Application of Count data models on Facebook metrics dataset.

Malvina Chrząszcz

Bedil Karimov

Abstract

Social networks are not only useful in social life, but are also efficiently useful for the firms from different sectors for their marketing purposes. In this paper we made use of Facebook dataset which is comprised of the 19 count variables where we have 12 post performance metrics and 7 input variables. To model such kind of dataset the traditional generalized linear models, namely Poisson, Quasi-Poisson and Negative binomial where these were followed by Zero-inflated regression models. We choose one of the post's performance metrics, in our case comments, as dependent and 7 inputs as regressors as was suggested in the relevant paper. By this way we want to show that there is a positive relationship between comments and Paid type of advertising as well as Page total likes. Here we also have found that for our dataset Negative binomial model appeared to be more appropriate for us, taking into account the obstacles that we had faced with other models.

Introduction

Social media emerged as a communication medium which grows rapidly where people create, share, place and label contents. Nowadays, social media platforms like Facebook, Twitter and Instagram became a giants and reformed the social structure. Thanks to the ease of use, speed, wide accessibility these platforms has begun to create trends in many areas like the formation of communities, the environment, the economy, politics and technology (Ulusoy., H. 2014).
Since we think that social media represents an environment where society constantly exchange their thoughts and ideas, this information can be gathered easily and be used by the companies from variety of sectors to make better decisions regarding their products and services, thus increase their profitability. As the number of users on social media sites continues to increase, so does the need for businesses to monitor and utilize these sites to their benefit (Fan., Gordon., 2014).
According to Statista Dossier (2014), the number of social network users will increase from 0.97 billion to 2.44 billion users in 2018, predicting an increase around 300% in 8 years. Taking into account this high frequency development of the social media platforms and their usage,

companies started to emphasize and spend more effort on building insights from the social media turning this aspect of business into an additional asset. Companies also can improve their spending as stated in Deloitte Digital (2015), that according to their survey of 3000 US consumers, the digital interactions are going to have an impact of 64 cents per each dollar spent in retail stores, which is actually a substantial amount. Henceforth, estimating the impact of the advertisement became an important aspect of modern company's strategy. In the literature, you can notice some studies that focused on studying the impact of the social media publications and how the users react to those specific type of publications. However, you can find fewer studies which are dedicated to forming the predictive systems which are used to predict the evolution of a post prior to its publication (Mor., Rita., Val 2015). Making use of such predictive analytics of the published posts can benefit the company's view about a particular post and make better decisions. Thus, the studied explanatory tool can improve brand building.

Statistical modelling made in this paper can serve as the basis of getting some additional predictive knowledge from the social media data. Since we have a count data on hand, which is a normal case in economics and social sciences (Zeileis., Kleiber., Jackman., 2007) our analysis will be basically based on usage of Generalized linear models and Zero Augmented models. We focused on finding the impact of the posts published on company's page using different metrics of how those posts were published and what were the consequences. We believe that the information that was gathered and analyzed produces some knowledge for the companies to make some more consideration to make a more beneficial publication. For our modelling we used a data with 500 posts that were made on renowned cosmetics company's Facebook page that were made in 2014. Using this data as an input for our statistical modelling, our analysis will be based on two hypotheses:

i) There is a positive relationship between the page total likes and comments
ii) Positive interaction between Paid post and comments

In the second part of the paper we will describe the data, its basis and whole structure. In Part 3 we will present the models and the methods that were utilized to analyze the data especially the basics on the technical aspects of the our modelling procedure. Following this we will discuss the results and how one can improve their marketing (branding) decisions based on the insights found. At the end, we will point out the main conclusions and the investigations.

Dataset

The specified problem is based on the count data related to social media, namely Facebook. The dataset that we chose is "Facebook metrics dataset" which is from UCI. These metrics are related to the posts that were made on of the Cosmetics company's Facebook page in 2014 where it comprises of 500 posts published. According to Statista (2019) nowadays Facebook is the mostly used social media platform with nearly 2.4 billion active users. Overall the dataset contains 19 variables which are mentioned in .

