Predicting Active Engagement Activities on Digital Blogs

Charles Schumacher

7/15/2018

Abstract

Digital media space is increasingly transacted in an auction-style format called programmatic media buying (or "RTB" short for real-time bidding) where marketers bid on the users they want to show ads in real-time based on their previous online behaviors and the context of the webpage that they are on at a given moment. This subset of media buying methods is rapidly becoming the dominant way to purchase advertising space online; over $33 billion was spent in these channels worldwide during 2017 and it is conservatively projected to hit $68 billion in spend during 2018.

For good reasons, strict privacy laws are in place that limit the types of data that can be collected, stored, and utilized for these types of marketing purposes. The signals that remain to describe an anonymous user's online interests and browsing behaviors are scarce, so innovative uses of machine learning techniques are crucial to the technology behind the platforms that enable the buying power. One particular opportunity to gleam additional insights into the vast sea of content on the internet and how users engage with it is to crawl the raw HTML documents. The types of information that will be available through these crawls vary depending on the publisher and type of website, and domains that feature user-generated content (typically abbreviated as "UGC") such as blog tend to have more available to be parsed from the HTML. Since UGC makes up a non-trivial portion of the internet, often features advertising space, and this sector of the industry continue to grow at a rapid rate, focusing additional efforts on predictive ability in this type of content is worthy of attention because of the potential to increase profit margins and a growing scale.

In this paper, the ability to predict the number of comments that will appear on a given blog post within the next 24 hours using multiple regression and Lasso regression techniques are evaluated based on historical data collected from a series of web crawls. They performance of each of these methods are comparable when trying to predict the number of comments that will appear on a blog in the next 24 hours, with each predicting certain areas of the distribution better or worse when trained on the raw target variable as well as a log transformation of the target variable. Despite less than optimal results, the overall performance is good enough to suggest that further exploration is warranted since there are several alternate algorithms and transformations not tested here which may capture this type of behaviors more effectively.

Introduction

The core idea behind this advent of this technology is efficient supply and demand, "paying the right price to show the right user the right ad on the right content at the right time". The ability to predict when a particular piece of content is going to gain attention, or go "viral" to some degree, allows for increased efficiency in the purchasing of ad spaces on the page housing the content. Knowing the likelihood of a spike in new users visiting the page, or at least sustainable levels of active engagement, can allow agencies buying programmatic media to purchase high efficiency ads: units displayed on pages that users actively engage with for greater periods of time and otherwise would be less valuable media space.

To build a predictive model around blog content, information needs to be scraped and processed from a massive number of those webpages. In order to gleam a potentially meaningful amount of information, the body of text that comprises the blog needs to be parsed along with the timestamp of when it was published, the timestamps of any comments on the post, the number of traceback links, and the timestamp of when the page crawl was completed. Using this small amount of raw information, additional features can be generated that can improve the predictive accuracy of the models.

Fortunately, a large dataset that contains this kind of information is publicly available at the UCI Machine Learning Repository, “BlogFeedback” (Buza), and is utilized for this analysis. The dataset contains the target variables, number of comments in subsequent 24 hours from web crawl, along with 280 predictor variables generated from the HTML of the page. 66 of these are continuous numerical variables which describe the total number of comments and traceback links in 48-24hr and 24-0hr leading up to the crawl, along with five number summaries for each feature generated related to the comment totals. 14 of these are binary dummy variables which encode the day of week the blog was originally published as well as when it was crawled. The remaining 200 variables are binary variables which indicates the presence or absence of a discriminatory word in the blog text (the exact words are not included). The training dataset included just over 52,000 observations, and 60 separate training sets were evaluated which ranged in number of observations from 87-300+.

Methodology

Since the ultimate goal for this type of predictive model is real-time utilization at high-scale, multiple regression was chosen for evaluation due to its fast training time; Lasso regression was chosen for its relatively fast training time and tendency for parsimony with predictor variables.

Exploratory analysis was conducted on the dataset in advance of model building to identify the need for any additional cleaning, and potentially obvious opportunities to eliminate predicator variables. Since this was a published data set, it was very clean and did not require handling of missing or broken values. However, the numerical predicator variables were very highly correlated since they were generated from a small group of original features, and a subset of the binary variables (the 14 day of week encodings) showed minimal correlation with the target variable (Figure 1) and did not improve model statistics in initial evaluation. The

multicollinear variables were treated in iterative multiple regression model tuning, and the day of week variables were dropped from the two of the three final versions of training dataset used.

