Statistics for Data Science Assignment3

Navanith Rayavarapu

November 19, 2018

Abstract

The Assignment is about the Analysis of the Blog feedback data set which was downloaded from the UCI Machine Learning Repository. There are **281 attributes** in the data set. All the attributes are either real or integer. The last attribute is the target attribute. Each blog post (or observation) was selected such that the post was published 72 hours before a selected base time/date. In the train data The basetimes were in the years 2010 and 2011 whereas the test data set contains the basetimes in the year 2014.

Problem Definition

Each observation in the data belongs to a blog post. Our goal is to predict the number of comments in the upcoming 24 hours (relative to a basetime) for a given blog post.

Methodology

The Analysis of the data can be done by two ways:

1. Descriptive Analysis

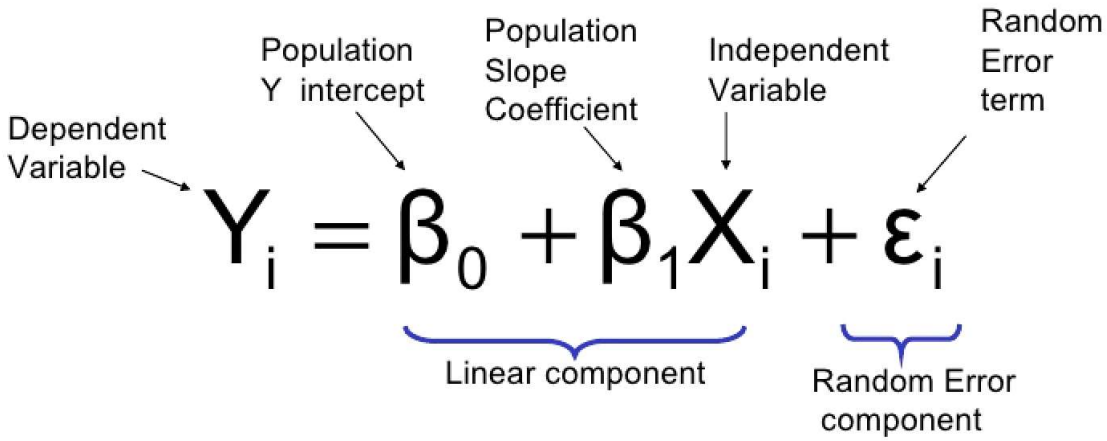
2. Predictive Analysis

In Descriptive Analysis we will find the

- number of blog sources.
- distribution of the target variable.
- number of blogs posted in each weekday.

In Predictive Analysis we will

- fit the regression line for for the data. The regression line formula is that



- find whether the fit follows homoscedasticity or in heteroscedasticity by drawing the plot between predicted values and the residuals.

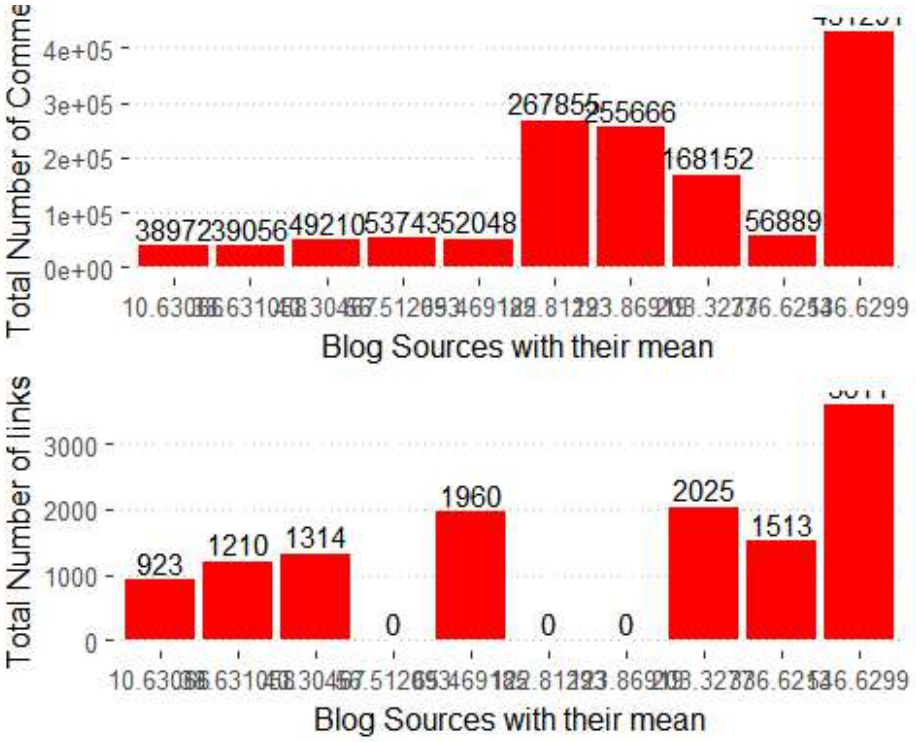
- find the performance metrics of the line. The performance metrics I used are
  1)R^2 value
  2)R^2 adjusted value
  3)finding rmse on the test data

