Data Analytics

&

Data Driven Decision

A.A. 2017-2018



Facebook Analysis

Data Analysis on “Facebook Comment Volume” Dataset

Davide Mariotti, **255558**

Stefano Cortellessa, **254260**

Luca Grillo, **254****377**

Indice

[1. Rework on UML Model: 3](#_Toc507163143)

[2. Performance Requirements: 3](#_Toc507163144)

[3. Execution Graphs 3](#_Toc507163145)

[3.1 Purification Process 3](#_Toc507163146)

[3.2 Quality Checking 4](#_Toc507163147)

[3.3 Data Collector 4](file:///C:\Users\Stefano%20Cortellessa\Desktop\Univaq\Magistrale\1_ANNO\Software_Quality_Engineering\2ndHomework\Doc\Water%20Control%20System.docx#_Toc507163148)

[3.4 Antileakage Process 5](#_Toc507163149)

[3.5 Optimization Process 6](#_Toc507163150)

[4. Overhead Matrix 6](#_Toc507163151)

[5. Classes of Jobs 7](#_Toc507163152)

[6. Queuing Network 8](#_Toc507163153)

## Introduction:

The aim of this project is to do a deep analysis on a dataset in order to achieve a scope. In our case we have chosen the "Facebook Comment Volume" dataset, a set containing a lot of useful information regarding a Facebook post (like number comments, number of shares, etc.).

The aim of our analysis is to answer at the following question: *May the length of a post influence its reading and consequently also the number of comments and shares it will receive?*

For answering to this problem, we walked through two different preliminary phases:

* Data Cleaning of the dataset,
* Exploratory analysis.

After these we performed two others Machine Learning technics named *Supervised Learning* and *Unsupervised Learning* to predict which will be the shares for a post in the future.

First of all, we considered a significant number of datasets, each of which very interesting as in terms of topics covered as for their attribute’s configuration. We gradually discarded some of them, keeping all sets that at the same time had a better structure and aroused our interest.

In the end we opted for "Facebook Comment Volume", a dataset containing a lot of useful information regarding a Facebook post.

The most important reason we have chosen the Facebook dataset was the argument factor; we were looking for a set in which we could apply all concepts learned during the course, but at the same time the dataset had to capture our attention and it had to be stimulant for subsequent studies and analysis. A second criterion was dimension; we were aiming for a sufficiently large dataset, both regarding entries and attributes. In fact, working with a large number of information, we are able to produce more consistent results for a large number of studies, like explorative analysis, supervised learning and unsupervised learning. A third parameter for the choice was the structure; the dataset on which we desired to apply our work had to contain significant values, namely attributes able to adequately describe the application reality of interest.

Some important domain specific concepts are discussed below:

***Public Group/Facebook Page***: It is a public profile specifically created for businesses, brands, celebrities etc.

***Post/Feed:*** These are basically the individual stories published on page by administrators of page.

***Comment:*** It is an important activity in social sites, that gives potential to become a discussion forum and it is only one measure of popularity/interest towards post is to which extent readers are inspired to leave comments on document/post.

***Share:*** It is another important activity in social sites, that allow people to share posts so that other people can read and comment on these posts.

## Dataset Description:

In the following table the meaning of each column is explained:

|  |  |
| --- | --- |
| **Column Name** | **Column Description** |
| **Page Popularity/likes** | It is a feature that defines users support for specific comments, pictures, wall posts, statuses, or pages. |
| **Page Checkin's** | Describes how many individuals so far visited this place. This feature is only associated with the places eg: some institution, place, theatre etc. |
| **Page Talking About** | This is the actual count of users who are ’engaged’ and interacting with that Facebook Page. The users who actually come back to the page, after liking the page. This include activities such as comments, likes to a post, shares by visitors to the page. |
| **Page Category** | This defined the category of source of document eg: Local business or place, brand or product, company or institution, artist, band, entertainment, community etc. The category is defined by an integer number. |
| **C1** | Total comment count before selected base date/time. |
| **C2** | Comment count in last 24 hrs w.r.t to selected base date/time. |
| **C3** | Comment count is last 48 hrs to last 24 hrs w.r.t to base date/time. |
| **C4** | Comment count in first 24 hrs after publishing the document, but before the selected base date/time. |
| **C5** | The difference between C2 and C3. |
| **min C1** | Define the min of the variable C1 grouped by pages. |
| **max C1** | Define the max of the variable C1 grouped by pages. |
| **mean C1** | Define the Mean of the variable C1 grouped by pages. |
| **median C1** | Define the Median of the variable C1 grouped by pages. |
| **standard\_deviation C1** | Define the Standard deviation of the variable C1 grouped by pages. |
| **min C2** | Define the min of the variable C2 grouped by pages. |
| **max C2** | Define the Max of the variable C2 grouped by pages. |
| **mean C2** | Define the Mean of the variable C2 grouped by pages. |
| **median C2** | Define the Median of the variable C2 grouped by pages. |
| **standard\_deviation C2** | Define the Standard deviation of the variable C2 grouped by pages. |
| **min C3** | Define the min of the variable C3 grouped by pages. |
| **max C3** | Define the max of the variable C3 grouped by pages. |
| **mean C3** | Define the Mean of the variable C3 grouped by pages. |
| **median C3** | Define the median of the variable C3 grouped by pages. |
| **standard\_deviation C3** | Define the Standard deviation of the variable C3 grouped by pages. |
| **min C4** | Define the min of the variable C4 grouped by pages. |
| **max C4** | Define the max of the variable C4 grouped by pages. |
| **mean C4** | Define the Mean of the variable C4 grouped by pages. |
| **median C4** | Define the Median of the variable C4 grouped by pages. |
| **standard\_deviation C4** | Define the Standard deviation of the variable C4 grouped by pages. |
| **min C5** | Define the min of the variable C5 grouped by pages. |
| **max C5** | Define the Max of the variable C5 grouped by pages. |
| **mean C5** | Define the Mean of the variable C5 grouped by pages. |
| **median C5** | Define the median of the variable C5 grouped by pages. |
| **standard\_deviation C5** | Define the standard deviation of the variable C5 grouped by pages. |
| **Base Time** | Selected time in order to simulate the scenario. Decimal (0-71) Encoding. |
| **Post length** | Character count in the post. |
| **Post Share Count** | These features count the number of shares of the post, that how many peoples had shared this post on to their timeline. |
| **Post Promotion Status** | To reach more people with posts in News Feed, individual promote their post and these features tells that whether the post is promoted (1) or not (0). |
| **H Local** | This describes the H hrs, for which we have the target variable/ comments received. |
| **Monday** | Indicates if the post was posted on Monday (0-1). |
| **Tuesday** | Indicates if the post was posted on Tuesday (0-1). |
| **Wednesday** | Indicates if the post was posted on Wednesday (0-1). |
| **Thursday** | Indicates if the post was posted on Thursday (0-1). |
| **Friday** | Indicates if the post was posted on Friday (0-1). |
| **Saturday** | Indicates if the post was posted on Saturday (0-1). |
| **Sunday** | Indicates if the post was posted on Sunday (0-1). |
| **Monday\_Base\_Time** | Indicates the day on which the post was published on selected base date/time. |
| **Tuesday\_Base\_Time** | Indicates the day on which the post was published on selected base date/time. |
| **Wednesday\_Base\_Time** | Indicates the day on which the post was published on selected base date/time. |
| **Thursday\_Base\_Time** | Indicates the day on which the post was published on selected base date/time. |
| **Friday\_Base\_Time** | Indicates the day on which the post was published on selected base date/time. |
| **Saturday\_Base\_Time** | Indicates the day on which the post was published on selected base date/time. |
| **Sunday\_Base\_Time** | Indicates the day on which the post was published on selected base date/time. |
| **Target Variable** | The number of comments in next H hrs (H is the variable 'H Local'). |

The dataset can be found at this link:

<http://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset>.

## Data Cleaning

Data Cleaning is one of the most important part of the project. It’s a process able of guaranteeing, with a certain level of reliability, the correctness of a large quantity of data. So it’s responsible of a correct analysis since its aim is to clean the dataset from the values that doesn’t make sense, to check if they are correct with respect to the ‘logic’ of the dataset and to check if they are different from *null.* Furthermore, in this phase all duplicates are deleted and empty columns as well.