Table 1 : Dataset's variables, variable types and data types

Variable	Type	Data type
Type	Categorization	Factor
Category	Categorization	Numeric
Page total likes	Categorization	Numeric
Post month	Identification	Numeric
Post weekday	Identification	Numeric
Post hour	Identification	Numeric
Paid	Categorization	Numeric
Lifetime post total reach	Performance	Numeric
Lifetime post total impressions	Performance	Numeric
Lifetime engaged users	Performance	Numeric
Lifetime post consumers	Performance	Numeric
Lifetime post consumptions	Performance	Numeric
Lifetime post impressions by people who have liked your page	Performance	Numeric
Lifetime post reach by people who like your page	Performance	Numeric
Lifetime people who have liked your page and engaged with your post	Performance	Numeric
Comments	Performance	Numeric
Likes	Performance	Numeric
Shares	Performance	Numeric
Total interactions	Performance	Numeric

The dataset is comprised of 2 typical features:

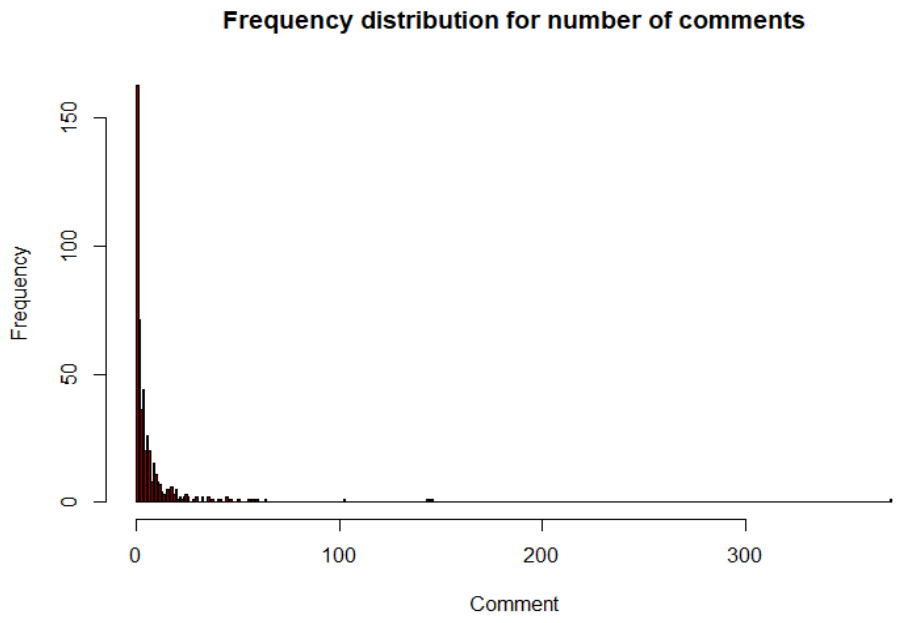
- Identification: variables which identify the posts
- Categorization: variables with the characteristics of the post
- Performance metrics: the variables showing the impact of each individual post

All of the variables mentioned above were exported directly from the company's Facebook page, except only two "category" and "total interactions". The former one presents the sum of the features that were directly exported from each post, namely comments, likes and shares. The latter one was created due to the request of the Facebook page managers. It basically reflect the information about the type of the campaign (i.e. link, status, video, photo) that was posted

on the company's page. It should be pointed out that the same dataset was used by the (Mor., Rita., Val 2015) where they used data modelling forming the predictive analytics.

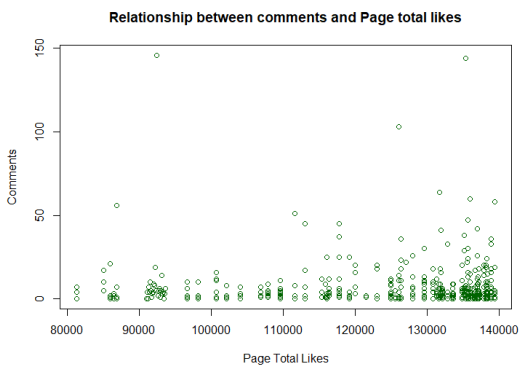
To better capture the data the following plots are presented. In the figures below we can grasp how our dependent variable is correlated with different set of predictors. It should be also mentioned that certain variables were factorized for analyses purposes. The plot below shows how the variable comment is distributed. It shows that there are large variations coupled with the large number of zeros. The focal point in the process of our analysis will be based on this step. It should also be mentioned that as we calculated the mean and variance, the latter was greater. Thus, as the numbers and the figure below suggest, we can conclude that there presents an overdispersion in our data, because it is believed that in Poisson model the variance and mean have to be equal to not have a dispersion problem with data.

