Project Title: **FACEBOOK COMMENT VOLUME PREDICTION**

Team Members: Yixi Chen (yca3160), William Ehrich (wde220), Chungyuen Li (clt2503)
Email Addresses: ecchen96153@gmail.com, wdehrich@gmail.com, lizoyu1@gmail.com
Course: EECS 349 - Machine Learning, Northwestern University, Spring 2017

Abstract

In our study, we seek to create a model to predict the number of comments a Facebook post will receive based on features of that post and features of the Facebook page on which the post was made. This endeavor is of great important from a scientific perspective because of the great potential to understand the thoughts and feelings of people based on their behavior on social media. From a commercial perspective, companies and other institutions use social media for marketing purposes. Therefore, greater insight into social media activity can lead to more effective marketing strategies. Experimentation was performed using three regression models: (1) regression tree, (2) neural network, and (3) Naïve Bayes. Additionally, single feature baselines were calculated and data transformations were performed in attempts to increase accuracy.

To measure accuracy we used three metrics: (1) Hits@10, (2) proportion of predicted events with a factor of 2 of the actual results, and (3) the mean squared error. The Hits@10 score is determined by providing the model with 100 feature vectors, generating predicted values and counting how many of the feature vectors in the predicted top 10 are in the actual top 10 with respect to comment volume. Our findings revealed that, regression trees and neural networks were able to yielded impressive Hits@10 scores of around 7. By comparison Naïve Bayes analysis yielded lower accuracies. Additionally, it was observed that the number post comments was best predicted by the comments on a corresponding page in the preceding 24 hours. Ultimately, user controllable features did not have impact on prediction accuracy and data transformations did not have a significant impact on prediction accuracies.

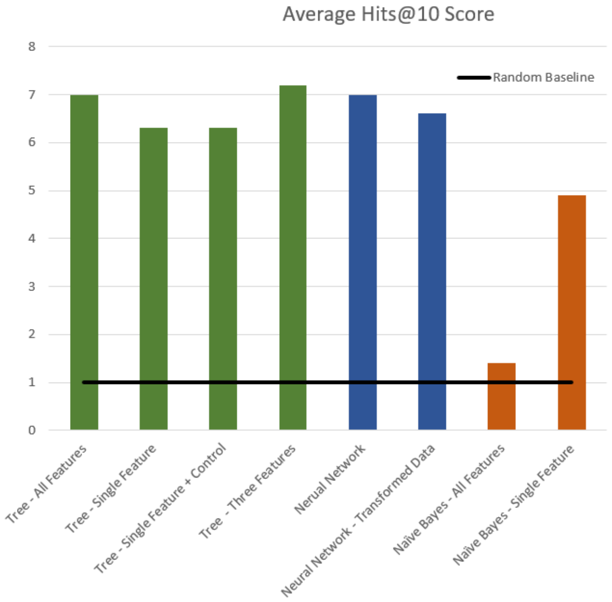

Figure 1: Average Hits@ 10 Scores for different combinations of regression models and feature selections

Introduction

Task:

The goal of our project is to predict the number of comments a Facebook post will receive based on characteristics of the page and information about the post. This pursuit is of great potential value because the number of comments on a Facebook post can be used as a proxy for interest in the subject of the page and relevance of the post's content. Therefore, by formulating a model to predict the number of Facebook comments based on page and post information, one can gain insight into the thoughts and feelings of people active on social media. This insight can, in turn, be used by advertisers, marketers, as well scientists studying social media.

Approach:

The data source for this project is information regarding a set of Facebook posts compiled by Kamaljot Singh, who obtained the data using Facebook's API for data crawling and donated the resulting data set to the University of California, Irvine Machine Learning Repository. The data set includes, for its training set, the number of comments, as well as 53 input attributes, of which 28 are base attributes and 25 are derived attributes, for 40949 sample Facebook posts.

In our investigation, we use three learners: regression trees, neural networks, and naïve bayes to predict the number of comments on a post based on the input attributes of the post. To measure the accuracy of our regressors, we use three metrics: (1) Hits@10, (2) mean squared error, and (3) the proportion of predicted values within a factor of two of the actual values. To calculate the Hits@10 metric, we take a test set of 100 example Facebook post, ranks the posts by the predicted number of comments and count how many of the posts in the predicted top ten are in the actual top ten.

Results:

Using Hits@10 as our most important accuracy metric, we determined that regression trees and neural networks yielded higher accuracies than did the Naïve Bayes regressor. In particular, we were able to obtain hits at 10 scores of 7.0 for both. To further investigate the data, we performed feature selection on the regression tree and found, somewhat unsurprisingly, that the best indicator of the number of comments on a post is the total number of comments on the corresponding Facebook wall in the most recent 24 hours. In an attempt to see if our study could generate actionable information, we examined the effect of post character lengths and days of posting, and found that we could increase the number of predicted values within a factor of two of the actual values, but could not increase the Hits@10 score. We also sought to increase the Hits@10 score by performing transformations on a single input feature baseline, but did not see increases in accuracy.

Experimental Design

Data

The data in the repository included the number of comments over variable amounts of time. To standardize our approach, we only included measurements of comments over 24 hour time periods. For our training set, we used 24-hour data from one set of data and randomly selected 10 sets of 100 instances of 24-hour data from another data set. We then performed 10 experiments and averaged the results.