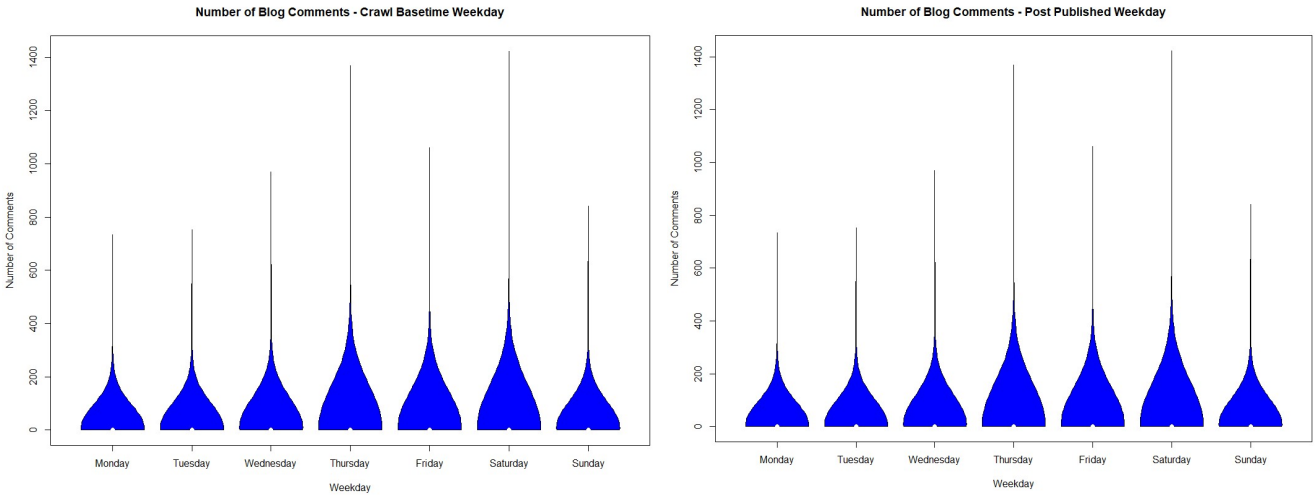

Figure 1: Violin plots of total number of comments on a blog post based on the basetime of the crawl and the day the blog was originally published. The majority of data points are very low, with a very small percentage of high comment posts

The distribution of the target variable is highly skewed with the majority of points occuring at 0 and very low values. Due to the nature of this kind of distribution (which may follows a beta distribution or combination of other distributions (Figures 3 and 4)), the changes to beta coefficients that are caused by large values in the end of the tail contribute to less effective predictions for the numerous small comment values at the lower end of the distribution. Additionally, outliers weren’t easily evaluated in the usual way of removing samples outside a

multiple of the IQR since zero values assume nearly the entire three lower-quartles. For this reason, models were evaluated with observations removed that had a target variable above an experimental threshold to reduce the impact of any potential outliers. Setting a comment threshold to less than 25 preserved 95% of the original data points (Figure 2), which was tested since removing ~5% of data points as outliers is generally acceptable. However, a target variable threshold of less than 750 (this preserved ~99% of the observations) produced more results with more optimal predictive power in initial tests.

