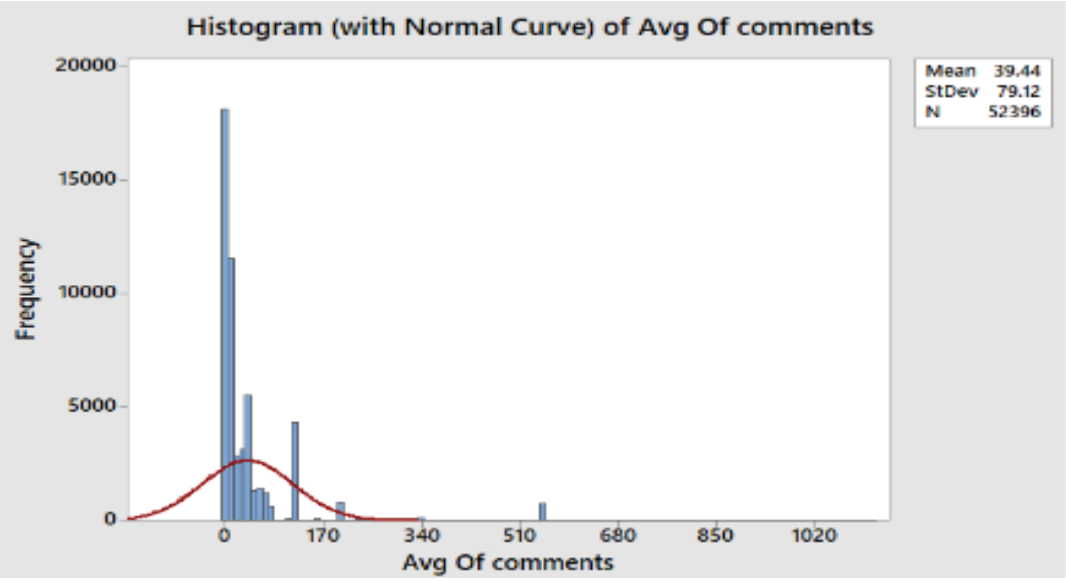
In the train data, the basetimes were in the years 2010 and 2011. In the test data the basetimes were in February and March 2012. This simulates the real-world situation in which training data from the past is available to predict events in the future. The train data was generated from different basetimes that may temporally overlap. Therefore, if you simply split the train into disjoint partitions, the underlying time intervals may overlap. Therefore, the you should use the provided, temporally disjoint train and test splits in order to ensure that the evaluation is fair.

