BUSN 41201 BIG DATA

Final Project: Facebook News

Prepared By:

Raghav Goel

Kim Janssen

Sreeranjani Ramprakash

Chong Tao

We pledge our honor that we have not violated the Honor Code during the preparation of this assignment.

Contents

0. Executive Summary

We analysed a year's worth of Facebook news posts from Seattle ABC-affiliate KOMO-TV in an attempt to predict the performance of posts based on their content and timing, and to generate insights about what drives increased performance. We used data visualisation, principal component analysis, cross-validated linear and categorical lasso regressions, CART and random forest models and examined treatment effects. Our specific goals were to predict and understand the drivers of impressions (the number of times each post is seen), link clicks (the number of click on links within a post, typically to visit the KOMO-TV website), video views (the number of times a video is viewed) and average viewing time. We found:

- Including video in a post ensures it gets more impressions, but also leads to fewer link clicks. Our guess is that viewers watch the videos within the Facebook platform and have little reason to click through to the KOMO-TV website.
- Longer headlines increase impressions and link clicks, as does posting more frequently over a 3 hour period.
- If it bleeds it leads: no topic ensures a post gets more impressions and link clicks than crime. Adding a little bit more crime into the main text of the post adds more than 13,000 impressions and 127 link clicks to the average post.
- Sport, weather and human interest — all local news staples — also drive impressions.
- Celebrity news, social issues and whimsy are all relative losers for KOMO-TV.
- No statistically significant evidence that paying Facebook to boost posts causes more impressions or link clicks (although we only had 24 paid posts to work with).
- Longer videos not only result in longer viewing time, but also more views. This surprised us.
- Posting in the small hours of the morning doesn't do much.
- The best 13.5% of posts account for half of all impressions.
- The worst 20% of posts account for just 4% of impressions.
- The worst 50% of posts generate just 3% of link clicks.
- But viral hits, duds and poor referrers are hard to predict with the models we built.

For each of the specific modeling challenges we set ourselves, the best model we built was:

- **Modeling impressions:** a 500-tree Random Forests model to predict log impressions, with out of boot R^2 of 0.28 and an in-sample R^2 of 0.87.
- **Modeling link clicks:** a 500-tree Random Forests model to predict log link clicks, with out of boot R^2 of 0.69 and an in-sample R^2 of 0.92.[1]
- **Modeling impression 'hits':** a 500-tree Random Forests model, tuned to minimize false positives predicts 71% of all hits before they happen but has a false positive rate of 0.79.
- **Modeling impression 'duds':** a 500-tree Random Forests model, tuned to minimize the false positive rate to under 0.07 predicts 5% of all duds before they happen.

[1] This model only marginally outperformed a 100-fold CV Lasso.

- **Modeling link click ‘poor referrers’**: a 100-fold cross-validated binomial lasso regression model, with 42 X variables, tuned to minimize false positives to under 0.03, predicts 21% of all poor referrers before they happen.
- **Modeling average video view time:** a 100-fold cross validated lasso on average view time, which selected 3 X variables and rejected 80 variables. It had an out-of-sample R^2 of 0.57.
- **Modeling video views:** a 500-tree Random Forests model to predict log views, with out of boot R^2 of 0.43 and an in-sample R^2 of 0.85.

1. Introduction

Legacy news media is increasingly dependant on social media in general, and Facebook in particular, for its audience. A 2016 Pew Research Center study found that 62% of Americans get news on social media, and that 44% of Americans get their news on Facebook.[2] Legacy broadcast and print news organizations that until recently enjoyed monopoly or oligopoly profits from their outsized power in local advertising markets must now fight in the fiercely competitive online space for eyeballs. No longer are they simply competing with local TV stations and newspapers. Now they are up against viral videos, meme factories, Hollywood, and "homemade" content from your work mate's views on politics to your aunt's vacation photos — in short, everyone and everything is competing for the attention of viewers legacy news organizations could formerly take for granted. Increasingly, legacy media needs to know: what drives viewership and engagement for news on Facebook?

[2] "News Use Across Social Media Platforms"
http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

2. Our Dataset

Facebook may have stolen away many of the viewers that local TV news could once rely upon, but does at least offer station managers far more data about users, what they like and what they don't like than traditional broadcast ever did. Facebook's "Insights" function allows commercial Facebook users to download data about the reach, impact, and reaction to every post made on their Facebook page.

For this project, we are using the data from a year's worth of Facebook posts from the Facebook news page of the Seattle ABC affiliate, KOMO-TV Channel 4.

Our goal is to build a model or series of regression models that predict the performance of a post on the KOMO-TV Facebook page. These models should also provide insight about what makes a post perform well and gives it a chance to become a viral hit, and, on the flipside, provide insights about the kind of posts likely to sink without trace.

The dataset comprises 8,639 records, each referring to a single post on the KOMO-TV Facebook page.[3] Within each record there are 92 variables, of which 8 describe the content of the post (The words used in the post, any links, photos or videos embedded in the post, the time and date of the post, and so on) and 84 describe the performance of the post (The number of impressions, the number of "likes," "shares," how long users spent watching any video associated with the post, and so on).

For simplicity's sake, we decided to treat the performance variables strictly as "y" or target values in our analysis.[4] To keep the scope of the project manageable (and since many of the 84 performance metrics are similar to one another, or can be constructed from a smaller set, if necessary) we narrowed our initial focus to what we believe to be the 8 most important performance variables for the majority of posts:

Impressions: the number of times users see the post, either in their feed or by going directly to the KOMO-TV page. An important metric for advertising.

Paid impressions: the number of times a user who would not otherwise have seen the post, sees it, because KOMO-TV has paid Facebook to boost the post.

Link clicks: the number of times users clicked on a link in a post. These links invariably take the user to the KOMO-TV website, where KOMO can earn advertising dollars.

Engaged users: the number of times users clicked anywhere on the post. This indicates a higher level of interest than an "impression," but includes users who do not comment, click on a link, like or share the post, which would presumably require a higher level of engagement.

Likes: the number of users who click to "like" the post.[5] The next step up in engagement.

---

[3] The data was originally encoded in "json" format but we converted it in Python into a "csv" format for analysis in R. An unedited sample record is in the Appendix.

[4] It is possible to imagine a model that, for example, uses the performance of the preceding 10 posts to predict the performance of the 11th post, but since we were more interested in developing insights about content, we forwent that approach, which nonetheless remains a potential avenue of future study. We'd need a more complete dataset, however, as the data we have now only has "lifetime" values for performance variables.

[5] Our data comes from 2014-2015, before Facebook allowed users to respond with emojis beyond "like."

Comments: the number of comments on the post. Represents a higher level of engagement, arguably.
Shares: the number of times the post was shared. Represents high engagement and virality.
Negative feedback: the number of users who choose to either hide that specific post, or to hide all KOMO-TV posts, as a result of that post.[6]

For the subset of 831 posts which contain video, we have an extra set of 24 performance variables. Again, for simplicity's sake we reduced this to the following 2 "y" values:

Average viewing time: the average number of milliseconds users spent watching the video.
Views: the number of viewers who clicked to view the video.

Having substantially narrowed the number of target variables, we wanted to expand the number of potential explanatory variables under consideration. By splitting some of the existing 8 variables that describe the content of each post, and by running a few simple algorithms, we ended up with in excess of 100 potential explanatory variables[7] that we had reason to believe may affect post performance. The basic potential explanatory variables were:

Message length: The number of words in the message, the main text of the post.
Message omegas: The text of the message's weighting on each of ten topics or 'principal components,' each of which represents a genre of story.[8]
Name length: The number of words in the name, the headline of the post.
Name omegas: The text of the name's weighting on each of ten topics or 'principal components,' each of which represents a genre of headline.
Description length: The number of words in the description, the sub-headline of the the post.
Description omegas: The text of the description's weighting on each of five topics or 'principal components,' each of which represents a genre of sub-headline.
Month: The month in which the post was created, broken into dummies.
Day: The weekday on which the post was created, broken out into dummies.
Hour: The hour (in Pacific time) when the post was created, broken out into dummies.
Hours to last post: The number of hours since the previous post
Posts in last 2 hours: The number of posts in the two hours immediately preceding the post.
Posts in last 3 hours: The number of posts in the three hours immediately preceding the post.

---

[6] While "hide all" is obviously much worse for KOMO-TV, the number of users selecting "hide all" is vanishingly small, so we do not feel much is lost by combining it with the lesser "hide" variable.
[7] Including dummies, that is.
[8] These and the other omegas are described in greater detail later.

Posts in last 6 hours: The number of posts in the six hours immediately preceding the post.
Posts in last 24 hours: The number of posts in the 24 hours immediately preceding the post.
Path 1: the first word after the first "/" in the "link" field. Typically represents the type of story (news/sports/weather etc.). Converted to dummies.
Path 2: the first word after the second "/" in the "link" field. Typically represents the sub-type of story (national/local/lifestyle etc.). Converted to dummies.
Post type: the type of post, which can take four values: **Text,** which contains only text and no link[9]**; Photo,** a Timeline photo or Facebook photo album; **Video**, which contains a video; and **Link**, a post with both text and a link to an external source (typically KOMO-TV's website).
Paid: binary field indicating if KOMO-TV paid Facebook to boost the post.