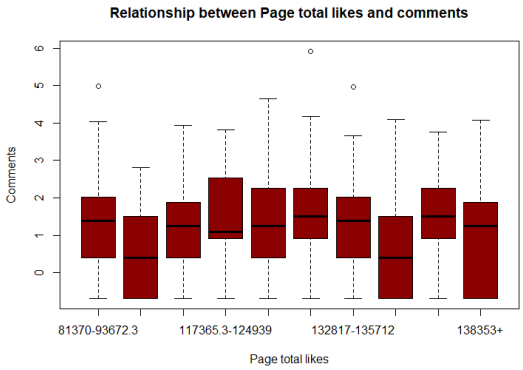
Figure 1: Frequency distribution for number of comments

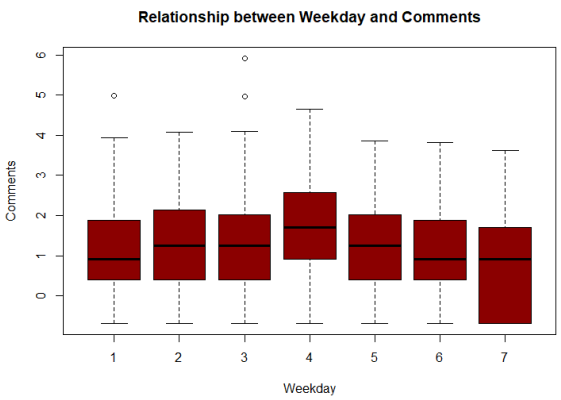


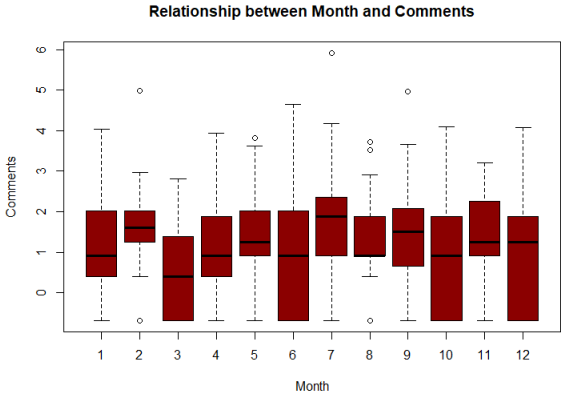
We proceed with the representation of the relationship of comments with respect to page total likes. From both of the plots below it can be deducted that there is no specific pattern that can be noticed between these two features.

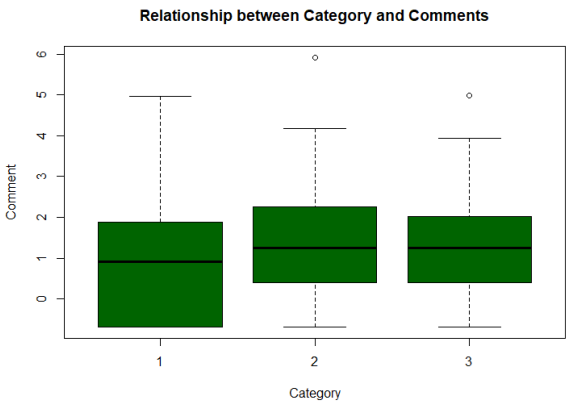
Figure 2: Relationship between Comments and Regressors