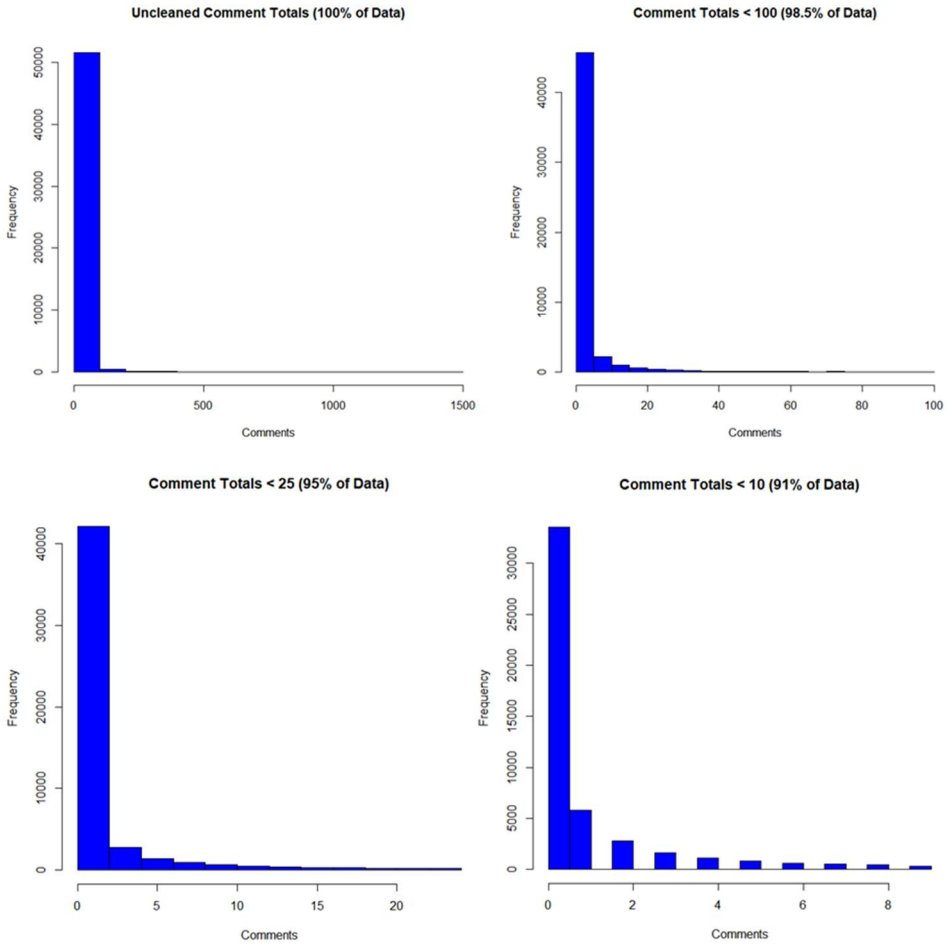

Figure 2: Histograms of the target distribution at various experimental thresholds for an upperbound

Using the comment threshold of 750, three versions of the cleaned testing dataset were produced and used to train the models: 1. All 280 feature variables, 2. Removing the 200 bag of words features and 14 day of week features, and 3. Removing the day of week features and features which were constant at 0 (four of the "minimum" columns describing collected numerical features). For each of these data sets at each comment threshold, multiple regression with two subsequent steps of tuning based on predictor significance was performed along with Lasso regression. Each of these models produced were also trained on the raw target variables as well as a natural log transformation of the variable in the form of :

$$\beta_1 * X_1 + \beta_2 * X_2 + \ldots + \beta_i * X_i = log(target + 1)$$

The purpose of the unit shift in this formula being to account for the target values of 0 and avoid infinity values.

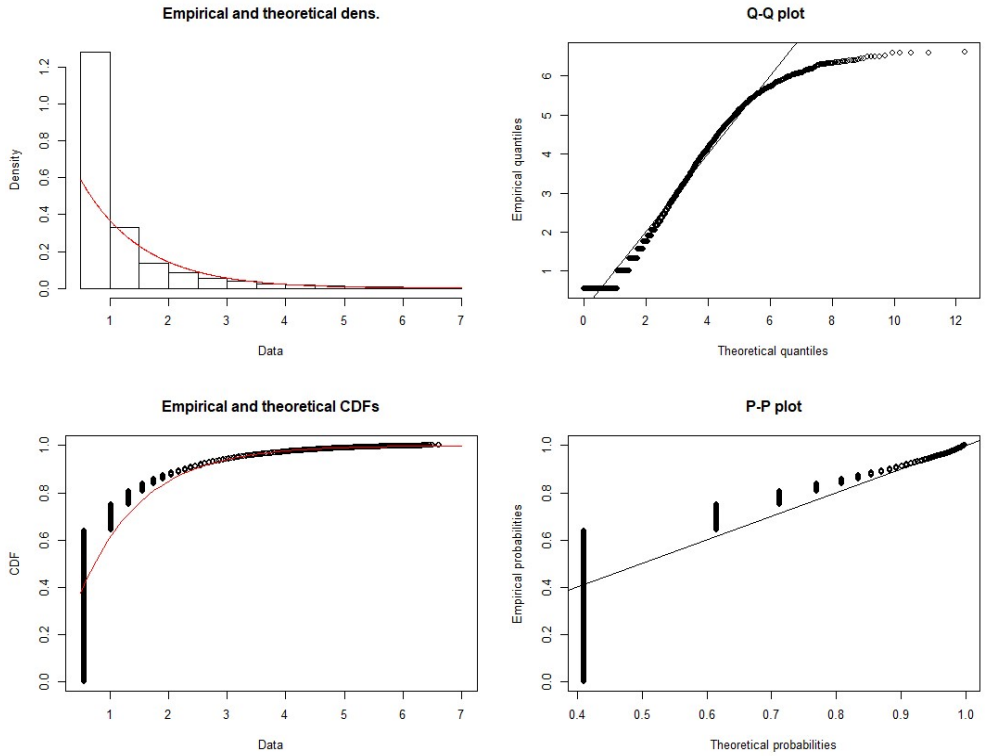

Figure 4: Experimental fit of log(target+1) distribution to the exponential distribution. Considerations for future work