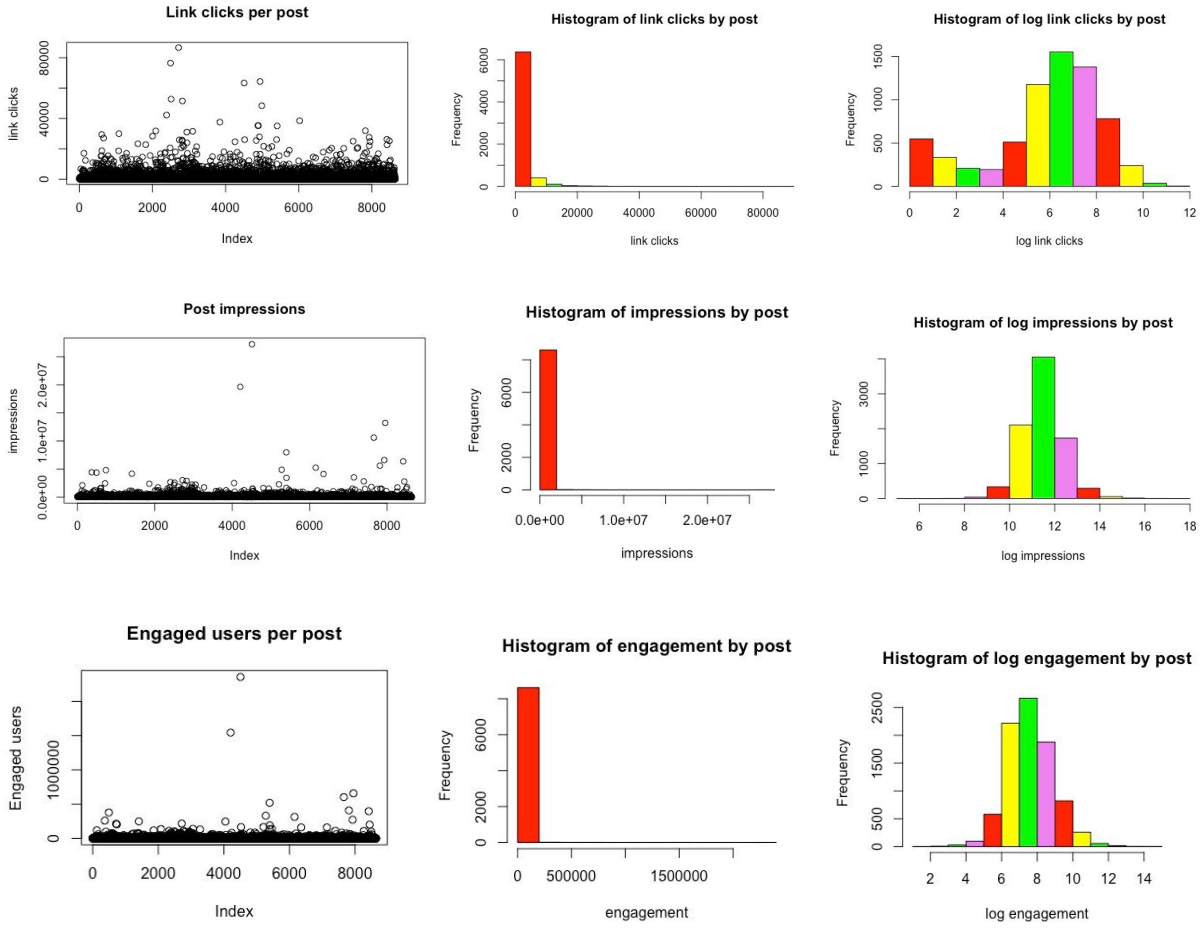
For video posts, we had an additional explanatory variable:

Video length: the length in milliseconds of the video.

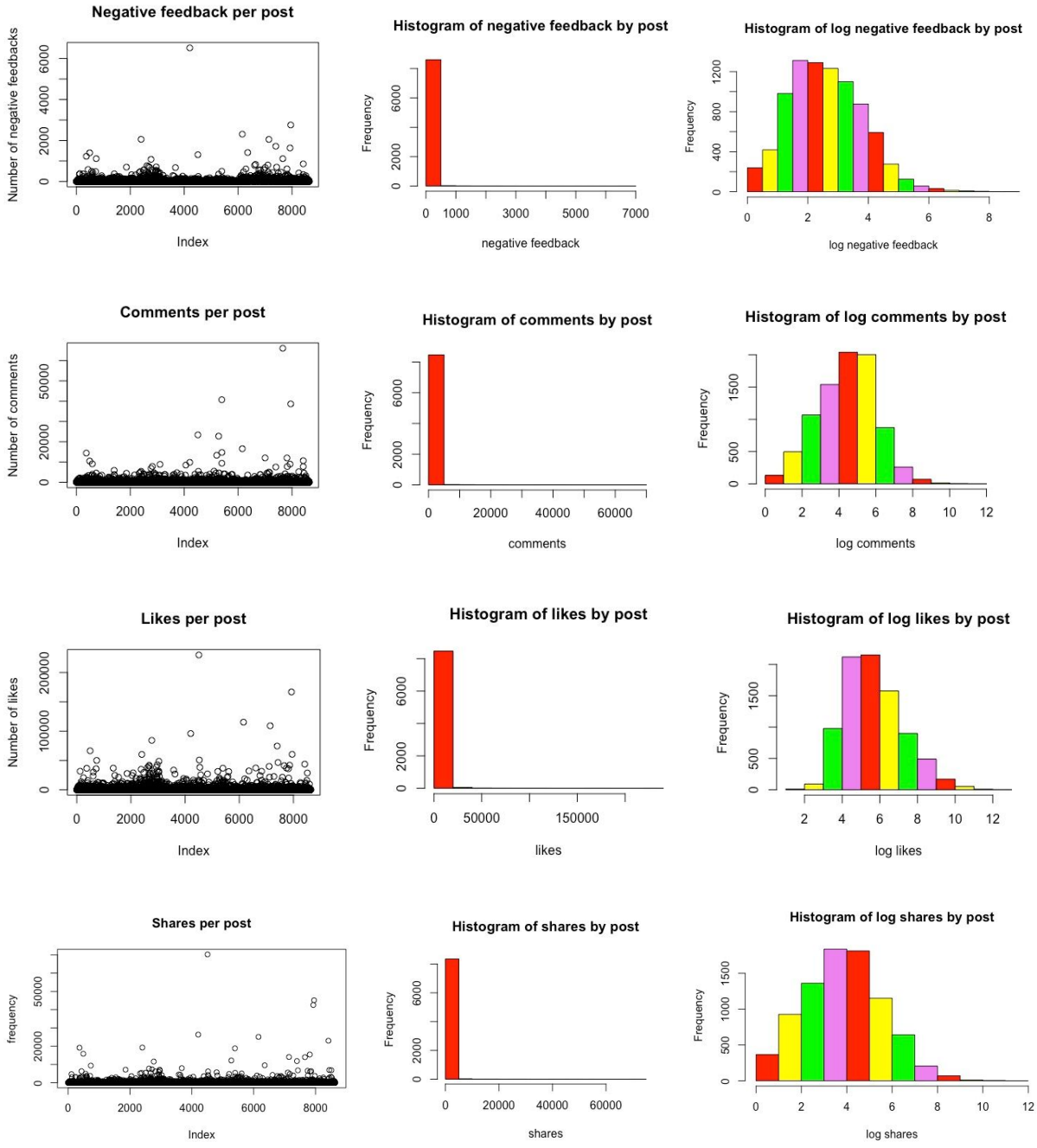
[9] In some cases, 'text' posts appear to have had a link appended to them after they were initially posted, though the "link" does not appear in the "link" field in the original facebook insights .json file.

3. Initial Analysis of Potential Y Variables

To make our project more manageable and to narrow the scope to the most relevant performance metrics, we began with an analysis of the distributions and correlations of the subset of seven potential 'y' variables we were considering. Each of these variables (link clicks, impressions, engaged users, negative feedback, comments, likes and shares) exhibited a strong left skew, with a handful of extreme outliers. Transforming each to a log scale resulted in near-Gaussian distributions:[10]

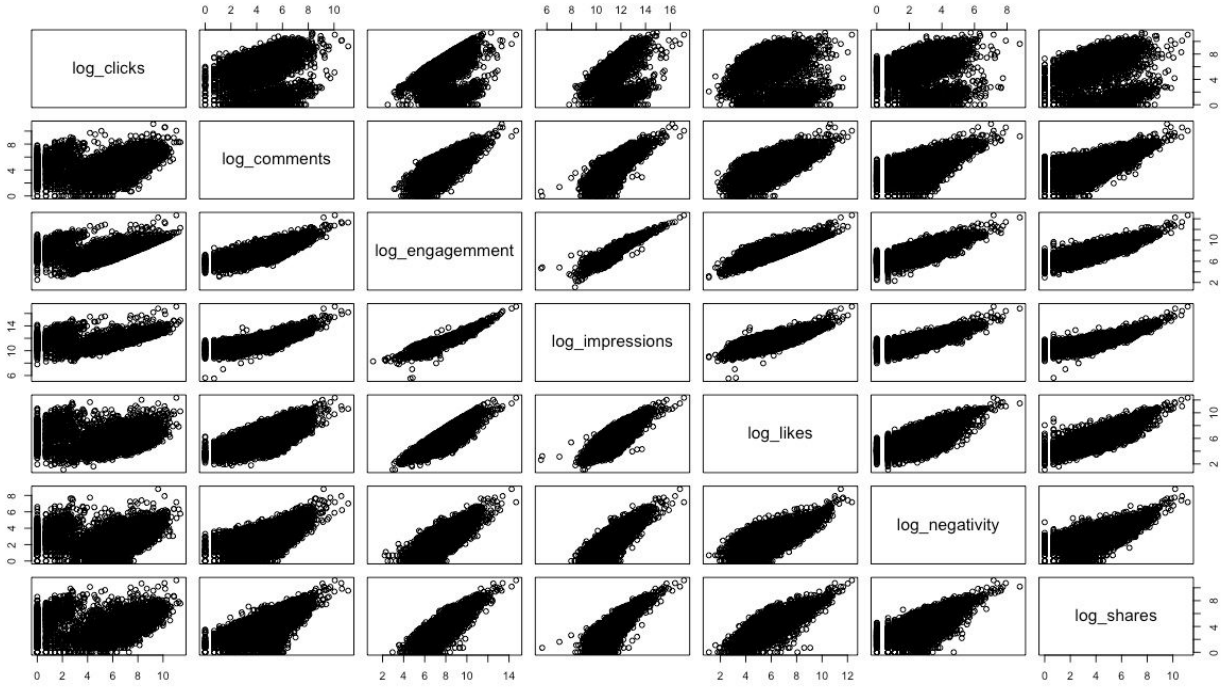


[10] In each of the following the plots, the left hand plot shows the untransformed distribution of the variable by index, the middle show a histogram of the untransformed variable, and the right hand plot shows a histogram of the log transformed variable.