In this phase, the following operation have been performed:

* All the columns have been checked to be sure that there are not *null* values in the dataset,
* All the values have been checked to be sure that they are all positive, specifically Page Checkin’s, Likes, Page Talking About, C1, C2, C3, C4, Base Time, Post Length, Post Share Count columns have been controlled,
* All the values in the days columns have been checked to be 0 or 1,
* All the values in Base Data Time have been checked to be valid,
* All values inside C1 have been checked to be correct. In particular that:
  + C1 >= C2 + C3
  + C4 <= C1
  + C5 = C2 - C3
* All useless columns have been removed (Post Promotion Status) and
* All duplicated rows have been removed.

## Exploratory Analysis

## Supervised Learning

## Unsupervised Learning

## Conclusions

Working on the dataset, the goal we set ourselves, that is to see how much influence a short post (with a number of characters less than 400) compared to a long one (with a number of characters more than 400) has, has been achieved by our analysis thanks to the development of various phases. The first has been the one referred ***to clean the dataset***, *eliminating all those columns that were repetitive or empty and checking that all numerical values within it were good and not corrupted*, in order to conduct a correct analysis. Successively has been possible to go through the ***exploratory analysis***. This showed us that for every week-day, at most, *a number close to 6000 posts are published*. ***The day when there are more publications is Thursday***, so exactly in the middle of the week, while ***the day when there are fewer publications is Monday***, with a number close to 5000 posts. Analysis of the comments was carried out, which showed that ***the period in which more comments are made and more likes are placed is between 48 and 72 hours after the publication of a post***. An analysis was also carried out on the number of likes that a post receives on average, from which it was found that more than 7000 posts, out of a total of 40948, receive less than 400. Another analysis that was carried out was about categories. In particular, has been analysed ***the category that shows the most interactions, which was found to be 'Professional Sports Team'***. After analyzing all these 'secondary' quantities that are related to what is our main purpose, the final analysis was initiated and completed. To do this, the posts with more than 400 characters were initially separated from those with less than the same number, identifying them respectively as 'Long Posts' and 'Short Posts'. This showed that about 90% of posts have less than 400 characters. This already mean that a post defined as short is certainly preferred over one defined as long. Referring to this, it was also possible to note that the former are shared about 90% more. Finally, deepening the study, it was noticed that this is done more in the ***24 hours following the sharing of the post for short posts***, while ***between 24h and 48h for long ones***. In addition to the number of shares, the number of comments received from these two types of posts was also analysed. The results were almost identical, clearly showing that short length posts are:

* preferred by users,
* have more interactions.

For this phase we have also tried to give an ***'ideal' number of characters to obtain the largest possible number of iterations (represented by likes, shares and comments) which was found to be about 107***. Finally, ***supervised and unsupervised learning*** was performed. In the first case, two different Machine Learinig prediction methodologies have been implemented:

* 'Decision Tree' and
* 'Neural Network',

both with the aim of predicting the number of shares that a post will have by identifying 4 different classes:

* A **low number** **of shares** that indicates a number between 1 and 49 and it is identified by the number '1'
* A **medium number of shares** that indicates a number between 50 and 199 and it is identified by the number '2'
* A **good number of shares** that indicates a number between 200 and 699 and it is identified by the number '3'
* An **high number of shares** that indicates a number grater then 700 and it is identified by the number '4'

In both cases, an ***accuracy level close to 78% were achieved***. Concerning the Unsupervised learning, we adopted a ***clustering algorithm called K-Means Clustering*** that clustered our data points into a number (8) of mutually exclusive clusters. After determining the right number of clusters through the elbow method, we interpreted the 8-cluster solution and plotted them.  
The resulting scatterplot shows that the seven clusters excluding the one more spread out are dense pack meaning that the observation within the clusters are pretty highly correlated between each other and within cluster's variance is relative low. This overlap means that there is not good separation between these seven clusters. On the other hand, the cluster with center at position around (8, -0.7) shows a better separation and the observations are more spread out indicating less correlation among the observations and high within cluster's variance.