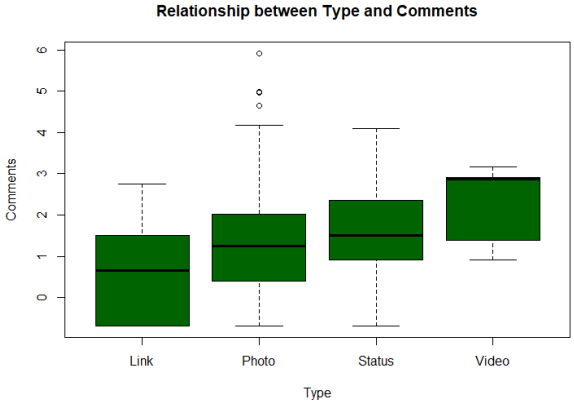


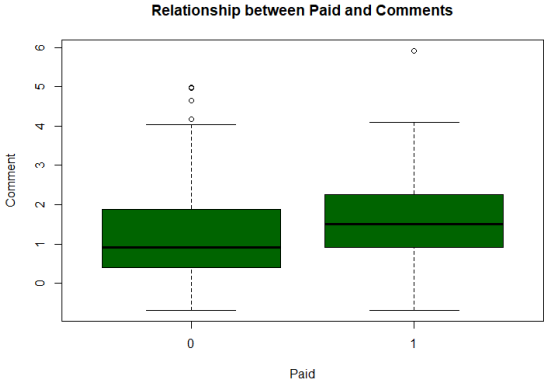












From other figures above we can notice and mention the following:

- On average there are more comments on the fourth day of the week, Thursday
- There is no deviating relationship between comments and month, where in July people appear to be more active
- There is an increasing relationship between Type of the post and comments, as the type of the post is increasing with respect to the size (i.e. photo to video)
- There is no substantial relationship between Paid and Category variables with comments

To avoid the high number of parameters during the regressing the models and for the sake of describing the overall results, we did two data transformations. For the Post Hour variable, we divided t into three groups (i.e. 0-8, 8-16, 16-24) and for the Post Month variable we divided them into 4 seasons respectively.

Methods and Models

In our short paper, we will make use of the Count Data Models which belong to the family of Generalized Linear Models (GLM). With the usage of these models you can analyze both linear and non-linear effects for any number and type of predictors with a discrete or continuous dependent variable. General Linear Models is comprised of two parts, the random part and the fixed part, as mentioned below. However in the Generalized Linear Models we have a third additional component which is the link (communication) function *g(.)* that transforms the expectation of a response into linear predictor:

General linear models:

i. $y_i \sim f(y_i|\theta)$

ii. $E(y_i) = \mu_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1i}$

Generalized linear models:

i. $y_i \sim f(y_i|\theta)$

ii. $E(y_i) = \mu_i = g^{-1}(\beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1i})$

Where $g()$:

iii. $g(\mu_i) = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1i}$

The analyses may include effects with many degrees of freedom for categorical predictors, effects with one degree of freedom for continuous predictors, and any combination of effects for continuous and categorical predictors. In the phase of building the models, for assessment and the testing of the hypotheses about their effects, the maximum likelihood method is used. Generalized linear models (McCullagh, Nelder, 1989; Venables, Ripley, 2002) make it possible to use various communication functions, the specific choice of which usually depends on the nature of the random distribution of the dependent y and its residuals. There are many possible combinations of binding functions in relation to distribution of the dependent variable, and several of them may be equally acceptable for simulated data. Therefore, when choosing a communication function can be guided by a priori theoretical considerations or by what the option "by eye" seems the most appropriate.

Since we possess with the count data, it is appropriate to proceed with using the models which are in line with the requirements for analyzing such data. The models which we will mention are Poisson, Negative Binomial and Zero Inflated Models.

Poisson model

This model is simply measures and predicts the number of occurrences of an event which can be expressed by:

$$P(x)= \lambda^x e^{-\lambda}/x! , \qquad x= 1,2,3\ldots.$$

Where λ is known as both the intensity parameter and the variance of the count. It also represents the amount of occurrences in a specific period of time.

$$\lambda = E[X]$$

$$\lambda = Var[X]$$

This case is also known as the equidispersion property of the Poisson model where mean equals variance. In practice, it is completely normal to face with the case of the overdispersion of the data, thus it is recommended to switch to the usage of the Negative Binomial Model (NB). However, in some situations you may encounter a case where there are more zeros than Poisson Model can predict. The solution for this is to use the Zero Inflated Poisson Model (ZIP).