Work Flow: -



Methodology: -

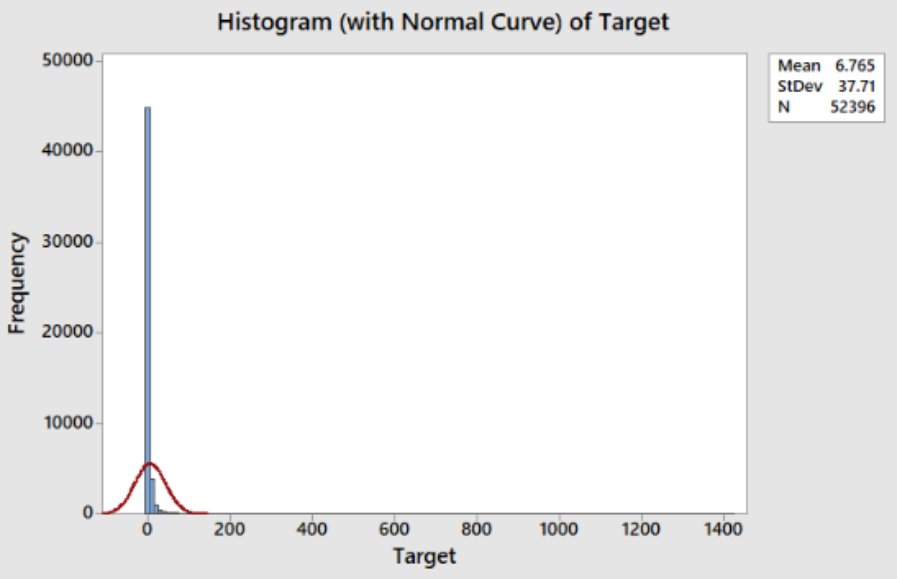
1.Descriptive Analysis: -

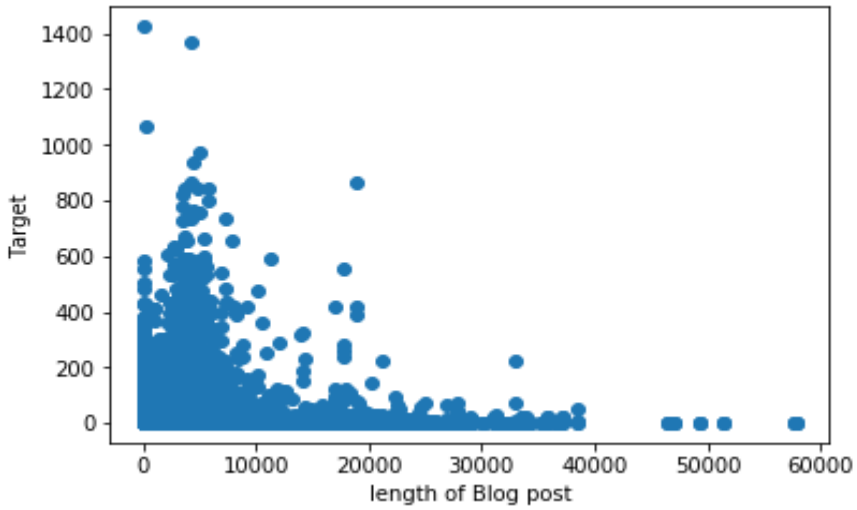


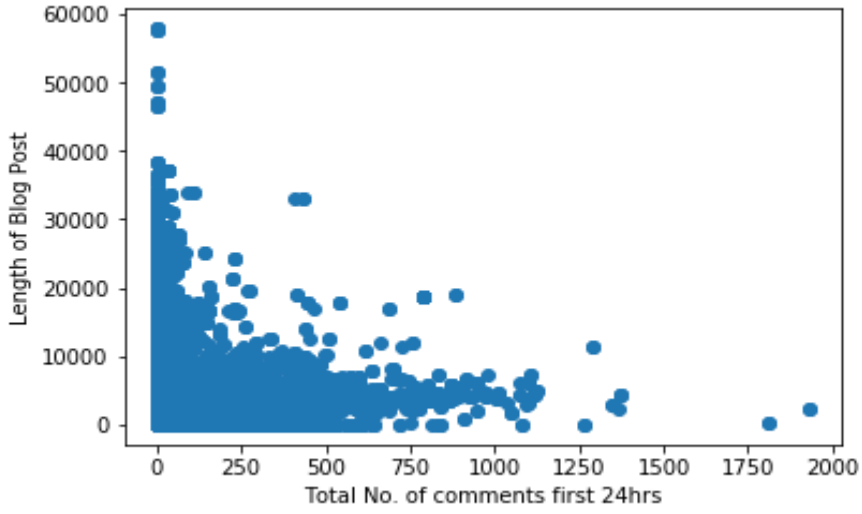
The above graph is histogram of the distribution of statistic average number of comments which follows normal distribution with mean = 39.44 and variance = 6259.9744.

Now we let's see the distribution of target class,

The below graph is histogram of the distribution of statistic average number of comments which follows normal distribution with mean = 6.765 and variance = 1422.0441.


Histogram (with Normal Curve) of Target
Mean 6.765
StDev 37.71
N 52396
Frequency
Target
50000
40000
30000
20000
10000
0
0
200
400
600
800
1000
1200
1400


Target
length of Blog post
1400
1200
1000
800
600
400
200
0
0
10000
20000
30000
40000
50000
60000


Length of Blog Post
Total No. of comments first 24hrs
60000
50000
40000
30000
20000
10000
0
0
250
500
750
1000
1250
1500
1750
2000

Coefficient of Determination:

- The **coefficient of determination** (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.
- We apply Linear regression to calculate the score.
- For the model we have trained and tested we got a value of R^2 =0.36476032257116975.

Coefficient of Determination →
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Sum of Squares Total →
$$SST = \sum (y - \bar{y})^2$$