Regression tree

To create a regression tree, a python project was created and the scikit-learn python library was utilized. To prevent overfitting, the maximum depth of the regression tree was set to 5.

Neural network

The neural networks were created from multilayer perceptrons and were implemented using the MLPRegressor in sklearn.neural_network. Because neural networks often yield inaccurate results when there is a high variance of training input attribute value ranges, the input feature vectors of the training set were scaled so that input values ranged from 0 to 1. Next, the input feature vectors of the testing sets were modified using the same scale. To maximize accuracy and performance, a number of parameters were set. First, the hidden layer size was set to (150, 5). As a result of this specficiation, there were two layers, in which the first layer had 150 perceptrons and the second had 5 perceptrons. We specified two layers because any function can be approximated to arbitrary accuracy by a network with two hidden layers [Cybenko, 1988], and it was found to achieve good accuracies from trial-and-error. Second, the learning rate was set to 0.001 as default to avoid learning too fast. Finally, early stopping was turned on, in which 10% of training data was used as validation set and its error was used to decide if and when to terminate the training.

Naïve Bayes

To create our naïve Bayes regressor, a Gaussian naïve Bayes classifier was used. In this approach, regression was performed by using the classifier to predict the most likely output value in the training set for every input vector in the testing sets.

Results

Regression Tree

When all input attributes were used in building the regression model, a Hits@10 score of 7.0 was obtained. An analysis of feature importance revealed that the greatest predictor of comment volume was the number of comments on a Facebook page in the most recent 24 hours. A single feature baseline accuracy was calculated by using only that input attribute for training and testing. The result was a surprisingly high Hits@10 score of 6.3. After obtaining this result we were interested to see if any user-controllable input attributes, such as post character length and day of posting, could have an impact on the number of comments. To do so, we added those input attributes to the single-feature baseline and observed no increase in Hits@10, but a decrease in mean squared error. It appears that user-controllable inputs have little predictive value concerning relative comment volume, but can increase the precision of predicted values. As an additional note, an attempt was made to increase the accuracy of the single-feature baseline by transforming the values of the input attribute, but not transformation led to an increase in Hits@10. The transformations attempted were as follows: square, cube, square root, reciprocal, natural logarithm.

Table 1: Regression Tree Results

Regression Model	Hits@10	Within Factor of 2	Mean Squared Error
Tree - All Features	7.0	0.806	8303.93
Tree - Single Feature	6.3	0.837	10508.93
Tree - Single Feature + Control	6.3	0.838	10343.72387
Tree - Three Features	7.2	0.804	8495.39

Neural network

Using all the input attributes to train and test, the neural network achieved an average Hits@10 score of 7.0 and an average mean squared error of 6534.32. In an attempt to increase accuracy, all of the attributes were individually transformed. After the transformations, the average mean squared error decreased and the number of predicted values within a factor of 2 of the actual values increased, but the average Hits@10 score actually decreased to 6.6.

Table 2: Neural Network Results

Regression Model	Hits@10	Within Factor of 2	Mean Squared Error
NN	7	0.817	6534.32
NN - Transformed Data	6.6	0.853	5803

Naïve Bayes

When all the input attributes were used to train and test the Naïve Bayes regressor, an average Hits@10 score of 1.4 was obtained. This score was barely above the random baseline accuracy of 1.0. However, after isolating the feature with the greatest predictive value, we were able to generate an average Hits@10 score of 4.9 using Naïve Bayes.

Table 3: Naïve Bayes Results

Regression Model	Hits@10	Within Factor of 2	Mean Squared Error
NB - All Features	1.4	0.783	26967.76
NB - Single Feature	4.9	0.628	17459.06

Analysis

The results of the regression tree models indicate that much of it predictive power derive from a single attribute, as the average Hits@10 score only declined from 7.0 to 6.3 when the all but the one input feature were removed. From a scientific standpoint, the findings of our study that the comment volume of a Facebook post can be predicted most accurately by the historical content volume of that post's page and the number of Facebook shares a post receives. From a practical standpoint, it appears that user controllable features such as character length and day of posting have relatively little impact on the comment volume, as we were not able to increase the accuracy of the single feature baseline by adding those input attributes. The results of the Naïve Bayes classifier indicate that due to the high correlation between post attributes, the assumption that the features are conditionally independent given the outcome is clearly false. Instead, using Naïve Bayes with one attribute closely correlated with the output can lead to acceptable results. With that said, single-feature Naïve Bayes cannot match the accuracy of the regression tree.

Conclusion

Our investigation has revealed that much of the comment volume of a post is determined by the features of that posts's Facebook page and is relatively unrelated to intrinsic features of the post. In particular, the number of posts on that page in the preceding 24 hours and the number of post shares largely predicts the amount of comments a post will receive. Among features that can be controlled by the user, the character length of a post and the day of posting are the most predictive, but their relative importance is small when compared to the page features. With that said, future work could be performed to examine the effect of promoting Facebook posts to see if such actions lead to greater comment volume. Such an approach would help determine if Facebook posts promotions are actually effective in increasing the exposure of a post.