# Facebook Analysis

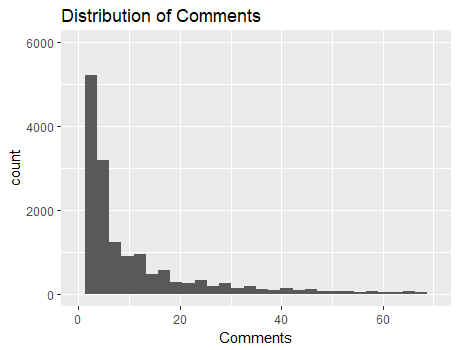
The Problem statement is to predict how many comments a post on a Facebook Page is expected to receive in the 24 hours after a given base time.

The Training Data Set has 63,830 data points and 41 variables including the dependent variable. The Dependent variable has been labelled as – “PsC24”. Some of the other Independent variables include –

1. Page Popularity
2. Page Views
3. Category of the Page Source/ Document
4. Day on which the Post was Posted

## Preliminary Data Analysis/Data Visualisation

Firstly, the Distribution of the Dependent Variable (No. of Comments “PsC24”) is analysed through a Histogram and the result obtained is as follows:



From the Histogram we can conclude that the Distribution of Comments is highly positively skewed as higher mass lies in the initial values and the tail is very long. Thus, clearly a high percentage of posts receive very few comments.

As the Comments can only take integer Values and the Distribution is clearly positively skewed thus it can be modelled using a Poisson Regression.

Next, the Distribution of Comments across different Categories is found out. The top 5 Categories in terms of comments is as follows:

1. Politician
2. Health/Medical/Pharmacy
3. TV Channel
4. Education Website
5. Actor/Director

Similarly the Categories having the lowest comments are as follows:

1. Cause
2. Sports/Recreation Activities
3. Home Décor
4. Undefined
5. Computers

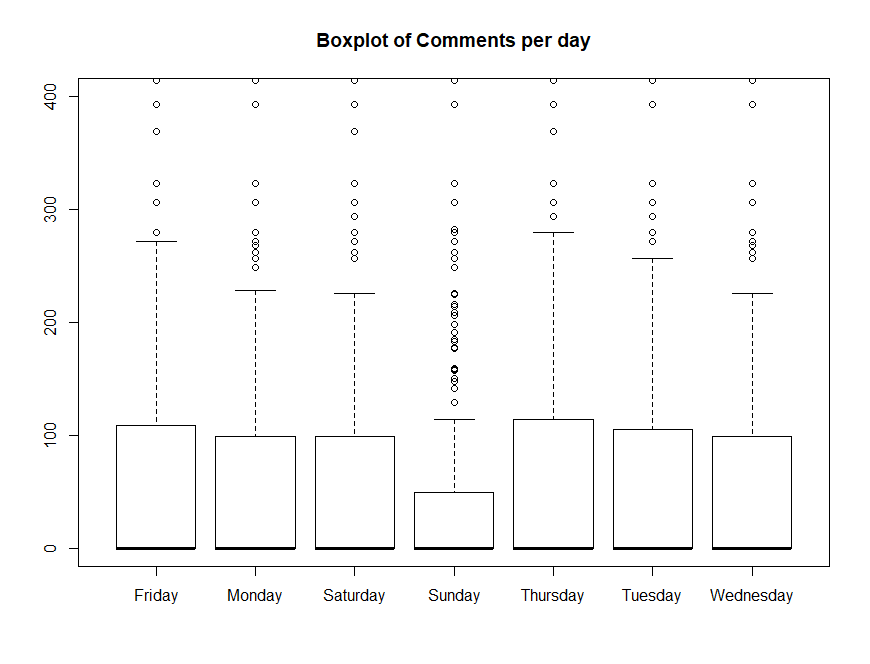
So, clearly the Categories which have the highest comments are also intuitively likely to have high comments. As Politicians and actors have large masses of people following them thus likely to gather a high number of comments.

The Frequency of Posts as per Days of Week:

1. Friday 7667
2. Monday 7567
3. Saturday 7132
4. Sunday 6226
5. Thursday 7567
6. Tuesday 7487
7. Wednesday 7418

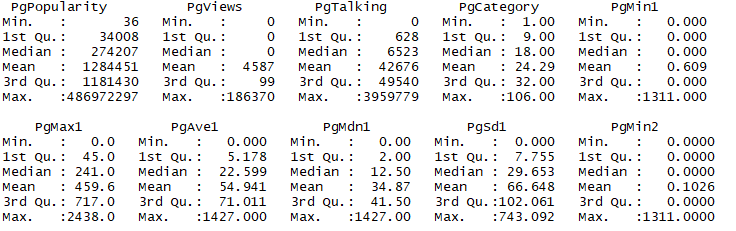
As we can see, the comments are mostly equally distributed across the week.

Next, the Distribution of Comments across each weekday is visualised with a boxplot. The result is as follows:



## Data Cleaning and Manipulation

A summary of a few variables is as follows:



From above we can clearly see that the maximum value and the third quantile for each variable are quite dispersed thus indicative of outlier values in the data set. Thus, for the modelling purpose we try to remove these extreme values.

For each Numeric variable I cut off rows where a particular variable’s value exceeds the 99th percentile.

## Model Training

The Data set was split into training and Validation data sets. The Model is fitted on the training data set and tested on the validation set.