Negative Binomial Model

The NB model appear to be more flexible compared to the Poisson Model. It is characterized as having greater mass to the right and the left of the mean. Moreover, it accounts for over- and under-dispersion, but at the cost of an additional parameter α where:

$$Var[y_i | x_i] = \lambda_i (1 + \alpha \lambda_i) > \lambda_i$$

Here we have three cases:

$\alpha > 0$ – overdispersion

$\alpha < 0$ – under-dispersion

$\alpha = 0$ – Poisson model

Hurdle models

The Hurdle model, which is also known as two-part model relaxes, is a model with two parts one generating the zeros (whether or not there is an event) and the positive (how many events) values. For instance, different features may effect whether or not you watch TV and how many times you watch TV in a week or a month. If $f_1(.)$ is the process resulting in zeros and $f_2(.)$ is the one resulting in positives, therefore g(y) can be specified as:

i. $f_1(0)$ if y=0,

ii. $(1-f_1(0)/1-f_2(0))*f_2(y)$ if $y \geq 1$

If the two process are equal (same) then this is the standard count data model.

Zero Inflated Poisson model

As mentioned before, there are some cases when there are more zeros then Poisson can predict, which is also known as the excess zero problem. ZIP model handles the excess zeros problem by letting the zeros to take place in two different ways, as a realization of the binary (k=0) and count processes (when binary variable k=1). For instance, you may like travelling or not. If you like travelling you may take one multiple amount of trips. But you may also like travelling and not have trip this year. In this way we are able to account and generate more zeros. If $f_1(.)$ is the process resulting in zeros and $f_2(.)$ is the one resulting in positives, then we can express zero inflated model, g(y), as:

i. $f_1(0) + (1- f_1(0))* f_2(0)$ if y=0,

ii. $(1-f_1(0))*f_2(y)$ if y ≥1

Modelling.

i. Poisson Model

As we implement Poisson model between comments and other regressors, what we have as the response are too optimistic significance levels of variables. This may arise due to the misspecification of the likelihood.

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.644e+00  5.918e-01  -7.848 4.22e-15 ***
Page.total.likes   4.938e-05  5.177e-06   9.539  < 2e-16 ***
TypePhoto          3.742e-01  1.391e-01   2.690 0.007141 **
TypeStatus         1.497e-01  1.494e-01   1.002 0.316520
TypeVideo          6.352e-01  1.754e-01   3.623 0.000292 ***
Category2          6.966e-01  4.680e-02  14.885  < 2e-16 ***
Category3         -4.592e-02  5.038e-02  -0.911 0.362059
Post.Month spring -5.787e-01  1.015e-01  -5.704 1.17e-08 ***
Post.Month summer -1.028e+00  1.365e-01  -7.534 4.92e-14 ***
Post.Month autumn -1.391e+00  1.543e-01  -9.013  < 2e-16 ***
Post.Weekday2      2.310e-01  7.223e-02   3.198 0.001383 **
Post.Weekday3      9.325e-01  6.390e-02  14.595  < 2e-16 ***
Post.Weekday4      4.488e-01  6.731e-02   6.668 2.60e-11 ***
Post.Weekday5      5.600e-02  7.359e-02   0.761 0.446693
Post.Weekday6      6.333e-02  7.343e-02   0.862 0.388483
Post.Weekday7     -4.819e-01  8.091e-02  -5.956 2.59e-09 ***
Post.Hour 8-16     1.386e-02  3.653e-02   0.379 0.704465
Post.Hour 16-24    4.448e-01  1.477e-01   3.011 0.002601 **
Paid1              4.673e-01  3.707e-02  12.605  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 7566.1  on 448  degrees of freedom
Residual deviance: 6196.5  on 430  degrees of freedom
AIC: 7434

Number of Fisher Scoring iterations: 6
```

As the next step we calculated the sandwich standard errors from which we can see that the variables *Type, Weekday, Hour and Paid* are no more significant. Moreover, the new standard errors for other variables seem more realistic.