As can be seen, the extreme outliers, representing ‘viral hits’ are, by definition, small in number, and are therefore likely to be hard to predict. For example, the mean number of link clicks for a linked post is 1,721, the median is 576 and there are a handful of extreme outliers, with the maximum number of link clicks per post at 86,640. The standard deviation is high at 3,762.

Given the similar distributions of the transformed performance variables[11] and our assumptions about how virality works (that is, that, likes, clicks, comments and shares work on Facebook's internal algorithms to increase impressions, which in turn increase likes, clicks, comments and shares, and so on) we suspected that they were highly correlated with one another. A matrix of correlation plots showed that this was, indeed the case:



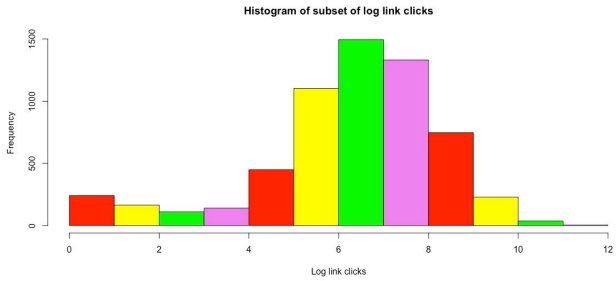
With the possible exception of log clicks (top left), log impressions (at center) seemed highly correlated with all of the other performance variables. While deciding which of the performance variables is most important is a chicken-and-egg type problem, given our assumptions about the recursive nature of virality, it makes sense to us that the number of eyeballs on a post is a key determinant of how often it is liked, shared, etc. Since "impressions" is also a key advertising industry metric, we decided to make log impressions our main 'y' variable.

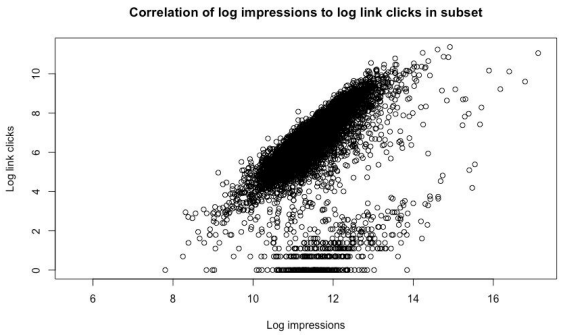
However, we were also interested in log clicks, for a couple of reasons: 1) it did not appear to be so strongly correlated with log impressions; and 2) it represents users who, by clicking on a link in a Facebook post, are typically being taken to KOMO's own website, which has advertising revenue implications for KOMO.

Even after being transformed, log link clicks remained bimodal, with a significant number of posts showing either zero or a very small number of users who clicked through to KOMO's website. This is partly explained by the fact that "text" and "photo" type posts had either no links in them, or appear to have had links appended to them after they were initially created. Removing "text" and "photo" type posts from the datasets left 6,606 rows and reduced most of the bi-modality from the distribution (below, left). However, some remained, and the correlation

[11] Log link clicks is, however, uniquely bi-modal, as we discuss later.

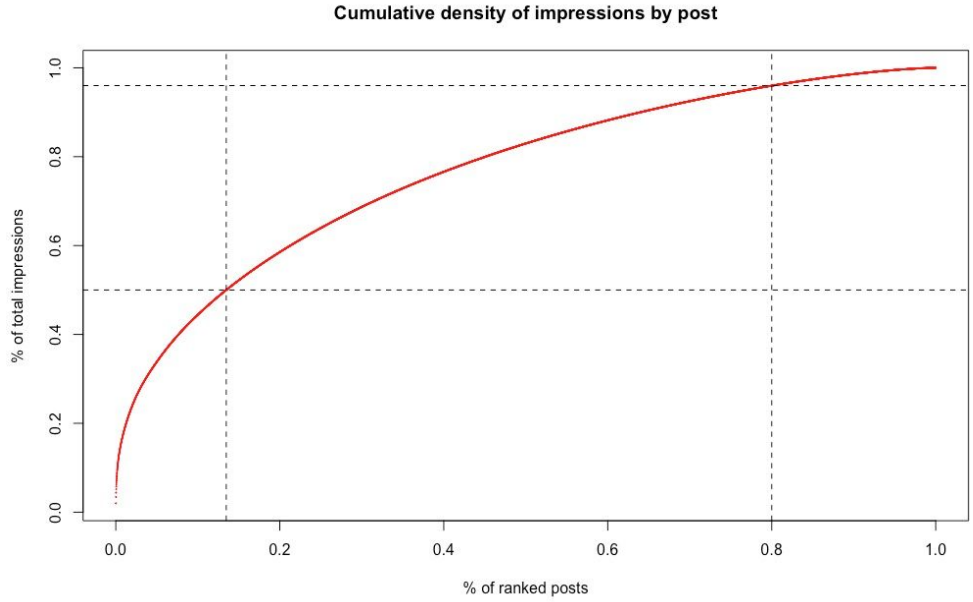
between log impressions and log link clicks for this subset of 6,606 posts seemed like it might reflect this interesting bimodality (below, right):



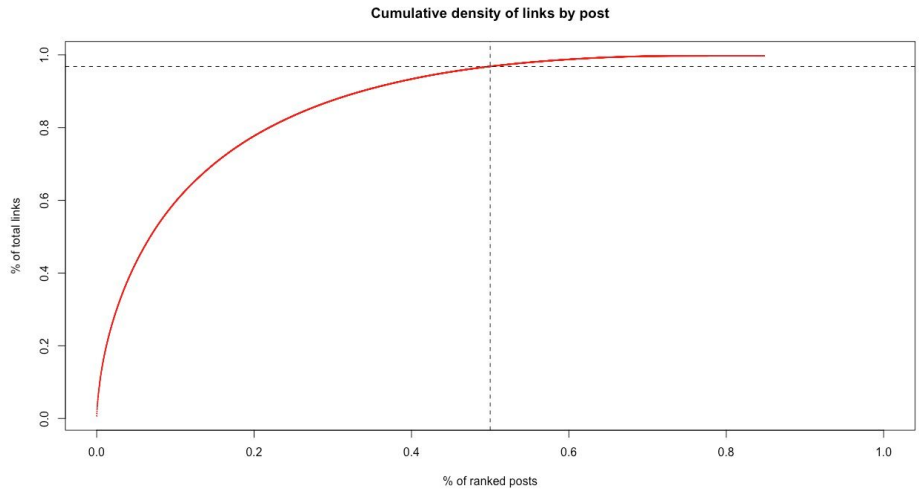


We therefore decided to retain log link clicks as a target variable in our analysis, along with log impressions. Throughout the rest of the project, when modeling log link clicks, we used the subset of 6,606 posts, whereas our models of log impressions were built with all 8,639 posts.

Finally, having selected impressions and link clicks as our targets, we plotted the cumulative density functions of each to try to understand the relative importance of big viral hits and of posts that went nowhere to KOMO. This showed us that half of all impressions come from just 13.5% of posts, and the worst 20% of posts add just 4% of impressions. We decided to call posts with more than 233,271 impressions (the top 13.5%) *hits*, and posts with less than 48,414 impressions (the bottom 20%) *duds.* The ability to predict hits and duds would potentially be of great use to KOMO-TV.



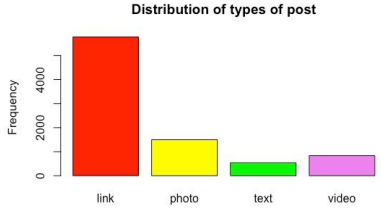
The situation is even more extreme when it comes to link clicks. Even if we look only at posts with links in them, the bottom performing 50% of posts account for only 3.2% of link clicks. We call these posts *poor referrers*. Predicting these posts might also be useful for KOMO-TV:

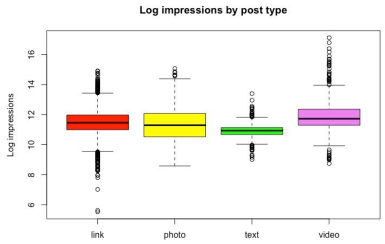

Cumulative density of links by post
% of total links
0.0
0.2
0.4
0.6
0.8
1.0
% of ranked posts

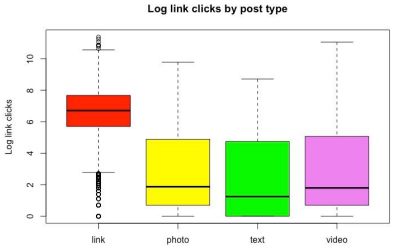
4. Initial Analysis of X variables

Now that we knew what our target variables were, we wanted to understand our potential explanatory variables a little better, both in terms of their own distribution and in a naive correlation with our target variables.[12]

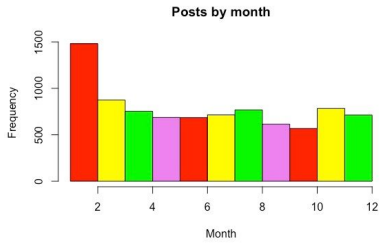
Post types: The vast majority of the posts (5,774) are "link" type posts, containing text and a link to KOMO-TV's website; there are 832 video posts, around 537 posts containing text only (and no link in the original post), and 1,496 photo posts. Video posts appear to outperform other posts when it comes to impressions, but posts without a video or photo generate a lot more traffic to KOMO-TV's website:[13]

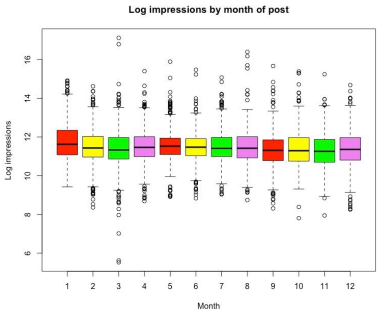


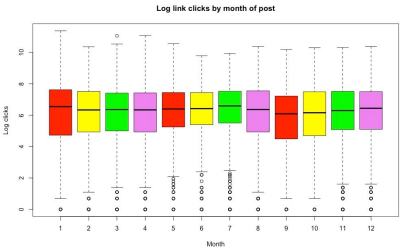




Months: KOMO-TV posted significantly more in January than it did in any other month. This is likely due to excitement around the Seattle Seahawks Super Bowl trip that year. There is surprisingly little variation from month to month in the number of impressions a post generates, nor does there seem to be much variation in link clicks per post from month to month:





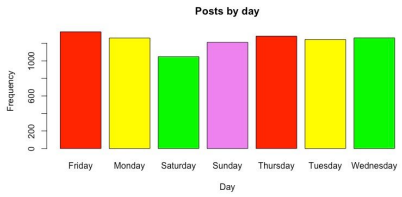


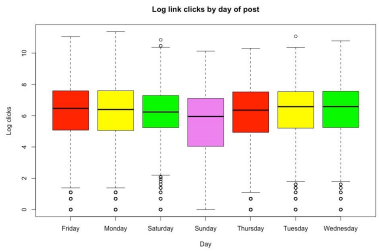
Days: Fridays and Thursdays saw the heaviest posting, and Saturday the lightest, though the variance is less than we might expect. Sunday looks like a bad days for impressions per post (and Thursday looks a little slow), though the variance here is also surprisingly small. Link clicks per post are also remarkably stable across the week:

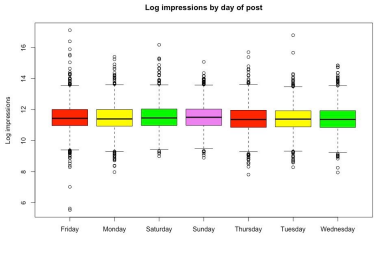
---

[12] In each of these analyses, the leftmost plot shows the distribution of the x variable, the center shows the x variable plotted against log impressions and the rightmost shows the x variable plotted against log clicks.

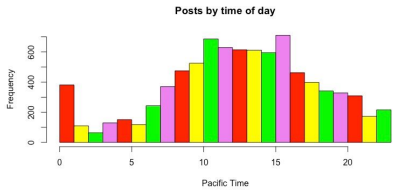
[13] The "text" field has a non-zero median because a handful of "text only" posts had links added to them after they were first posted.

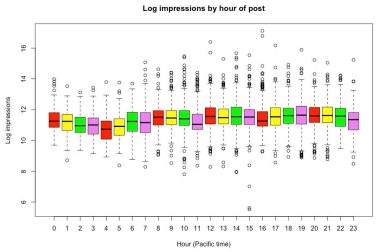


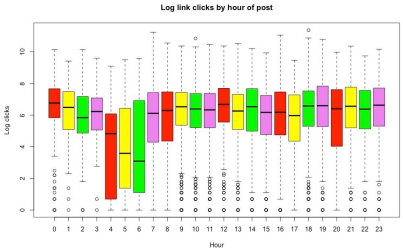




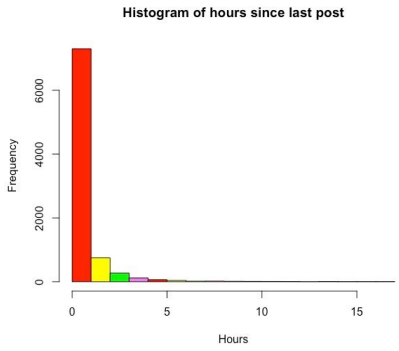
Hour: There are bimodal peaks in posting during the day, around the 10 a.m. and 3 p.m. hours. These may coincide with TV broadcasts, or reflect the production schedule more generally. The time of the day is more revealing than the month when it comes to predicting impressions. We observe declining performance from 10pm to 4am, and peaks at noon and 7pm. Posts between 4 and 7am get very few link clicks and there is some variance observable through the rest of the day:

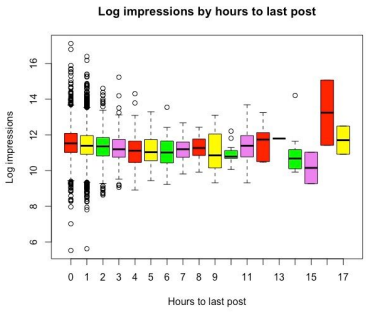


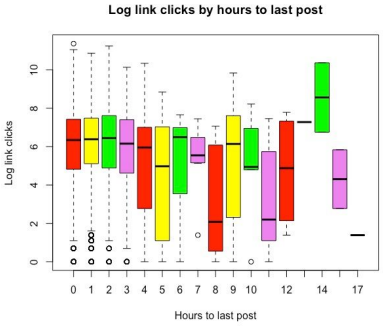




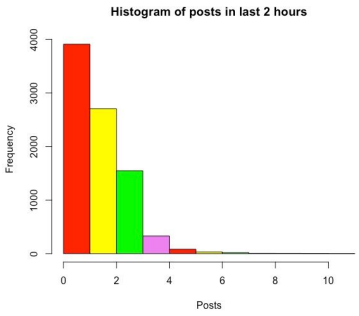
Hours to last post: Most posts are made within an hour of another post, creating a heavy left skew to the distribution. Though this may favor a log transformation, we kept it untransformed for consistency with the other time variables and for ease of interpretation. More frequent posting seems associated with higher impressions and link clicks:

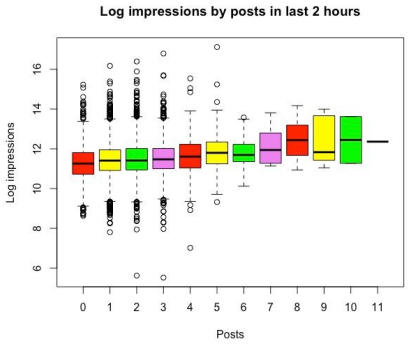


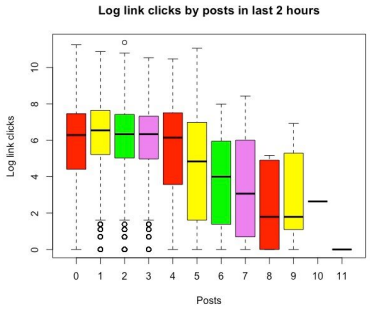




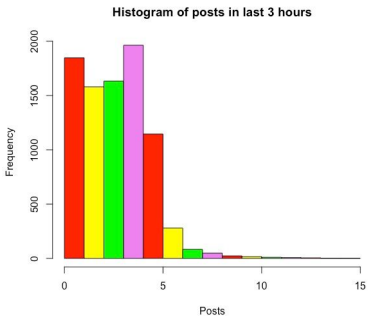
Posts in past 2 hours: Whether it's over the last 2, 3, or 6 hours, increasing the number of posts seems to increase the number of impressions and decrease the number of link clicks each post gets. This phenomenon may be explained by Facebook's surfacing algorithm favoring frequent posting, and KOMO's audience's finite time for reading stories. More frequent posting would therefore result in more eyeballs on KOMO posts, which would be competing with each other, resulting in cannibalization and fewer link clicks per post.

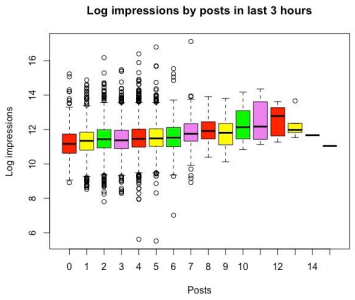


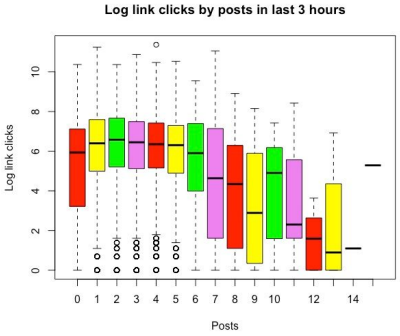




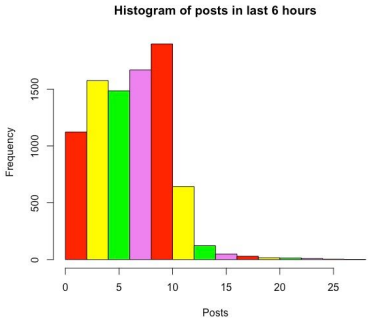
Posts in past 3 hours:

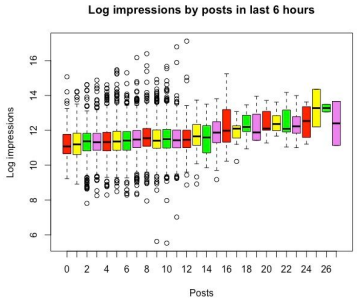


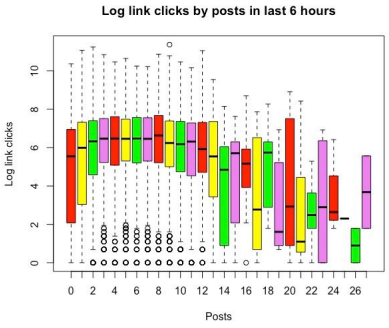




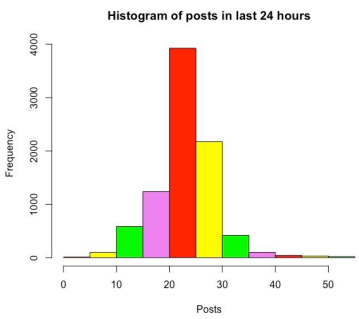
Posts in past 6 hours:

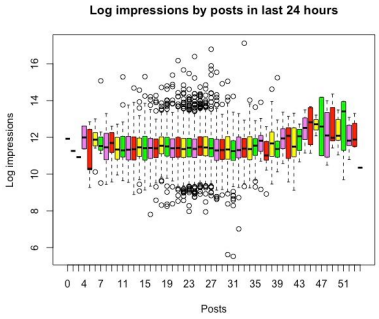


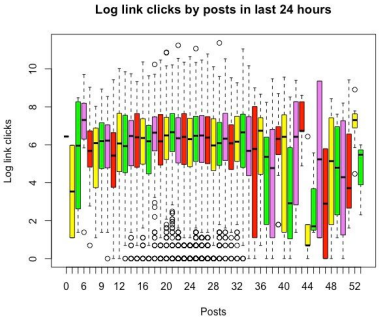




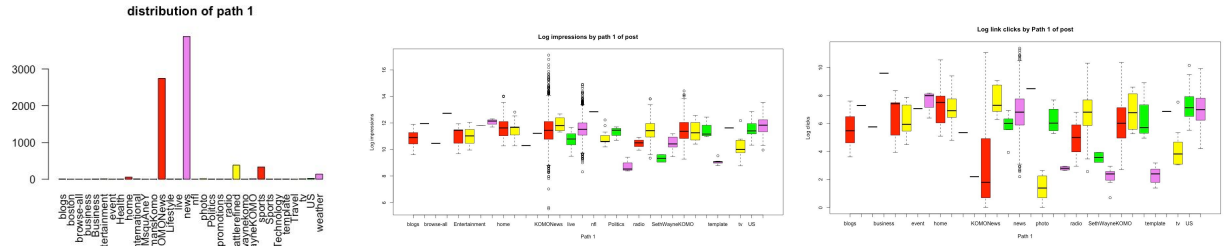
Posts in past 24 hours: The pattern is harder to see here:



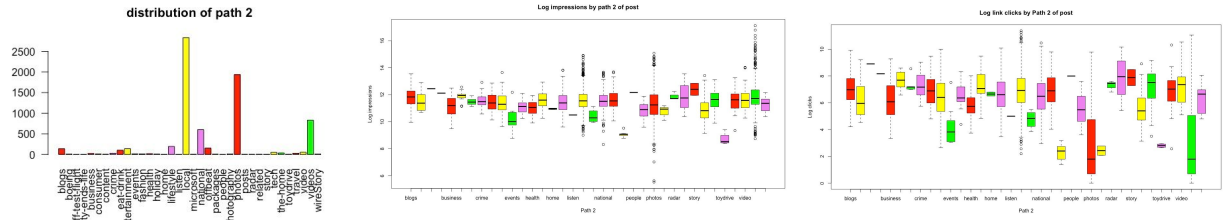




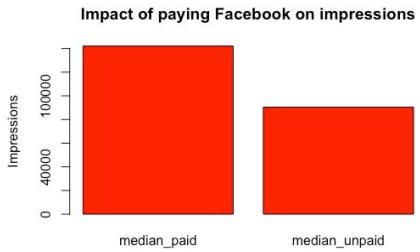
Path 1: The vast majority of posts are routed to “Komo News” and “News” for path 1, and there are a large number of paths that are rarely used. There is large variance in the impressions per post by path 1, but two largest path 1 factors (News, in purple, at the center, and KOMONews, in red near the center), show no noticeable difference. However KOMO-News pathed posts generate almost no link clicks. Following this analysis we removed all Path 1 factors with less than 20 observations.



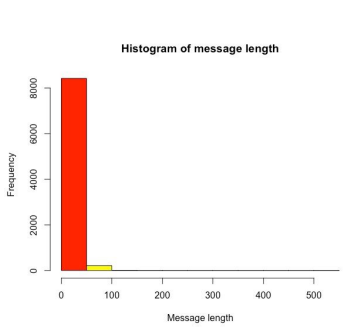
Path 2: The majority of posts are routed to “local,” “photos” and “videos” on path 2, and again there are many paths that are rarely used. Path 2 shows some promise as a predictor of impressions, and also as a predictor of link clicks. Following this analysis we removed all Path 2 factors with less than 20 observations.

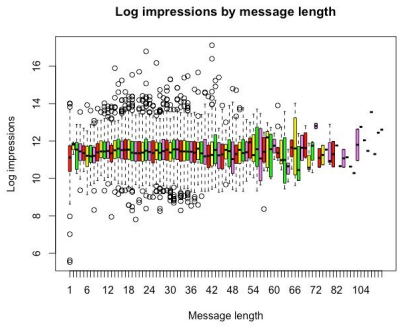


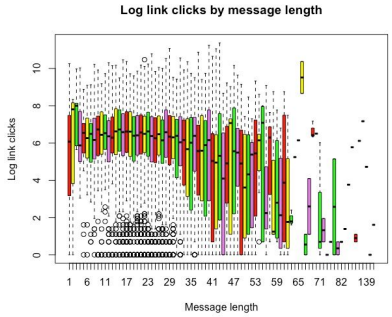
Paid: There are only 24 posts that KOMO-TV paid Facebook to boost: too few to show up on a chart next to the 8,615 unpaid posts. However, it appears on the surface that KOMO-TV is getting something for its money: the median paid posts gets 57% more impressions than the median unpaid post, though there is likely a sample bias here, since KOMO-TV is unlikely to pay to boost poor posts:



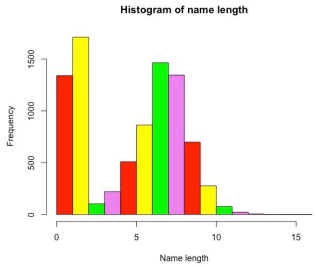
Message length: On this scale, message length looks like it could benefit from a log transformation, however for consistency with the other “length” variables we left it untransformed. It has no noticeable correlation with impressions but longer messages are associated with low link clicks.

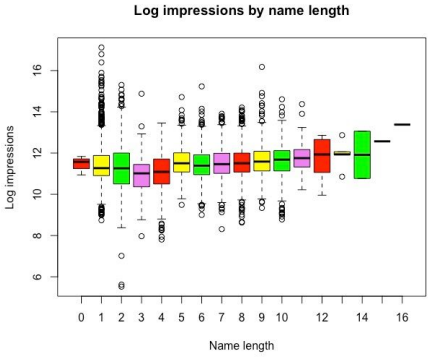


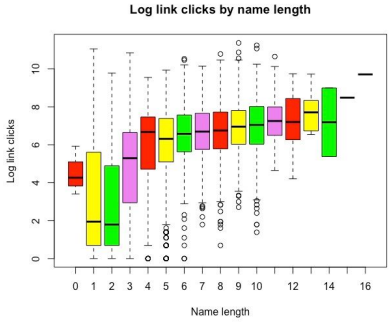




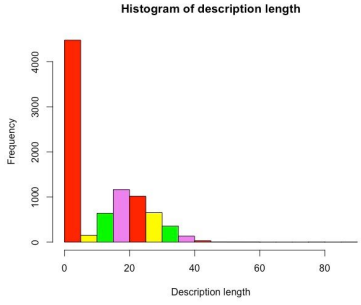
Name length: in contrast, long headlines seems to result in more link clicks, and slightly more impressions.

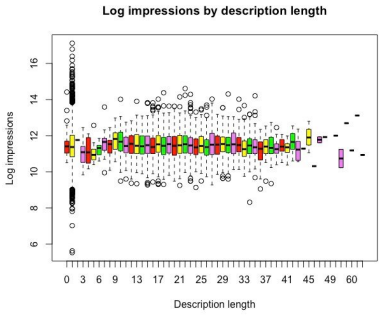


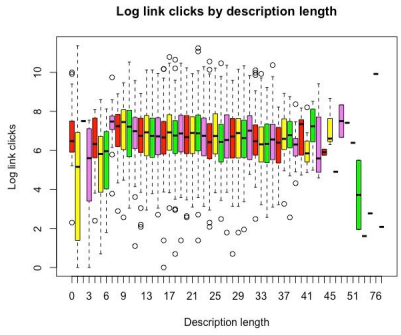




Description length: does not appear to matter much.







Message omegas: these were built with R’s “Topics” package, which used stochastic deviance minimization to group words which appear together most commonly in the “message” field into 10 principal components, selecting K, the number of principal components, with the highest Bayes Factor. We stripped all words of less than three letters and all punctuation, and removed all words that appeared less than five times in the dataset before building the topics. By observing the top 50 most common words in each factor we were able to determine common

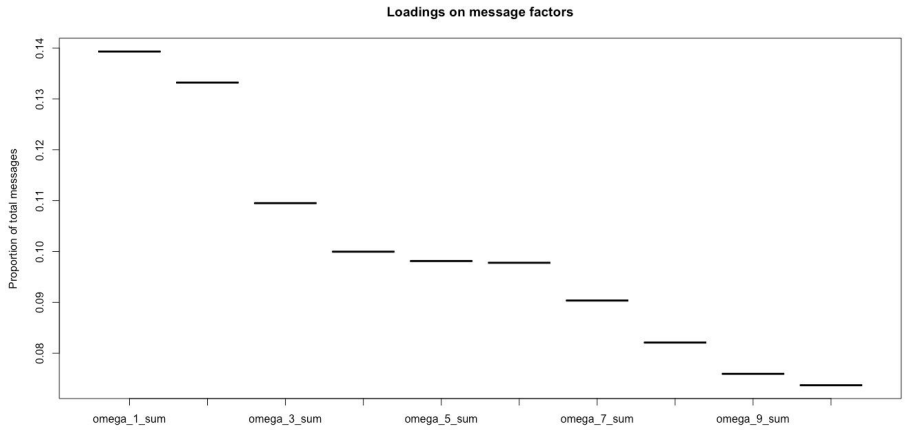
themes or genres for each of the topics. Sorting the data by the weighting, or “omega” on each of the 10 factors then looking up the individual posts allowed us to confirm that our interpretations appeared correct and that the topics model very cleanly separated the text into topics. For example, the posts with the highest omega 1 scores (representing “crime”) were all crime stories.