Results and Discussion

Results of Descriptive Analysis

Number of blog Sources

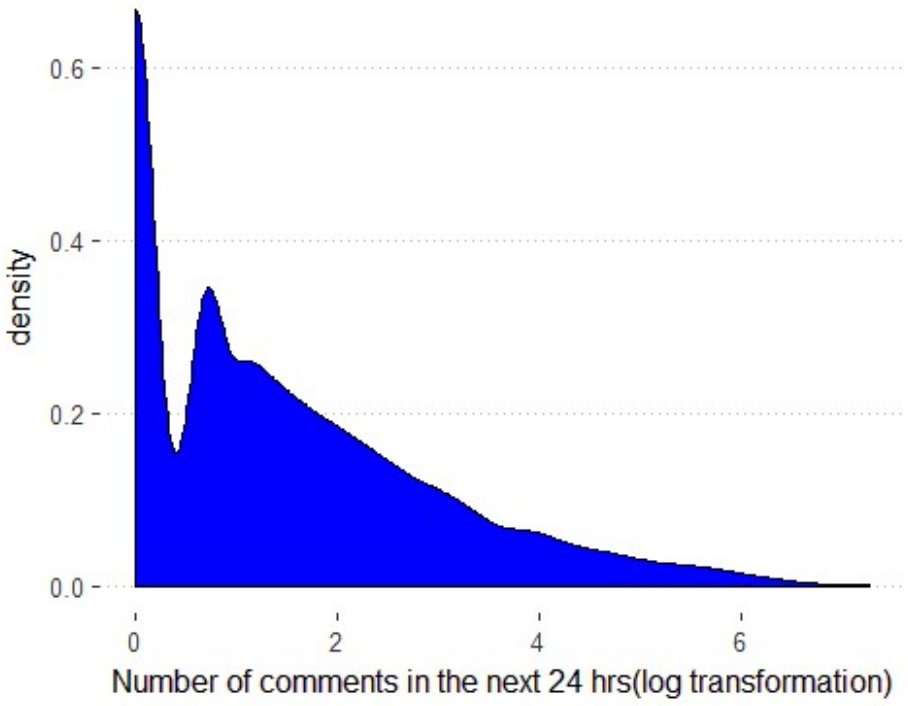
There are 433 blog sources in the train set. The following bar graph shows the top 10 (based on the number of comments) blog sources with their total number of comments.



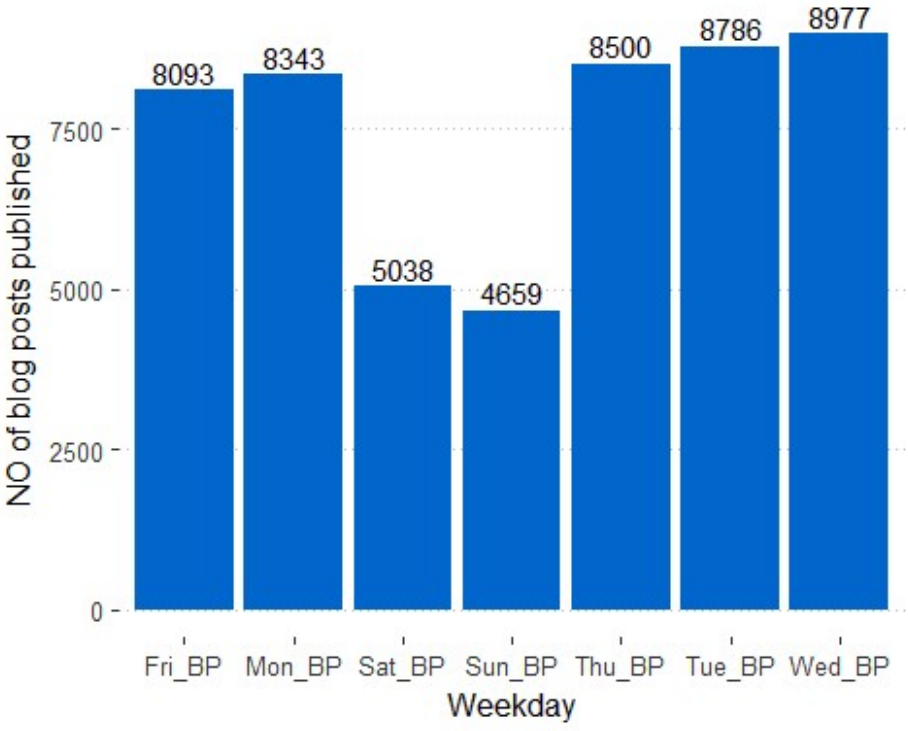
From the above bar graphs we can see that the blog source with mean (546.6299) of total number_comments has the highest number of comments and links. The number of blog posts in that blog source are 4314

distribution of the target variable

Following plot is the density plot of the log transformed target variable.

