Feedback Prediction for Blogs

Akash Barnwal
RU ID: 171004409

Varnith Chordia
RU ID: 175009968

Abstract

In this paper we try to predict the number of comments a blog post will receive depending on the past comments on the blogs. We have centered our analysis to two statistical methods - linear regression and poison regression. Using linear regression we used different selection methods and transformations to achieve the best candidate model. In this paper we have shown methods of stepwise selection. We applied a Poisson regression as well to the dataset and compared the prediction performance with linear regression.

1 Introduction

Feedback is a gift. With the world caught in the whirligig of technology and social media, feedback is a blessing in disguise for any business. Facebook comments, Twitter post, YouTube videos are much more than entertainment. Companies would inevitably need to understand the types of content that is likely to elicit the most user engagement as well as any underlying patterns in the data such as temporal trends.
The project revolves around with the idea of predicting the number of comments a blog would receive depending on the past comments. The dataset under consideration is available on "UCI Machine learning repository".
The paper proceeds as follows: Section 2 includes a short Literature Review; Section 3 describes the dataset used, Section 4 discusses the Experiment, Section 5 discusses the Result and Discussions, Section 6 discusses the Conclusion and Future Work.

2 Literature Review

This project delves into same dataset what Buza has taken into consideration for analysis. He has considered two performance metrics: Area under the Cure and blog post forecasting of top ten blog pages with highest number of feedbacks. Buza inspects various models such as a *multilayer perceptron model, RBF-networks, regression trees (REP-tree, M5Ptree)*.
Our analysis is focused on Linear Regression and a more generalized linear model for count termed "*Poisson Regression*". The constraint with linear regression was it was predicting negative counts of comments for a post. We tried various methods to stabilize the data through *log transformation, square root transformation* but couldn't achieve the assumptions. Hence we extended our analysis to Poisson Regression. Since Poisson regression is used to model count data, we expect it to perform better than Linear Regression.

3 Dataset

The data set is available at *UCI Machine learning repository* and comprises of list of train dataset and test datasets. The train data contains *52,397* observations with *281* features associated including target variable (number of feedbacks). The following features are present in the dataset

- **Basic Feature**: Number of links and post in the previous 24 hours, number of links and post in previous 48 hours, summary statistics (mean, median, standard deviation) of each of the post.
- **Textual Feature**: The most discriminating bag of words used in blogs.
- **Weekday Feature**: Binary Indicator feature that describes which day of the week the blog was posted.
- **Parent Feature**: A page X is a parent of page Y, if Y is a reply to X. Parent features are number of parents, minimum, maximum and average number of feedbacks that the parents received.

The train dataset comprises features of the blog, extracted from the year 2010-2011, and we have considered test data for Feb-2012.

4 Experiments

The effectiveness of Linear Regression and Generalized Linear Model for counts on different types of features was explored throughout the project. We trained the models on train data and predicted on test data. The data contained

4.1 Linear Regression

For linear regression, Adjusted R-Square was used to assess the fit of different models. We built a linear regression model on the entire training data for basic analysis. The model had an adjusted R-square of *.367*. This does not show a great fit so transformation of the response variable and variable selection deemed important. The different transformations that were taken were square root and log transformation. The transformations did not show any significant changes in the adjusted R square. The model assumptions of constant variance and normality were still being violated.

To reduce the dimensionality in the data, we selected significant variables from the 276 features present. Stepwise selection method coupled with Bayesian Information Criteria (BIC) was appropriate. BIC is useful as it tends to hone in one model as the number of observations grows. As a result BIC picks a smaller model compared to Akaike information criterion (AIC). Due to large computation time in running the stepwise selection method, we considered random sampling.

Random sampling is a method of selecting random observations from a population dataset. Each observation has an equal chance of getting selected thus eliminating any chance of bias. In R before taking any random sample we assign set. seed (n) so that the same dataset is used across different models for comparison. In this method a random sample of *10,000 observations* is taken which is *~20%* of the original population data. We ran four models for different types of features of the data.

The 3 types of data on which the regression model was implemented are as followed:

- Basic Features-Retained only basic features from dataset
- Basic & Parent Features-Retained basic and parent features
- Basic & Weekday Features-Retained basic and weekday features

The final model was built on the entire dataset. The segmentation helped to identify the effect of features in the model.