```
           Estimate  Std. Error z value Pr(>|z|)
(Intercept)       -4.6445e+00  2.4639e+00 -1.8850  0.05943 .
Page.total.likes   4.9377e-05  2.1482e-05  2.2985  0.02153 *
TypePhoto          3.7425e-01  3.0485e-01  1.2276  0.21959
TypeStatus         1.4969e-01  5.0985e-01  0.2936  0.76906
TypeVideo          6.3524e-01  4.7941e-01  1.3250  0.18516
Category2          6.9660e-01  3.4976e-01  1.9916  0.04641 *
Category3         -4.5922e-02  2.0401e-01 -0.2251  0.82190
Post.Month spring -5.7875e-01  4.0494e-01 -1.4292  0.15294
Post.Month summer -1.0281e+00  6.4122e-01 -1.6033  0.10887
Post.Month autumn -1.3908e+00  6.7119e-01 -2.0721  0.03826 *
Post.Weekday2      2.3100e-01  2.9261e-01  0.7894  0.42985
Post.Weekday3      9.3254e-01  5.0787e-01  1.8362  0.06633 .
Post.Weekday4      4.4878e-01  3.1752e-01  1.4134  0.15754
```

```
Post.Weekday5      5.5999e-02  3.0203e-01  0.1854  0.85291
Post.Weekday6      6.3326e-02  2.9678e-01  0.2134  0.83103
Post.Weekday7     -4.8185e-01  3.0742e-01 -1.5674  0.11702
Post.Hour 8-16     1.3856e-02  2.1034e-01  0.0659  0.94748
Post.Hour 16-24    4.4480e-01  3.5814e-01  1.2420  0.21425
Paid1              4.6731e-01  3.1465e-01  1.4852  0.13750
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As we implement the Quasi Poisson model using the same variables, we have an estimated dispersion around 21, which suggests that there is an overdispersion problem.

ii. Negative binomial

As we move into the implementation of Negative Binomial model, which helps us to account for the overdispersion parameter, we end up with having only one insignificant variable which is *Type*. As shown in Table 2 there coefficients of the NB is different from the Poisson and Quasi Poisson. Also the extend to which the variables are significant are more substantial. Therefore, in terms of forecasted means Negative Binomial gives very different results.

iii. Zero-inflated Regression

A different way of augmenting the negative binomial count model with additional probability weight for zero counts is a zero-inflated negative binomial regression. It should be mentioned that to avoid facing the error regarding the conditional singularity, here we have avoided the Page total likes variable. As it can be seen from the Table 2 the numbers are significantly different from the previous models and it is also inferior to NB if we account for the likelihood parameter of the two.

Table 2: General table with the model’s coefficients

	P	QP	NB	ZIPois
(Intercept)	-4.6E+00	-4.64E+00	-5.64E+00	-
Page.total.likes	4.9E-05	4.94E-05	5.90E-05	1.02677
TypePhoto	3.7E-01	3.74E-01	5.41E-01	0.56716
TypeStatus	1.5E-01	1.50E-01	4.53E-01	0.24876
TypeVideo	6.4E-01	6.35E-01	1.02E+00	0.57662
Category2	7.0E-01	6.97E-01	4.66E-01	0.54906
Category3	-4.6E-02	-4.59E-02	1.40E-02	-0.17683
Post.Month spring	-5.8E-01	-5.79E-01	-8.91E-01	0.2827
Post.Month summer	-1.0E+00	-1.03E+00	-1.36E+00	0.20341

Post.Month autumn	-1.4E+00	-1.39E+00	-1.74E+00	0.02701
Post.Weekday2	2.3E-01	2.31E-01	7.28E-02	0.20733
Post.Weekday3	9.3E-01	9.33E-01	6.10E-01	0.93387
Post.Weekday4	4.5E-01	4.49E-01	4.73E-01	0.36046
Post.Weekday5	5.6E-02	5.60E-02	-1.39E-02	0.08256
Post.Weekday6	6.3E-02	6.33E-02	1.69E-02	0.03281
Post.Weekday7	-4.8E-01	-4.82E-01	-5.28E-01	-0.34884
Post.Hour 8-16	1.4E-02	1.39E-02	6.67E-02	0.02798
Post.Hour 16-24	4.4E-01	4.45E-01	4.44E-01	0.2956
Paid1	4.7E-01	4.67E-01	3.90E-01	0.35981
no. params	18'	18'	18'	17'
AIC	7434'		2605'	6664
log L	-3278'		-2565	-3278'

