**STATISTICS FOR DATA SCIENCE**

**ASSIGNMENT-3**

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

**Contents**

1. Abstract
2. Problem Definition
3. Dataset Description
4. Methodology
   1. Descriptive Analysis
   2. Predictive Analysis
5. Results and Discussion
6. Conclusion

***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: -***

***A close up of text on a white background

Description generated with high confidence***

***Methodology: -***

***1.Descriptive Analysis: -***

**A screenshot of a social media post

Description generated with very high confidence**

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.

**A screenshot of a cell phone

Description generated with very high confidenceA close up of a map

Description generated with high confidenceA screenshot of a cell phone

Description generated with high confidence**

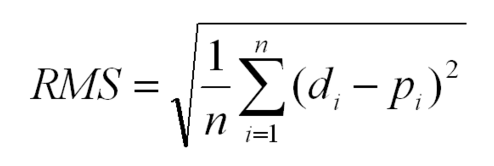
**Coefficient of Determination:**

* The **coefficient of determination** (denoted by R2) 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 R2 =0.36476032257116975.

A screenshot of a cell phone

Description generated with very high confidence

* Root mean squared error of the test data is 25.44855395442049



* Correlation between actual data and predicted data is 0.5565369634165751

A picture containing object

Description generated with very high confidence

***2.Predictive Analysis: -***

**Original vs Predicted values**

***A close up of text on a white background

Description generated with high confidence***

**Test of Assumptions: -(Results and Discussion)**

**Heteroskedasticity and linearity: -**

**Residuals vs Predicted values**

A close up of a map

Description generated with high confidence

**Residuals vs Original values:-**

***A close up of a map

Description generated with very high confidence***

* 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: -**

**A close up of a map

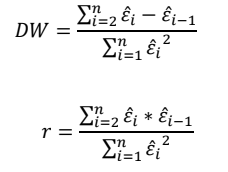
Description generated with high confidence**

**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.



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.**A screenshot of a cell phone

Description generated with very high confidence**

* Choose a lambda 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.

**A screenshot of a cell phone

Description generated with very high confidence**

**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..