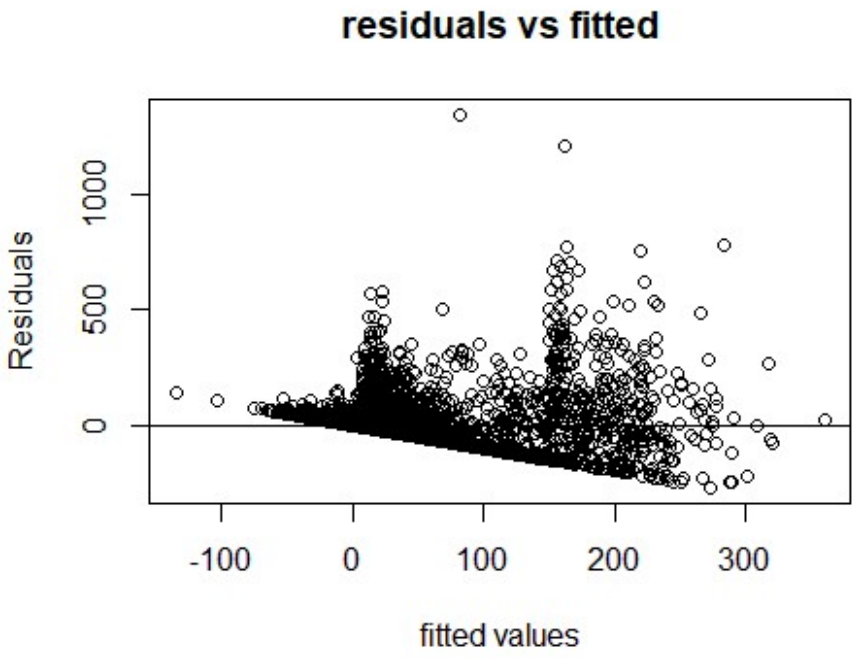


Number of blog posts in weekday

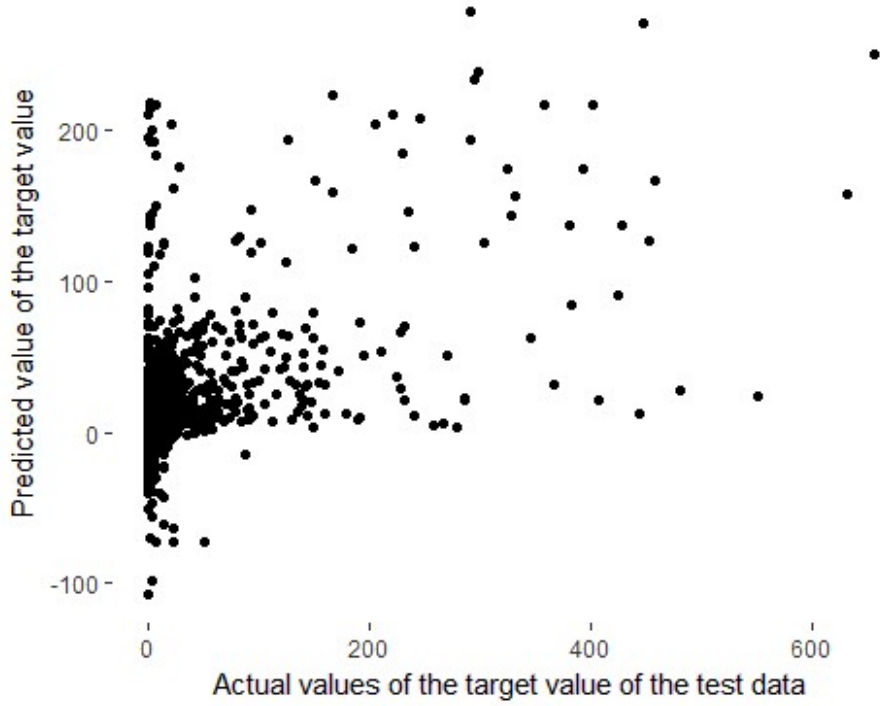


From the above bar graph we can say that, On saturday and sunday the number of blog posts published were less compared to the remaining weekdays.

Results of Predictive Analysis



From the above **residuals vs fitted** plot we can say that the regression fit has heteroscedasticity because the variance of residuals varies with the fitted value.



r^2 value of the fitted data is 0.3647607
adjusted r^2 value of the fitted data is 0.3617397
rmse of the of test data is 25.4476641
correlation between actual values vs predicted values is 0.55654

Conclusion

From the results of the regression line we can say that the data isn't good for the linear regression may be becuase there are redundant variables are there. And also most people are busy on saturday and sunday that they aren't interested to post blogs on these days.