Sum of Squares Regression →
$$SSR = \sum (y' - \bar{y'})^2$$

Sum of Squares Error →
$$SSE = \sum (y - y')^2$$

- Root mean squared error of the test data is 25.44855395442049

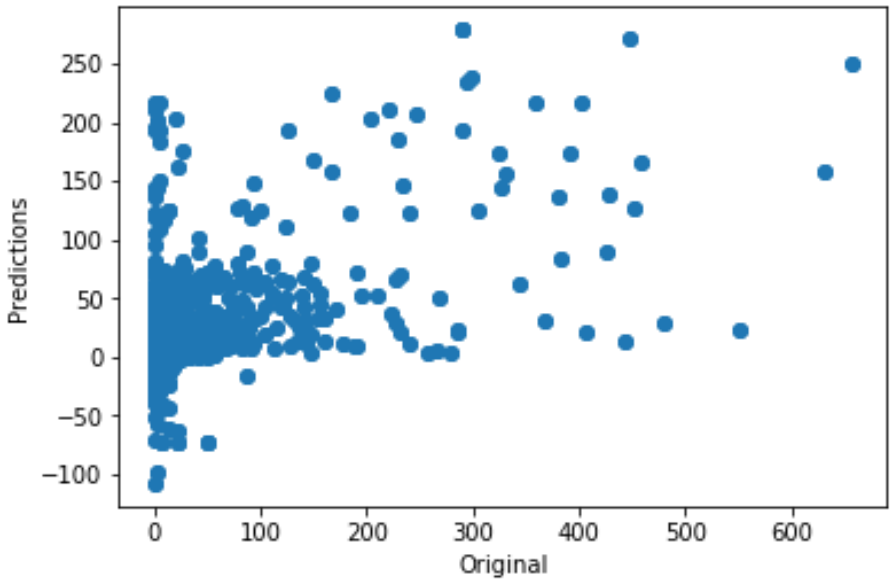
$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - p_i)^2}$$

- Correlation between actual data and predicted data is 0.5565369634165751

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\ s_x s_y}$$

2.Predictive Analysis: -

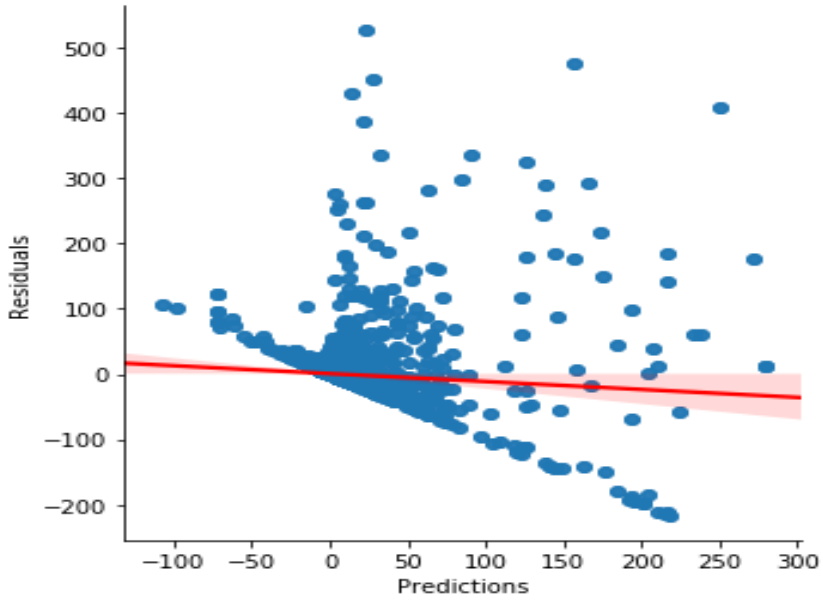
Original vs Predicted values



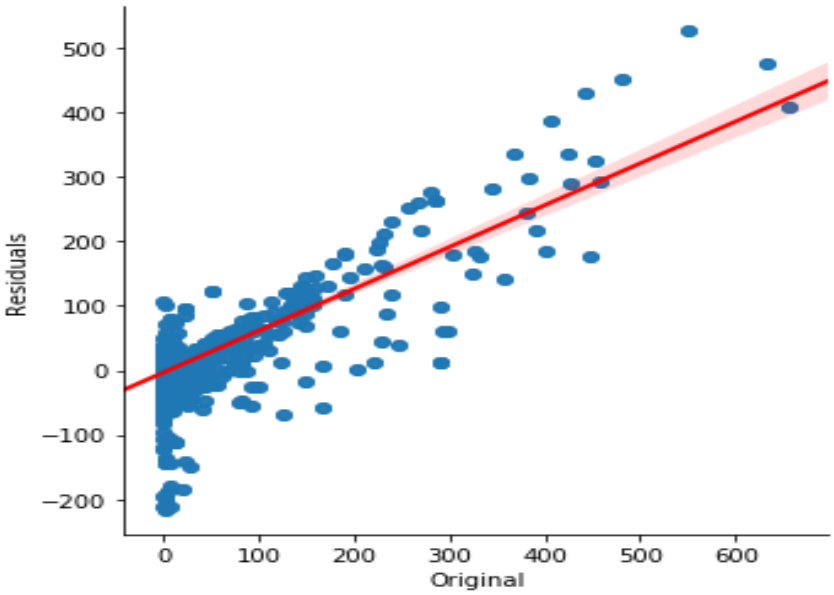
Test of Assumptions: -(Results and Discussion)