4.2 Linear Regression vs. Poisson Regression

Linear regression works well with a continuous response variable. Predicting the count of comments for a blog in next 24 hours is a discrete variable. Linear regression assumes the normal distribution of errors and a transformation of response variable can help achieve normality. However the above models did not achieve any sort of normality, this is because the response is skewed towards zero (more than 50% of the blogs had 0 comments). Secondly, it predicts negative values for count data which is theoretically not possible. A Poisson model is similar to an ordinary linear regression with two exceptions. First, it assumes that the errors follow a Poisson distribution and not a normal distribution. Second, rather than modeling Y as a linear function of the regression coefficients, it models the natural log of the response variable ln(Y) as a linear function of the coefficients.

Distribution of Target Variables

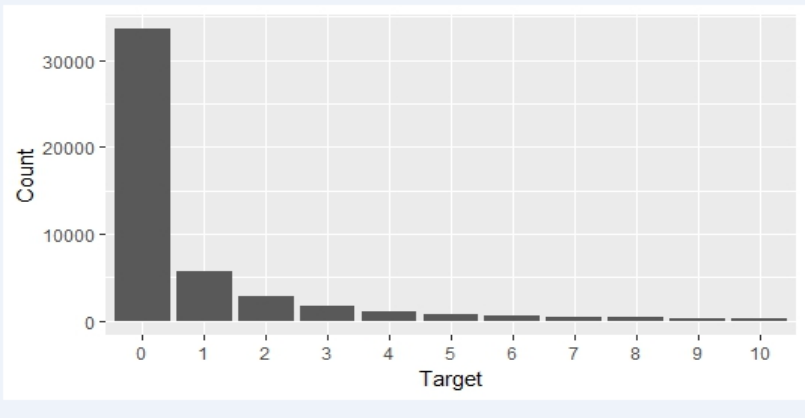

Figure 1

4.3 Poisson Regression

Poisson regression models are generalized linear models with the logarithm as the (canonical) link function, and the Poisson distribution function as the assumed probability distribution of the response. The response variable in Poisson regression is a count (Y) or can also be the rate (Y/t) where 't' represents the time or space. It is used to model count data or contingency tables.

Generalized linear model for counts with assumptions:

$$g(\mu)=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k = x_i^T\beta$$

Random component: Response Y has a Poisson distribution that is $y_i \sim$ Poisson (μ_i) for i=1... N_i where the expected count of y_i is E(Y) =μ.

Systematic component: Any set of $X = (X_1, X_2, \ldots X_k)$ are explanatory variables.

***Natural log link*:** log (μ) = α + βx for a single explanatory variable.
This is referred to as "Poisson log linear model". This can be written as $\mu_1 = e^{\alpha}e^{\beta x}$

Interpretation of Parameter Estimates:

exp (α) = effect on the mean of Y, that is μ, when X = 0
exp (β) = with every unit increase in X, the predictor variable has multiplicative effect of exp(β) on the mean of Y, that is μ

- If β = 0, then exp (β) = 1, and the expected count, μ = E(y) = exp (α), implies Y and X are not related.
- If β > 0, then exp(β) > 1, and the expected count μ = E(y) is exp(β) times larger than when X = 0
- If β < 0, then exp(β) < 1, and the expected count μ = E(y) is exp(β) times smaller than when X = 0

Poisson regression can be used for modeling count data. Using the same methodology of fitting four different models in linear regression, Poisson method was implemented using stepwise selection method coupled with BIC.

4.4 Naïve Method

This is one of the simplest method of estimation. We took the mean across the response which is assumed as the predicted value of the entire response. In financial data sometimes the mean of the response maybe considered as the best predicted value. Hence we took this into consideration.

5. Results & Discussions

We have used RMSE (Root Mean Square Error) method to evaluate the prediction of the two methods on the test data. The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. R-squared is a relative measure of fit whereas RMSE is an absolute measure of fit.

The below table summarizes the output:

Type of regression	Basic	Basic+ Parent	Basic+ Weekday	All Features
Linear	9.22	9.21	9.35	10.22
Poisson	10.17	10.12	10.64	10.16

It can be seen that the least RMSE was observed when we regressed least squares on only basic features. But the difference was not greatly significant between Poisson and Linear Regression. This is because the response variable present in test data were mostly skewed having zero comments

and there was no blog with more than 30 comments present. Due to which the RMSE seems to be hardly differentiable.