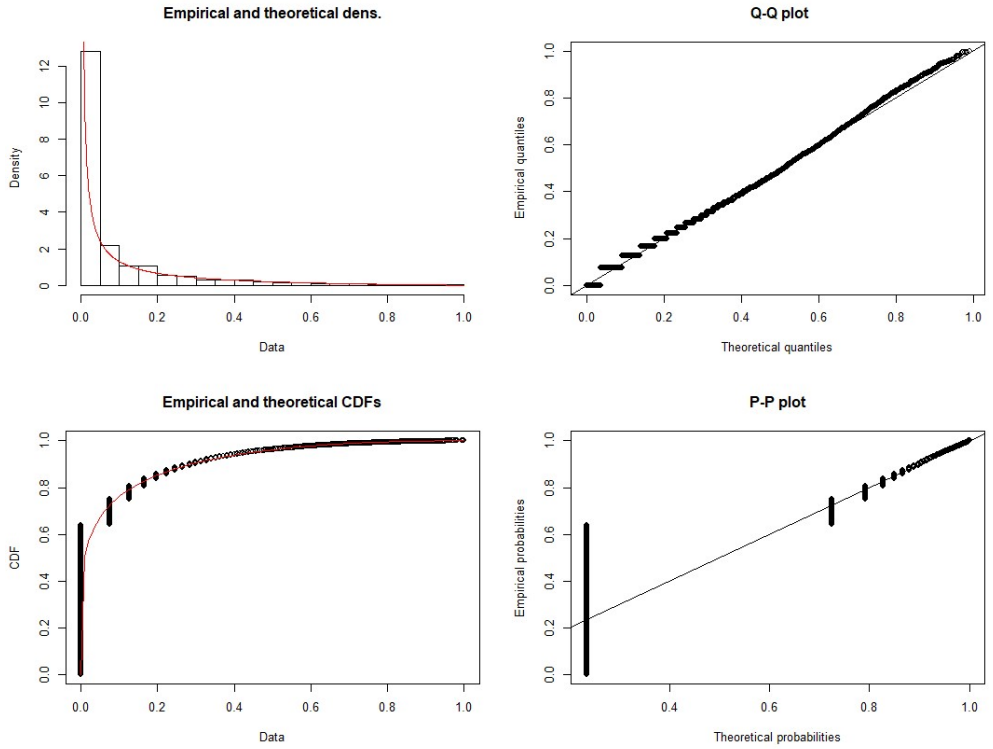

Figure 3: Experimental fit of log(target+1) to beta distribution. This is a potentially good match and main consideration for future work

Results

The primary metrics of consideration for evaluation of predictive accuracy for the multiple regression and Lasso regression models is root mean square error ("RMSE"). A table displaying identifying information on the models and the five number summary of RMSE results from evaluating 60 separate test datasets can be found in Figure 5, in ascending order based on median RMSE. The models which were trained on a log-transform of the target variable had their predicted values reverse-transformed before computing the RMSE on the 60 validation sets.