The top 50 words for each topic were:

```
[1] 'driveway', 'stabbing', 'vehicular', 'murdersuicide', 'stabbed', 'knife', 'lifethreatening', 'seconddegree', 'firstdegree', 'booked', 'unconscious', 'shoplifting', 'she
d', 'suspect', 'grabbed', 'homicide', 'arrested', 'robbery', 'murder', 'attempted', 'suspects', 'boy', 'assaulted', 'raped', 'suspected', 'accomplice', 'victim', 'custody',
'officers', 'attacked', 'police', 'prosecutors', 'accused', 'womans', 'hid', 'deputies', 'detectives', 'suspicion', 'taxi', 'suspicious', 'charged', 'fleeing', 'amber', 'sa
feway', 'killing', 'wang', 'argument', 'father', 'arrest', 'felony' (13.9)
[2] 'palouse', 'patchy', 'damp', 'thundershowers', 'blustery', 'midweek', 'mainly', 'thereafter', 'breezy', 'thundershower', 'lowlands', 'wrap', 'shubha', 'ridge', 'gorlin'
, 'dreamy', 'totals', 'tirumale', 'larry', 'meg', 'cloudy', 'advisory', 'sunbreaks', 'gusty', 'travelingbhat', 'sreedharan', 'sigma', 'highs', 'snowfall', 'lenticular', 'th
understorms', 'foggy', 'meteorologist', 'sunshine', 'lunar', 'beauty', 'mild', 'gorgeous', 'backwoods', 'odonnell', 'skies', 'clouds', 'showers', 'photography', 'mid', 'set
h', 'pittman', 'wet', 'shannon', 'wayne' (13.3)
[3] 'gays', 'stutzman', 'guidelines', 'expanded', 'violates', 'upheld', 'beliefs', 'carbon', 'supreme', 'reform', 'samesex', 'approved', 'legalize', 'revenue', 'minimum', '
immigration', 'appeals', 'tax', 'opponents', 'state', 'wage', 'proposed', 'inslee', 'governor', 'ruling', 'states', 'lawmakers', 'projects', 'budget', 'leaders', 'marijuana
', 'law', 'transportation', 'shellfish', 'lesbians', 'islamic', 'jay', 'measure', 'recreational', 'recall', 'legalized', 'campaign', 'tobacco', 'gay', 'gov', 'laws', 'natio
ns', 'legal', 'policy', 'resources' (11)
[4] 'httpbitly1xiremk', 'rookie', 'willson', 'xlix', 'warmups', 'locker', 'norwood', 'rallying', 'fumble', 'autographs', 'seavsphi', 'jermaine', 'hauschka', 'giants', 'reli
ve', 'offense', 'sfvssea', 'bobby', 'defending', 'pregame', 'cowboys', 'helfet', 'kearse', 'louder', 'halftime', 'panthers', 'yards', 'matchup', 'homeward', 'adoptable', 'f
uzzy', 'baldwin', 'caturday', 'kitties', 'rams', 'pals', 'interception', 'cardinals', 'gamewinning', 'humane', 'preseason', 'oakland', 'packers', 'adoption', 'furry', 'touc
hdown', 'glendale', 'chancellor', 'wagner', 'gohawks' (10)
[5] 'lied', 'luckthe', 'itthe', 'luckanswer', 'luck
answer', 'what', 'itanswer', 'playingits', 'itanswer', 'answer', 'liveonkomoanswer', 'qotd', 'whoknew', 'survey', 'question', 'playing', 'answer', 'fries', 'recent', 'avera
ge', 'breaka', 'time', 'admit', 'americans', 'spends', 'doing', 'panda', 'women', 'why', 'valentines', 'common', 'teach', 'almost', 'curious', 'annoying', 'entertaining', '
cutest', 'reveals', 'fetch', 'lifetime', 'walks', 'boss', 'relationship', 'vacations', 'cute', 'pup', 'paper', 'shopping', 'loves', 'refuse' (9.8)
[6] 'yesler', 'westbound', 'directions', 'sr99', 'boil', 'delays', 'northbound', 'units', 'lanes', 'semi', 'spreading', 'crews', 'fire', 'fires', 'updating', 'blocked', 'so
uthcenter', 'brush', 'twisp', 'reopened', 'okanogan', 'sr167', 'developing', 'link', 'closed', 'homes', 'evacuation', 'dot', 'firefighters', 'blocking', 'flames', 'repairs'
, 'leak', 'i90', 'large', 'traffic', 'germanwings', 'chelan', 'complex', 'fuel', 'debris', 'burned', 'destroyed', 'precaution', 'bees', 'repair', 'omak', 'buildings', 'fire
fighting', 'jet' (9.8)
[7] 'elwes', 'westley', 'pippa', 'probably', 'theyre', 'really', 'donate', 'everybody', 'kind', 'exactly', 'petersen', 'bloody', 'theaters', 'joke', 'ann', 'arent', 'salvat
ion', 'burke', 'cake', 'nothing', 'wait', 'seems', 'plaza', 'priceless', 'stops', 'troubled', 'interested', 'toy', 'refund', 'flat', 'painting', 'expecting', 'mode', 'under
stand', 'kindness', 'fanfest', 'childrens', 'chat', 'tweets', 'bar', 'sense', 'questions', 'recorded', 'floor', 'having', 'heart', 'own', 'chocolate', 'chicken', 'chemicals
' (9)
[8] 'arquettes', 'for
support', 'hollywoods', 'sugarland', 'httpbitly1xjtqwt', 'serena', 'httpbitly1khovyw', 'pick
help', 'flo', 'httpbitly16uino3', 'patricia', 'charityif', 'givelove', 'httpbitly17bnxug', 'kristian', 'bankerscare', 'bankerscarecom', 'fooddrive', 'humanitarian', 'gravit
y', 'chideo', 'oscarwinning', 'sponsored', 'harvest', 'oscars', 'actress', 'bush', 'bradley', 'tulip', 'oscar', 'tulips', 'oscarsonkomo', 'artist', 'oscars2015', 'seattlear
ea', 'paintings', 'mmarkovichkomotvcom', 'awards', 'wines', 'restaurants', 'contestant', 'inspiring', 'childhood', 'foundation', 'cobain', 'brain', 'rent', 'vip', 'academy'
, 'microsoft' (8.2)
[9] 'baes', 'jaylen', 'bae', 'httpbitly1vzyfn2', 'hatch', 'lewismcchord', 'indict', 'mphs', 'darren', 'school', 'marysvillepilchuck', 'pilchuck', 'marysville', 'students',
'missouri', 'high', 'disneyland', 'classmates', 'ebola', 'grand', 'korea', 'ferguson', 'fryberg', 'teacher', 'abc', 'king', 'jury', 'emotional', 'infected', 'tested', 'spok
eswoman', 'jake', 'tragedy', 'powell', 'elementary', 'polls', 'resigned', 'election', 'httpbitly1bubo8t', 'allegations', 'cerebral', 'palsy', 'cosby', 'brown', 'tens', 'kit
sap', 'protestors', 'infection', 'protests', 'staff' (7.6)
[10] 'rips', 'koopmans', 'zoo', 'dutch', 'bamboo', 'chai', 'brad', 'royal', 'kelly', 'goode', 'jackson', 'shells', 'elephant', 'woodland', 'calf', 'greenpeace', 'angeles',
'stunt', 'orca', 'duchess', 'gets', 'drill', 'sanctuary', 'rig', 'alltime', 'arctic', 'aquarium', 'defiance', 'pod', 'caves', 'capitol', 'swam', 'paris', 'train', 'port', '
elephants', 'queen', 'hottest', 'creates', 'dramatic', 'cubs', 'dies', 'birthday', 'rush', 'encounter', 'set', 'humpback', 'anne', 'escaped', 'oil' (7.4)
```

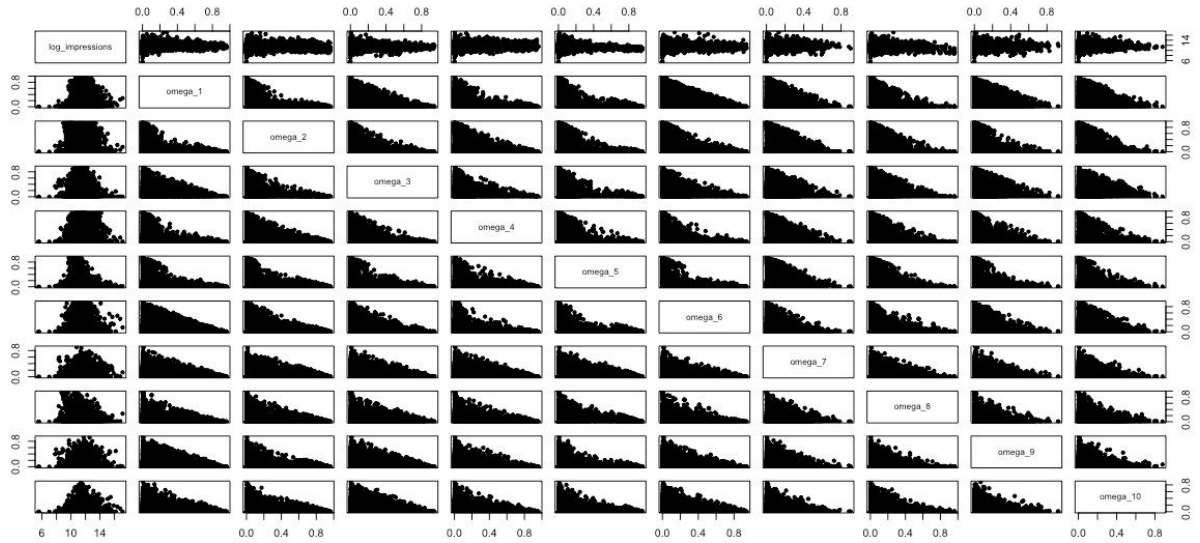
We interpreted these groups as: 1: Crime; 2: Weather; 3: Social issues; 4: Sports; 5: Whimsy; 6: Traffic; 7: Human interest; 8: Celebrity; 9: Schools[14]; and 10. Animals.