Here we can see the results for the models I used:

### LASSO

The model significantly reduces the dimension of the data. Hence, the final model obtained is:

(Intercept) 5.423825960

PgAve1 0.005462312

PgMdn2 0.314267880

PsCLast24 0.157943220

PsCDiff 0.020810141

BaseTime -0.122112525

PsShares 0.001307933

The Mean Squared Error (the lower it is the better is the model) was 1172.

### Linear Regression/Stepwise

The Linear Regression using stepwise regression was performed and the final model obtained was:

PsC24 ~ PgPopularity + PgViews + PgTalking + PgCategory + PgMax1 +

PgAve1 + PgMdn1 + PgSd1 + PgMin2 + PgMax2 + PgAve2 + PgMdn2 +

PgSd2 + PgMax3 + PgAve3 + PgMdn3 + PgSd3 + PgMin4 + PgMax4 +

PgAve4 + PgMdn4 + PgSd4 + PgMin5 + PgMax5 + PgMdn5 + PgSd5 +

PsCBefore + PsCLast24 + PsCLast48.24 + PsCFirst24 + BaseTime +

PsShares

The Mean Squared Error (the lower it is the better is the model) was 1136.

This is the model I used to do my prediction. You can find the results in Pred\_Fb.txt

# Spam Analysis

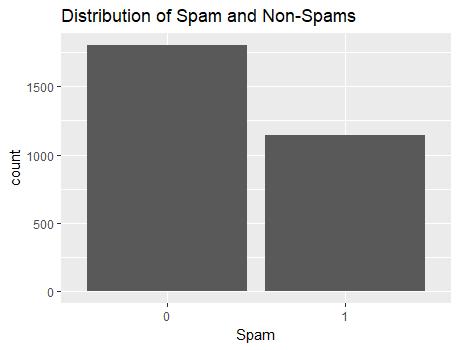
The Problem statement is to predict whether an e-mail received is a spam or not.

The Training Data Set has 3680 data points and 58 variables including the dependent variable. The Dependent variable has been labelled as “spam”. Some of the other Independent variables include:

1. Percentage of a particular word occurring in the E-mail
2. Uninterrupted sequences of capital letters
3. Percentage of signs occurring in the email.

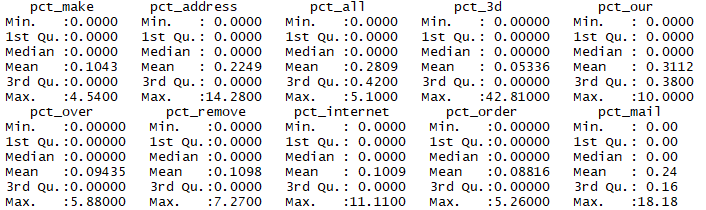
## Preliminary Data Analysis/Data Visualisation

Firstly, the Distribution of the Dependent Variable is analysed through a Bar Plot and the result obtained is as follows:



From the bar plot we can clearly see that the number of e-mails which are not spam are clearly higher than the number of emails which are spams.

The summary of some of the variables in the data set is as follows:



## Model Training

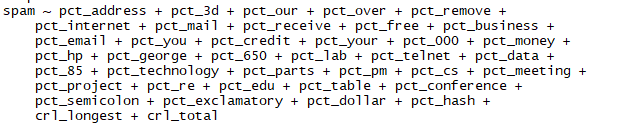
The data set is split into training and validation data sets. The training data set contains 80% of the observations and the rest 20% are in the validation data set.

### Logistic Regression

Since the problem is a Binary Classification problem a natural choice for such a problem is the Logistic Regression Model.

Next, a Backward Stepwise Regression is run to remove unnecessary variables from the model. The criteria used for model selection is the AIC.

The Stepwise Regression leads to a Model:

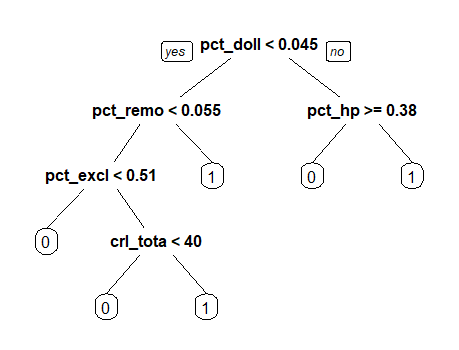


The number of variables reduces to 39 from 57.

### Decision Tree

The Classification problem as specified above can also be solved using a decision tree.

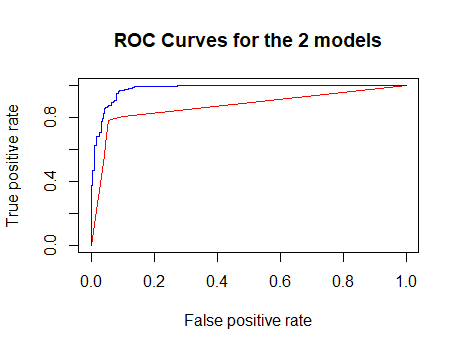
A Decision Tree was fitted to the data set and the result obtained is as follows:



The Important Variables identified from the Decision tree are:

1. Percentage occurrence of “doll”
2. Percentage occurrence of “remo”
3. Percentage occurrence of “hp”
4. Percentage occurrence of “excl”
5. Total length of uninterrupted sequences of capital letters

The ROC Curve for the two models is as follows:



The Blue line refers to the Logistic Regression and the red refers to Decision Tree Model.

The logistic regression has a higher area under the curve, so it has been chosen as model to do the prediction. You can find the results in Pred\_Spam.txt