Conclusion

As we compare the models taking into account the outputs in Table 2**,** the result shows that there are great differences between the GLMs and the zero-augmented model. Among the GLMs as was mentioned above the Poisson and Quasi Poisson are more or less similar to each other, but Negative Binomial differs substantially as it accounts for overdispersion. The major difference that appears here is the change of the sign in third Category and the negative sign in Weekday 7. Overall we can mention that the estimated mean functions are different. Additionally to the mean functions we will use AIC and the likelihood parameters to account for the differences.

	ML-Pois	NB	ZIPOISS
logLik	-3698.000	-1283.000	-3278.000
Df	19.000	20.000	36.000
AIC	7453.046	2625.766	6664.739

Taking into account the AIC and likelihood parameters presented above, we can conclude that the Negative Binomial is outperforming the other ones. The negative binomial already improves the fit dramatically. This also reflects that the over-dispersion in the data is captured better by the negative-binomial-based models than the plain Poisson model. Additionally, it will be beneficial to mention how different set of models are taking into account the zeros. Here, the detected zero counts are contrasted to the expected number of zero counts solely for the models which are based on the likelihood.

Obs	ML-Pois	NB	ZIPOISS
97	10	111	97

The basic Poisson model is not appropriate in detecting and modelling the zero counts. However, the negative binomial model is much better in modeling the zero counts. The expected number of zero counts in the zero-inflated models matches the observed number. In summary, negative binomial model leads to the best results (in terms of likelihood) on this data set. In terms of zeroes prediction Zero-inflated Poisson model performs the best. Thus as a last step, we chose the Negative Binomial Model as the best one and we proceeded into the variable selection process, where we removed two variables, namely Type and Post Hour. As can be noticed both of the hypotheses that we had in the beginning do hold. We have a significant positive relationship between comment and the Page Total Likes. As we calculate the adjusted r squared for the model, it gives us a value of 0.13 which is very low and may suggest that there explanatory variables are negligible or insignificant. But we can also improve our model by increasing the sample size.

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.663e+00  1.793e+00  -2.602 0.009280 **
Page.total.likes   5.551e-05  1.639e-05   3.388 0.000705 ***
Category2          4.269e-01  1.680e-01   2.541 0.011049 *
Category3          7.503e-03  1.652e-01   0.045 0.963779
Post.Month spring -8.227e-01  3.510e-01  -2.344 0.019079 *
Post.Month summer -1.294e+00  4.725e-01  -2.738 0.006185 **
Post.Month autumn -1.703e+00  5.263e-01  -3.235 0.001216 **
Post.Weekday2      1.110e-01  2.527e-01   0.439 0.660369
Post.Weekday3      6.103e-01  2.540e-01   2.403 0.016260 *
Post.Weekday4      4.639e-01  2.502e-01   1.854 0.063730 .
Post.Weekday5     -2.344e-02  2.549e-01  -0.092 0.926739
Post.Weekday6      2.178e-02  2.429e-01   0.090 0.928527
Post.Weekday7     -5.275e-01  2.420e-01  -2.180 0.029284 *
Paid1              3.800e-01  1.482e-01   2.565 0.010320 *
```

Reference:

Fan, D., Gordon, M.D., (2014) *Unveiling the power of social media analytics*.
Communication of the ACM

Mor, S., Rita, P., Val, B., (2015) *Predicting social media performance metrics and evaluatio n of the impact on brand building: A data mining approach*
Journal of Business Research

Zeileis, B., Kleiber, C., Jackman, Simon., (2007) *Regression Models for Count Data in R*

McCullagh, P. and J. A. Nelder. (1989) *Generalized Linear Models*. Second ed.
London: Chapman and Hall.

Dossier, Statista (2014). *Social media & user-generated content—Number of global social network users 2010–2018*
Statista Dossier 2014.

Retrieved from:
from http://www.statista.com/statistics/278414/number-of-worldwide-socialnetwork-users/