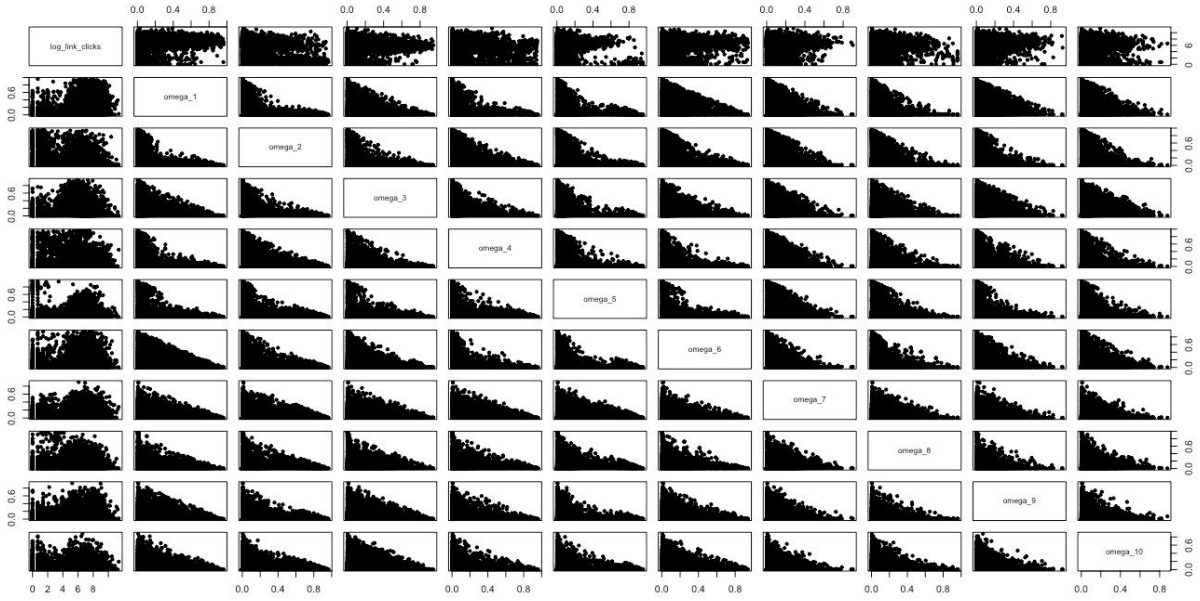
The topics decrease in frequency of use from 1 to 10, with crime making up 13.9% of all the messages and animals accounting for just 7.4%:



[14] This was the hardest of the topics to describe — though it included a lot of school words there were also unrelated words in there.

Although we had great hope for these topics, plotting the omegas against log impressions (top chart, top row, below) and log link clicks (bottom chart, top row) did not immediately yield any clear association, except in variance:





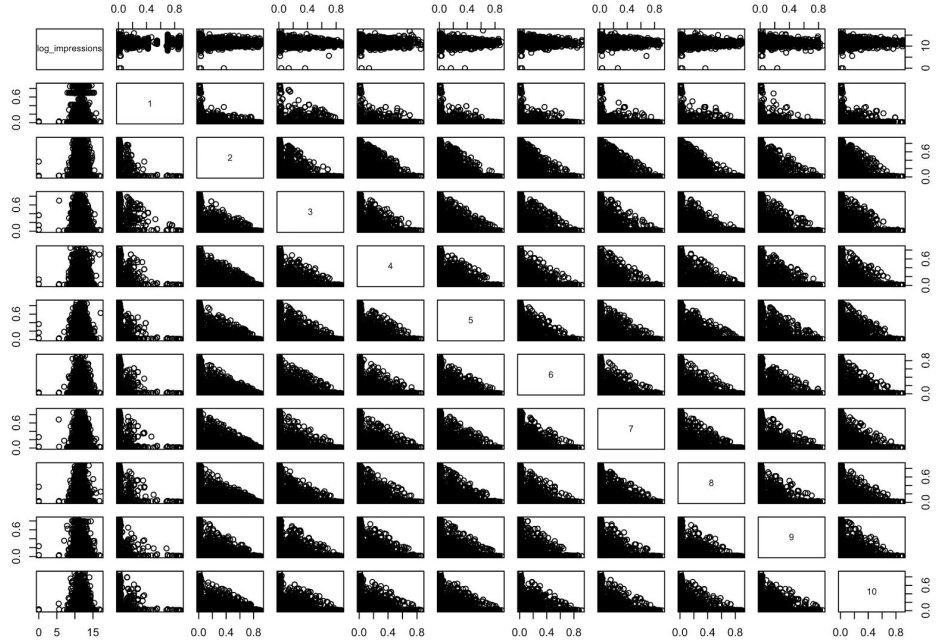
Name words omegas: these were again built with R's "Topics" package, following the same methodology as for the message texts. Again Topics broke the text in the "name" or headline field into 10 principal components. These were slightly harder to interpret than the message omegas, perhaps because the name field contains fewer words than the message.

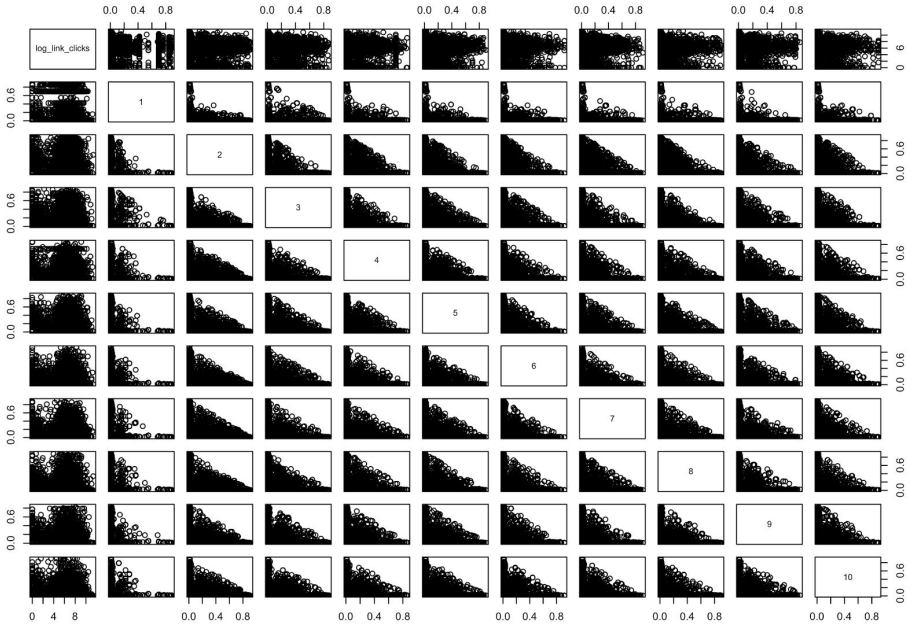
The top 50 words for each topic were:

[1] 'woodinvilles', 'homeward', 'humane', 'society', 'adoptable', 'post', 'timeline', 'photos', 'newss', 'cats', 'pet', 'pets', 'komo', 'from', 'sunsets', 'stunning', 'sunset', 'cover', 'photo', 'friend', 'moon', 'earth', 'brilliant', 'classic', 'escapes', 'rocks', 'hanging', 'abandoned', 'cliff', 'carpet', 'unique', 'steve', 'celebrate', 'thank', 'seahawk', 'flowers', 'valentines', 'banned', 'selfies', 'championships', 'canadian', 'sendoff', 'alive', 'dna', 'videos', 'visit', 'seen', 'autistic', 'spirit', 'academy' (20.4)
[2] 'vehicular', 'standoff', 'fleeing', 'homicide', 'police', 'suspect', 'attempted', 'hitandrun', 'man', 'driver', 'arrested', 'car', 'stabbed', 'fatal', 'bail', 'offender', 'charged', 'sought', 'child', 'chase', 'robbery', 'suspects', 'drunk', 'officers', 'publics', 'assault', 'pleads', 'shoots', 'critically', 'dies', 'headon', 'husband', 'burglary', 'girl', 'killed', 'soldier', 'teen', 'toddler', 'gunfire', 'woman', 'highspeed', 'kidnapping', 'driveby', 'amber', 'leg', 'manhunt', 'hurt', 'homeowner', 'troopers', 'guilty' (11.1)
[3] 'balance', 'politically', 'extension', 'packers', 'alltime', 'chancellor', 'nfc', 'contract', 'seth', 'wayne', 'international', 'the', 'russell', 'marshawn', 'wilson', 'championship', 'day', 'sounders', 'defiance', 'best', 'lynch', 'richard', 'seattle', 'teachers', 'west', 'host', 'opener', 'strike', 'draft', 'huskies', 'festival', 'earth', 'streak', 'parade', 'sherman', 'space', 'training', 'chiefs', 'weeks', 'warmest', 'fans', 'america', 'pride', 'point', 'things', 'stars', 'scenes', 'viewer', 'celebrate', 'valentines' (10.1)
[4] 'uploads', 'mobile', 'region', 'school', 'state', 'puget', 'agreement', 'million', 'students', 'island', 'parents', 'oregon', 'sound', 'into', 'western', 'mercer', 'teacher', 'house', 'pole', 'lapse', 'bainbridge', 'pool', 'buy', 'wins', 'high', 'kurt', 'behind', 'steve', 'sold', 'sale', 'eastern', 'shooter', 'taxes', 'shows', 'steal', 'university', 'fair', 'spike', 'bulldog', 'happy', 'captured', 'wanted', 'ashes', 'plans', 'home', 'cross', 'female', 'crashes', 'dark', 'delivery' (9.6)
[5] 'drilling', 'billion', 'knox', 'arctic', 'rig', 'gov', 'elephants', 'inslee', 'amanda', 'wage', 'higher', 'ceo', 'marriage', 'mariners', 'blames', 'minimum', 'zoo', 'group', 'for', 'against', 'transportation', 'airline', 'pacific', 'american', 'announces', 'drill', 'supreme', 'pit', 'lawsuit', 'race', 'tunnel', 'employees', 'over', 'request', 'hernandez', 'tax', 'court', 'seeks', 'file', 'workers', 'suit', 'issue', 'pay', 'build', 'planes', 'beat', 'drivers', 'viaduct', 'sues', 'woodland' (9.2)
[6] 'flo', 'rida', 'hang', 'newborn', 'fame', 'there', 'bowl', 'super', 'surprise', 'cubs', 'christmas', 'pays', 'toy', 'puppy', 'local', 'gets', 'santa', 'kitten', 'trip', 'interview', 'cat', 'blue', 'kelly', 'drive', 'test', 'loud', 'snoqualmie', 'patriots', 'hall', 'stuck', 'california', 'jail', 'golden', 'macklemore', 'get', 'gifts', 'away', 'lion', 'little', 'thief', 'try', 'cafe', 'just', 'dad', 'perform', 'win', 'measles', 'gift', 'horse', 'attention' (8.3)
[7] 'caves', 'climber', 'avalanche', 'quickcast', 'bag', 'nepal', 'everest', 'ice', 'rainier', 'cause', 'session', 'four', 'earthquake', 'ship', 'service', 'special', 'collapse', 'authorities', 'recall', 'two', 'since', 'church', 'south', 'cnncom', 'finds', 'quake', 'king', 'glitch', 'released', 'july', 'raid', 'cave', 'baltimore', 'ties', 'bay', 'members', 'national', 'farm', 'head', 'wildlife', 'fireworks', 'broken', 'severely', 'animals', 'call', 'likely', 'popular', 'enumclaw', 'lanes', 'blast' (8)
[8] 'twisp', 'evacuation', 'august', 'explodes', 'brush', 'warnings', 'wenatchee', 'outages', 'windstorm', 'strong', 'walla', 'crews', 'chelan', 'burns', 'wildfire', 'fallen', 'rare', 'firefighters', 'debris', 'fire', 'square', 'fires', 'wildfires', 'cub', 'grows', 'smoke', 'evacuated', 'hill', 'county', 'area', 'seattles', 'capitol', 'news', 'facility', 'arrives', 'pioneer', 'pierce', 'building', 'columbia', 'pounds', 'than', 'okanogan', 'flames', 'power', 'crazy', 'battle', 'gas', 'damage', 'massive', 'flee' (7.9)
[9] 'closings', 'openings', 'metro', 'holiday', 'week', 'ebola', 'wasnt', 'shut', 'medicine', 'tonight', 'health', 'background', 'this', 'not', 'ferguson', 'room', 'therapy', 'pentagon', 'snow', 'despite', 'football', 'bust', 'dogs', 'use', 'where', 'vacation', 'college', 'prosecutor', 'night', 'care', 'halloween', 'together', 'other', 'marijuana', 'cuts', 'order', 'hes', 'grand', 'transit', 'testing', 'korea', 'latest', 'really', 'fall', 'restaurant', 'says', 'check', 'playing', 'morning', 'bathroom' (7.7)
[10] 'institutes', 'tribeca', 'album', 'untitled', 'arquettes', 'givelove', 'patricia', 'sr99', 'tulips', 'experience', 'film', 'semitruck', 'york', 'academy', 'indiana', 'those', 'silence', 'food', 'des', 'moines', 'traffic', 'would', 'safeco', 'owner', 'live', 'long', 'religious', 'penalty', 'france', 'life', 'need', 'thrift', 'old', 'emerald', 'worker', 'plays', 'light', 'year', 'skagit', 'takes', 'northern', 'cant', 'unveils', 'construction', 'cargo', 'urban', 'wont', 'campaign', 'military', 'lights' (7.6)