Model Name	Algorithm	Predictor Variables	Target Variable	Min RMSE	1st Q. RMSE	Median RMSE	Mean RMSE	3rd Q. RMSE	Max RMSE	Num Model Vars
LRfit1.1	Lasso Regression	Full Predictors	Y	2.999333428	14.3920959	20.44281774	22.38189071	28.55528214	51.97402991	6
LRfit3.1	Lasso Regression	Minus Day of Week Vars	Y	2.999333428	14.3920959	20.44281774	22.38189071	28.55528214	51.97402991	6
LRfit2.1	Lasso Regression	Minus Bag of Words and Day of Week Vars	Y	3.085721752	14.71706619	20.46264245	22.47306132	28.89958117	52.13679643	5
LRfit2	Lasso Regression	Minus Bag of Words and Day of Week Vars	Y	3.979550343	14.43675701	20.61661226	22.36522342	28.42085383	51.42447077	16
LRfit3	Lasso Regression	Minus Day of Week Vars	Y	3.898691595	14.47417564	20.65203998	22.37574014	28.39134791	51.45482187	36
LRfit1	Lasso Regression	Full Predictors	Y	4.197888361	14.48656338	20.72490467	22.38214991	28.41012037	51.51027987	61
MRfit2.1	Multiple Regression	Significant Predictors From MRfit2	Y	5.267885889	16.07879762	21.04606399	23.06560536	27.99560041	51.96981147	42
MRfit2.2	Multiple Regression	Significant Predictors From MRfit2.1	Y	5.290358991	16.04134513	21.06694723	23.10238019	28.21101683	52.01569853	29
MRfit1	Multiple Regression	Minus Bag of Words and Day of Week Vars	Y	4.788002987	15.16882891	21.12355459	22.5853827	28.61352574	51.30442204	66
MRfit1.1	Multiple Regression	Significant Predictors From MRfit1	Y	5.292498324	16.01637393	21.15853381	23.1027079	28.22309123	52.040886	17
MRfit1.2	Multiple Regression	Significant Predictors From MRfit1.1	Y	5.29412392	16.00891805	21.17955731	23.11974287	28.216566	52.03374432	16
MRfit0	Multiple Regression	Full Predictors	Y	5.455949491	15.18738195	21.32696426	22.72262254	28.46059591	51.44646358	280
MRfit2	Multiple Regression	Minus Day of Week Vars	Y	5.15649177	15.39259429	21.34043917	22.76171754	28.61525547	51.4139805	262
LRfit5.1	Lasso Regression	Minus Bag of Words and Day of Week Vars	log(Y+1)	2.208563406	14.94803206	22.69175349	50.25380423	31.52592952	1000.802405	36
LRfit6.1	Lasso Regression	Minus Day of Week Vars	log(Y+1)	2.183255561	14.68992813	22.72934239	47.53968882	30.94690321	940.5087574	200
LRfit4.1	Lasso Regression	Full Predictors	log(Y+1)	2.160598576	14.68296181	22.87052073	46.93555514	31.54975359	906.8361558	208
MRfit5.2	Multiple Regression	Significant Predictors From MRfit5.1	log(Y+1)	2.074405746	14.75075742	23.20874239	42.65454072	33.00566937	692.8091504	90
MRfit5.1	Multiple Regression	Significant Predictors From MRfit5	log(Y+1)	2.074597534	14.71071416	23.25104537	42.82168203	32.92776189	704.7531598	95
MRfit0.1	Multiple Regression	Full Predictors	log(Y+1)	2.07973473	14.93519399	23.58129181	43.33160763	32.57567616	728.8378205	280
MRfit5	Multiple Regression	Full Predictors	log(Y+1)	2.07973473	14.93519399	23.58129181	43.33160763	32.57567616	728.8378205	280
LRfit4	Lasso Regression	Full Predictors	log(Y+1)	2.12638935	15.16131494	23.98372144	46.19149293	33.77875452	863.86232	255
LRfit5	Lasso Regression	Minus Bag of Words and Day of Week Vars	log(Y+1)	2.174900805	15.49769989	23.99845129	49.73019972	35.0911294	967.2139682	48
LRfit6	Lasso Regression	Minus Day of Week Vars	log(Y+1)	2.14862578	15.17879775	24.04675936	46.76791384	33.63059235	903.2223753	242
MRfit4.1	Multiple Regression	Significant Predictors From MRfit4	log(Y+1)	2.182793566	16.74658395	24.24865623	64.70977249	36.44564205	1729.557279	92
MRfit4.2	Multiple Regression	Significant Predictors From MRfit4.1	log(Y+1)	2.199099349	16.75766055	24.27000312	64.81748547	36.64786186	1730.630147	87
MRfit4	Multiple Regression	Minus Day of Week Vars	log(Y+1)	2.192555435	17.1945802	24.29108665	65.63486608	36.96584362	1765.72766	262
MRfit3	Multiple Regression	Minus Bag of Words and Day of Week Vars	log(Y+1)	2.218162847	17.78480946	24.55030262	76.5547003	39.0242572	2133.699989	66
MRfit3.1	Multiple Regression	Significant Predictors From MRfit3	log(Y+1)	2.218785145	18.50354572	24.63208399	76.53618725	39.19940442	2095.424769	45
MRfit3.2	Multiple Regression	Significant Predictors From MRfit3.1	log(Y+1)	2.219248171	18.53216203	24.65454471	76.71843434	39.20249002	2105.498926	44