Heteroskedasticity and linearity: -

Residuals vs Predicted values



Residuals vs Original values:-

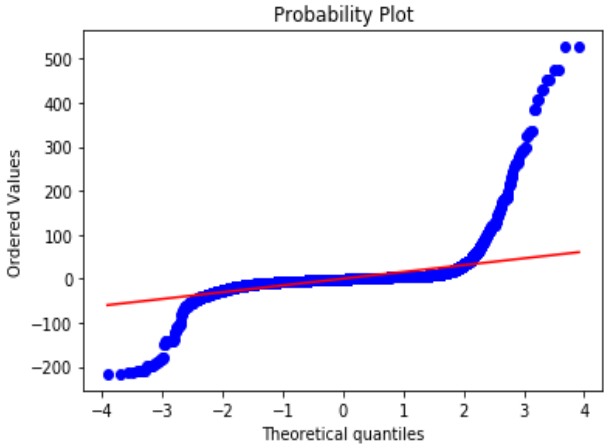


- From above residuals vs predicted values we can say that regression fit has heteroskedasticity because variance of residuals varies the fitted values.
- In above plot we have funnel type plot i.e., it has heteroskedasticity which is contradicting our assumption that our regression model should follow homoskedasticity.
- Remedy against violation of this assumption is to transform Y or apply Box-Cox method. Remedy for violation of assumption that plot of residuals and predicted values has non-linearity is to transform Y, X or both.

Normality check: -

- In case of Multiple linear Regression, errors follow Normal distribution.
- Here we use P-P plot for residuals to check distribution.
- We use 100(i-1/2)/n for P-P plot.
- If this Normality check isn't followed by model, then use Box-Cox method to solve normality problem.
- If you solve normality then, heteroskedasticity problem will also get solved.

P-P plot: -



Uncorrelated errors: -

We assume errors are uncorrelated. It should be true for regression model.

Durbin-Watson test

In order to check for uncorrelated errors we have to use Durbin-Watson test.

$$DW = \frac{\sum_{i=2}^{n} \hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}}{\sum_{i=1}^{n} \hat{\varepsilon}_i{}^2}$$

$$r = \frac{\sum_{i=2}^{n} \hat{\varepsilon}_i * \hat{\varepsilon}_{i-1}}{\sum_{i=1}^{n} \hat{\varepsilon}_i{}^2}$$

Correlation should be “0”

- DW = 2(1-r)
  - If r=0 then DW =2, No correlation
  - If r>0 then DW<2, positive correlation
  - If r<0 then DW>2, negative correlation
    r value: 0.0387522 , DW= 1.9224956

Remedy for Heteroskedasticity and Normality:

Box-Cox method: -

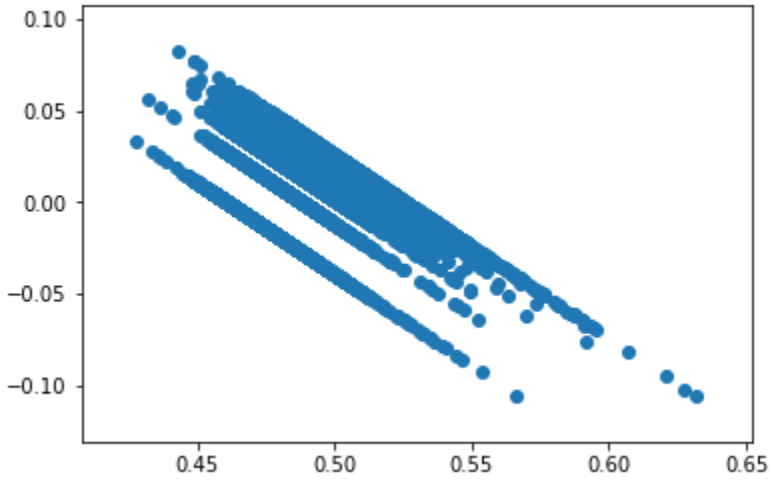
The Box-Cox transformation is a family of power transform functions that are used to stabilize variance and make a dataset look more like a normal distribution.

In this method, we'll transform dependent variable i.e., Y.

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y} \ln y, & \lambda = 0 \end{cases}$$

$$where\ \dot{y} = ln^{-1}[(1/n \sum_{i=1}^{n} ln\ y_i)]$$

- Choose a λ value such that it gives minimum SSE(sum squared error) and substitute into above transformation.
- We'll get transformed value of dependent variable.
- After transformation we have to apply regression on training data and new Y value. We'll plot predicted values vs residuals.



Conclusion: -

For given dataset, we have seen that our assumptions that are linearity, homoskedasticity and Normality are not obeyed, so we made some remedies and changed data made this model perform better. As, the results obtained were not satisfying we need to train this dataset with better model like Random Forest Regressor, etc..