Comparing the naïve method to the regression, the regression methods performed significantly better than naïve model which had an RMSE of 15.59.

Each model had different set of variables which were significant. Some of the important variables that were present across different models were length of the blogpost, time of blogpost and number of links etc.

6. Conclusions and Future Work

With the output shown, we can see that Poisson regression can be used for discrete response variable. There are certain feature learning methods such as SVM, Boosting and Random Forest which can be used to make a comparison. We can extend the work by trying certain other combinations of features to run this model too.

Acknowledgement: We thank Prof Han Xiao for his guidance throughout the project.

References:

[1] Buza, Krisztian. "Feedback prediction for blogs" Data Analysis, Machine Learning and Knowledge Discovery. Springer International Publishing, 2014. 145-152.

[3] Robert Tibshirani, Guenther Walther and Trevor Hastie. "Estimating the number of clusters in a data set via the Gap statistic." Journal of the Royal Statistical Society, B, 63:411-423, 2001.

[4] Olshausen, Bruno A. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images." Nature 381.6583 (1996): 607-609.

[5] Thai T. Pham, Camelia Simoiu "Unsupervised Learning for Effective User Engagement on Social Media"

Appendix:

Steps Followed:

- The data has *281* features out of which we removed *4* variables since all the values were 0. There were *214* categorical variables and *63* continuous variables.
- We ran different selection method using linear regression model on all the remaining features as followed:

Type of Regression Model	Features	Adjusted R square
Square Root	Basic + Parent	0.3801
Normal	Basic + Parent	0.3801
Square Root Step wise	Basic + Parent	0.3813
Normal Stepwise	Basic + Parent	0.3813
Square Root	Basic	0.3590
Stepwise BIC Square Root	Basic	0.3592
Normal	Basic	0.3500
Normal Stepwise	Basic	0.3592
Square Root	Basic + Weekday	0.3804
Square Root Stepwise	Basic + Weekday	0.3817
Normal	Weekday	0.3804
Stepwise	Weekday	0.3817

- **Plot Results:**

<u>Residuals vs. Fitted without Transformation</u>

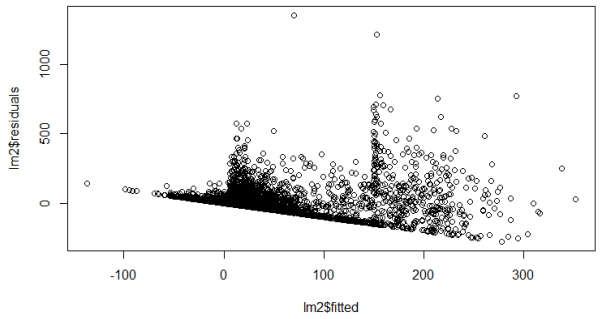

Figure 2

There is a violation of constant variance in the given plot.

Residual vs. Fitted With Square Root Transformation

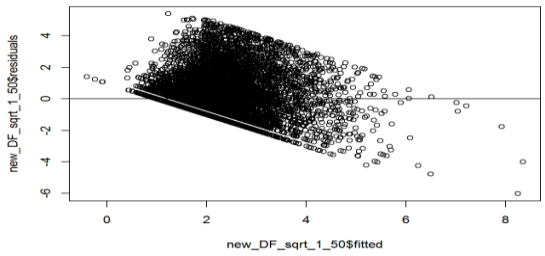

Figure 3

Residual vs. Fitted with Log Transformation

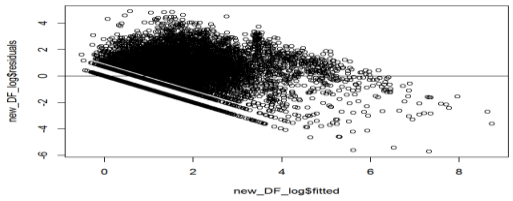

Figure 4

QQ plot of Residuals

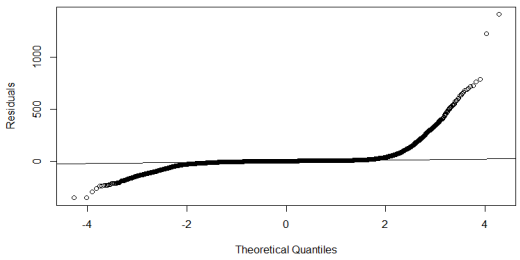

Figure 5

- There is a violation of normality.