Figure 5: Table of RMSE results from predictions on 60 test data sets. Sorted in ascending order based on median RMSE

While the spread of values in median RMSE is not very large, the models which were trained on the raw target variable produced a more optimal RMSE when evaluating the testing data than any of the models trained on a log transformation of the target. The most effective model was utilizing Lasso regression and turned out to be highly parsimonious with only 6 predictor variables having non-zero beta values. Notably, the models trained on a log-transformed target all produced lower minimum RMSE close to 2, while the other models had a minimum RMSE between 3 to 5.5. However, the models with a log-transform fared poorly on the other end of the spectrum with maximum RMSE in the testing corpus being between 600-2100, and the non-transformed models had a maximum RMSE around 52.

Model Name	Algorithm	X Set	Y Set	Min RSE	1st Q. RSE	Median RSE	Mean RSE	3rd Q. RSE	Max RSE	Num Model Vars
LRfit1.1	Lasso Regression	Full Predictors	Y	2.22938698	11.77291798	18.87031856	21.1035868	26.42031403	75.92954574	6
LRfit1	Lasso Regression	Full Predictors	Y	3.120265846	11.82718135	19.5633147	21.18650826	26.51003638	73.63963534	61
MRfit0	Multiple Regression	Full Predictors	Y	4.066624648	12.31004839	19.4355952	21.5590683	26.83614064	73.54383586	280
LRfit3.1	Lasso Regression	Minus Day of Week Vars	Y	2.34925925	12.44305119	20.30271206	22.72111225	27.81171798	88.54830572	6
LRfit3	Lasso Regression	Minus Day of Week Vars	Y	3.053690933	12.38587094	21.08469765	22.81547616	29.38240584	86.00624268	36
MRfit2	Multiple Regression	Minus Day of Week Vars	Y	4.051322553	13.0346399	20.60005287	23.26864853	28.52813192	86.00055636	262
MRfit1.2	Multiple Regression	Significant Predictors From MRfit1.1	Y	5.776362763	17.14915379	22.90492722	24.82117811	29.79691754	55.72201799	16
MRfit1.1	Multiple Regression	Significant Predictors From MRfit1	Y	5.809271621	17.2895523	23.004333	24.92257191	29.93017395	55.98707911	17
MRfit2.2	Multiple Regression	Significant Predictors From MRfit2.1	Y	6.278501016	18.41872447	24.96281431	26.52546996	31.52118576	59.35452381	29
MRfit2.1	Multiple Regression	Significant Predictors From MRfit2	Y	6.917073048	20.16369784	27.34828064	28.66822833	33.94485259	63.77744276	42
MRfit1	Multiple Regression	Minus Bag of Words and Day of Week Vars	Y	8.211357379	23.26849858	33.29601979	34.5165905	41.98443866	99.23406098	66
LRfit2	Lasso Regression	Minus Bag of Words and Day of Week Vars	Y	6.927507955	23.35859677	31.77362821	34.60106318	42.20744743	102.826515	16
LRfit2.1	Lasso Regression	Minus Bag of Words and Day of Week Vars	Y	5.371552095	23.39270143	33.03421028	34.89923332	43.913037	103.3358706	5
MRfit0.1	Multiple Regression	Full Predictors	log(Y+1)	1.550142744	12.69736859	21.57623117	46.76799993	29.47433868	1058.976802	280
MRfit5	Multiple Regression	Full Predictors	log(Y+1)	1.550142744	12.69736859	21.57623117	46.76799993	29.47433868	1058.976802	280
LRfit4	Lasso Regression	Full Predictors	log(Y+1)	1.580532756	12.77593918	21.36208382	50.3824847	29.71964665	1248.247301	255
LRfit4.1	Lasso Regression	Full Predictors	log(Y+1)	1.605960274	12.65986975	20.96898634	51.4739768	28.64762214	1310.342815	208
LRfit6	Lasso Regression	Minus Day of Week Vars	log(Y+1)	1.682933595	13.69739559	23.26844572	56.06035611	32.00243202	1457.169452	242
LRfit6.1	Lasso Regression	Minus Day of Week Vars	log(Y+1)	1.710057733	13.65179843	22.08476423	57.2343635	30.71153	1517.323605	200
LRfit5	Lasso Regression	Minus Bag of Words and Day of Week Vars	log(Y+1)	3.78601634	24.11590445	35.64434745	71.60634267	51.25612672	1202.117182	48
LRfit5.1	Lasso Regression	Minus Bag of Words and Day of Week Vars	log(Y+1)	3.844615408	22.95922974	34.6083639	72.18499498	50.02906903	1243.863102	36
MRfit4	Multiple Regression	Minus Day of Week Vars	log(Y+1)	1.72263424	14.88963925	23.90733768	84.10727809	34.17204854	2868.364062	262
MRfit5.2	Multiple Regression	Significant Predictors From MRfit5.1	log(Y+1)	5.172818856	32.93981575	48.64126218	89.58299246	78.61070385	954.9715106	90
MRfit3.2	Multiple Regression	Significant Predictors From MRfit3.1	log(Y+1)	2.965595111	23.93895514	32.15584245	92.29065847	48.01340426	2401.902625	44
MRfit3.1	Multiple Regression	Significant Predictors From MRfit3	log(Y+1)	2.991809279	24.05741403	32.35087198	92.5733038	48.27473893	2398.638899	45
MRfit5.1	Multiple Regression	Significant Predictors From MRfit5	log(Y+1)	5.556008425	38.12002599	51.11457116	93.36531237	82.03811062	996.6714767	95
MRfit3	Multiple Regression	Minus Bag of Words and Day of Week Vars	log(Y+1)	3.804117899	27.68479968	37.06215866	105.9017445	54.86546168	2641.187984	66

Figure 6: Table of RSE results from test set corpus, sorted in ascending order by mean RSE. This cleanly divides the untransformed and log-transformed models here

Discussion

The spread of performance as measured by RMSE for the multiple regression models built is not significantly large; the difference of max RMSE observed when comparing the model using only 6 predictor variables (LRfit1.1) and one using all 280 (MRfit0) is around 0.5, and the difference of min RMSE observed between these two models is around 2.5. The beta values in MRfit0 are larger with the top three being over 200,000 and the maximum reaching 2.94e7 ("Total number of trackback links in 24 hours before basetime crawl"). Unsurprisingly, this model utilizing 280 predictor variables suffers from significant variance inflation factors and extremely high beta coefficients beyond what makes sense. The beta values in LRfit1.1, however, are significantly lower with the max being only 1.58 ("Average number of comments on blog post in first 24 hours after being published").

It is a mildly curious result to find a model suffering with such intense multicollinearity and variance inflation produce meaningfully comparable results to the most optimal model in the set which eliminates most of those symptoms and contains reasonable beta coefficients that align with reasonable target values. Given the extreme distribution of target values with a long tail, it can be seen that the LRfit1.1 model with 6 non-zero betas generally underestimates in predictions compared to MRfit0, in particular in situations where the actual target value is high (Figure 7).

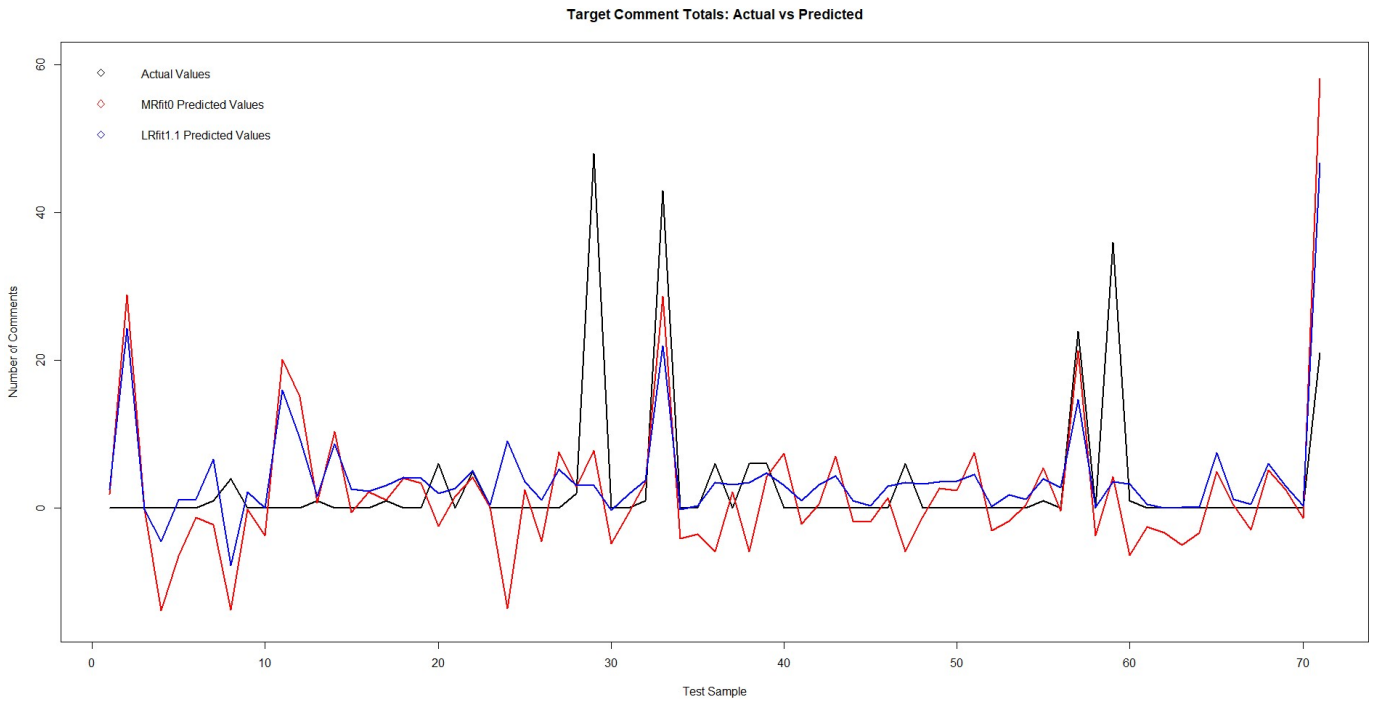

Figure 7: Actual vs predicted values for top two models by median RMSE

On the other hand, given the shape of this target distribution (Figure 2) it is not entirely surprising that the tuned models all seem to reach a comparable level of performance; a level which is less than ideal considering that 91% of the comment totals are less than 10 and the mean RMSE is over 20 for the best model. Given the distribution-fit estimations (Figure 4) it appears that there may be other distribution that are far more optimal to train the models on than a linear form of regression. This is noted as a starting place for future work.

Future Considerations

There are certain limitations of this dataset which may also be contributing to the mixed results seen when training linear models on the target: the fact that the authors didn't include frequency metrics for the word, only whether they a notably significant one was present or not, and they only included lookback times of 48 hours for comments and trackback links. UCI did not note whether the data collectors did modelling of this dataset themselves first and determined that these were the best way to structure the variables, so it may be worth testing the hypothesis that a different lookback window could provide more precise predictions, or that knowing the specific word frequencies on the page could provide more predictive power. All that is required to generate a similar dataset to what has been explored here is the collection of: raw text of the blog post for parsing important phrases, the total number of comments with timestamps, and the total number of trackback links with timestamps. All of the other metrics can be generated from these three raw features which can be easily collected from the HTML of a page.

If additional models are fit to alternate distributions such as log-exponential or beta during future iterations of this analysis, and they produce more favorable results, a time vs accuracy trade-off will need to be considered for evaluating if the regression models can be scalable to a level of utility for programmatic media, given the simultaneously high-scale and low-latency nature of the computing environment. Next steps in evaluating this would be to collect additional data independently that can be evaluated with the model to determine the intervals at which it will need to be retrained; once it is trained it may be very fast to make new predictions, but

if it needs to be trained frequently and customized to fit the needs of each advertiser and the type of content that they want to buy ads on it could be a more complex issue and ultimately favor additional tuning/transformation to a linear regression approach.

Conclusions

Linear models are typically very fast to train when compared to other machine learning algorithms. While the analysis described in this paper has not produced definitively useful models, it highlights that there is potential to utilize such regression methods in order to forecast active engagement on subsets of internet content, using only publicly available signals from the HTML of the webpage. Given the scale of inventory available for purchase programmatically on the internet (for reference, each buy-side company sends out 1-9 million bids per second) and the pace at which it evolves, a model that can predict active user engagement in the next 24 hours and which is also fast to train will have a high ceiling of potential value. In a world where billions of dollars worth of media is dynamically priced in real-time based on internet traffic metrics, knowing the domains where users are going before the companies that monetize the content realize the traffic surge will create a tremendous opportunity for efficiency in terms of profit margin and marketing efficacy on behalf of the advertiser.

Work Cited

Buza, Krisztian. "BlogFeedback Data Set." *UCI Machine Learning Repository*, 2014, archive.ics.uci.edu/ml/datasets/BlogFeedback.