We interpreted these groups as: 1: Nature and beauty; 2: Crime; 3: Sports; 4: Education?; 5: Industry/Courts; 6: Unclear; 7: Outdoors; 8: Fire; 9: Unclear; and 10. Unclear.

Plotting the name omegas against log impressions and log link clicks showed little association except in variance:





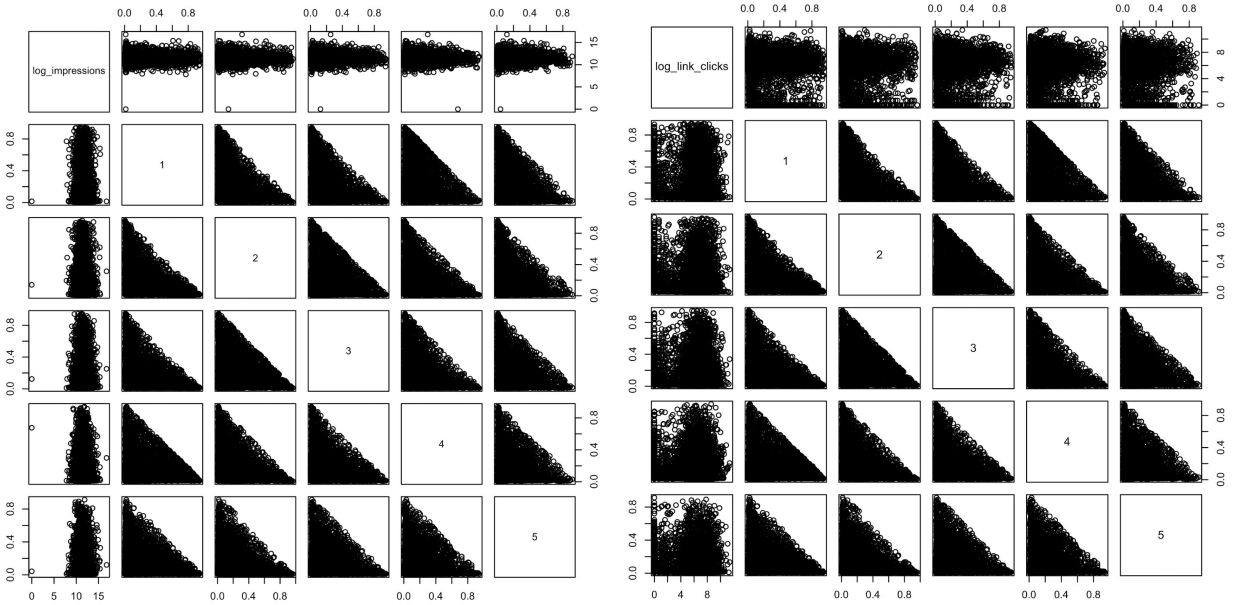
Description omegas: these were again built with R's "Topics" package, following the same methodology as for the message and name texts. In this case Topics broke the text in the "description" or sub-headline field into 5 principal components. As with the names omegas, these were harder to interpret than the message omegas, again perhaps because the description field contains fewer words than the message, and also perhaps because it is not always used.

The top 50 words for each topic were:

[1] 'manhunt', 'assaulted', 'murdersuicide', 'argument', 'wound', 'abducted', 'stabbed', 'clerk', 'stabbing', 'kidnapping', 'familys', 'hitandrun', 'suspicion', 'lacey', 'robber', 'offender', 'sheriffs', 'gunpoint', 'deputies', 'daughter', 'recovered', 'woman', 'man', 'homicide', 'standoff', 'robbed', 'driveby', 'drunk', 'police', 'suspected', 'robbing', 'superior', 'detectives', 'tried', 'armed', 'thurston', 'officers', 'vehicular', 'crook', 'invasion', 'son', 'loose', 'teenager', 'unarmed', 'puyallup', 'jail', 'kidnapped', 'attempted', 'car', 'investigating' (23)
[2] 'channel', 'mean', 'imagine', 'sounds', 'else', 'kitten', 'youre', 'puppy', 'lapse', 'weve', 'know', 'theres', 'watching', 'really', 'whole', 'eclipse', 'how', 'devices', 'maybe', 'great', 'moon', 'watch', 'jobs', 'these', 'see', 'summers', 'little', 'like', 'although', 'week', 'version', 'dont', 'things', 'want', 'thing', 'kick', 'eat', 'heres', 'think', 'thats', 'gallery', 'you', 'maps', 'display', 'puget', 'our', 'too', 'nothing', 'look', 'sight' (21)
[3] 'ninth', 'comicon', 'leisurely', 'gals', 'premiere', 'homered', 'innings', 'syris', 'pitcher', 'carpet', 'anime', 'handsome', 'earned', 'inning', 'lets', 'award', 'forever', 'enter', 'convention', 'studio', 'felix', 'nelson', 'victory', 'dempsey', 'mariners', 'gentle', 'win', 'walks', 'match', 'academy', 'auction', 'enjoy', 'sounders', 'nfc', 'ncaa', 'franchise', 'woodland', 'single', 'included', 'green', 'festival', 'licensed', 'names', 'clint', 'mix', 'elephants', 'beat', 'slow', 'bullpen', 'film' (20.2)
[4] 'tested', 'havoc', 'lewismcchord', 'tuition', 'measles', 'brush', 'wildfires', 'blaze', 'ebola', 'budget', 'northbound', 'firefighters', 'route', 'battling', 'raging', 'rainfall', 'senate', 'resources', 'elson', 'bacteria', 'immediate', 'firefighting', 'health', 'disease', 'smoke', 'fire', 'complex', 'twisp', 'rain', 'patients', 'measure', 'serious', 'troy', 'worker', 'efforts', 'areas', 'medical', 'effect', 'harborview', 'wildlife', 'shoreline', 'atop', 'swept', 'dangerous', 'began', 'critical', 'state', 'action', 'thick', 'education' (18.8)
[5] 'snarl', 'rogerson', 'stephanie', 'tipskomo4newscom', 'clapper', 'knox', 'knoxs', 'copilot', 'italys', 'seaports', 'malaysia', 'planes', 'amanda', 'offshore', 'dutch', 'todo', 'jan', 'comments', 'brady', 'print', 'seattletacoma', 'pst', 'updated', 'footballs', 'drilling', 'derailed', 'amtrak', 'communications', 'intelligence', 'rig', 'contract', 'email', 'avalanche', 'patriots', 'everest', 'shells', 'activists', 'cargo', 'contact', 'arctic', 'airlines', 'transfers', 'democratic', 'philadelphia', 'airport', 'prime', 'sightseeing', 'talks', 'sr99', 'ship' (17)

We were able to determine that Topic 1 was crime; Topic 3 appears to be a combination of movies and sports; and that Topic 4 seems to have some kind of weather/environmental theme, but again, with the exception of Topic 1, the divisions are not clear.

Plotting against log impressions (left chart, below, top row) and log link clicks (right chart, below, top row) did not demonstrate any clear association except in variance:

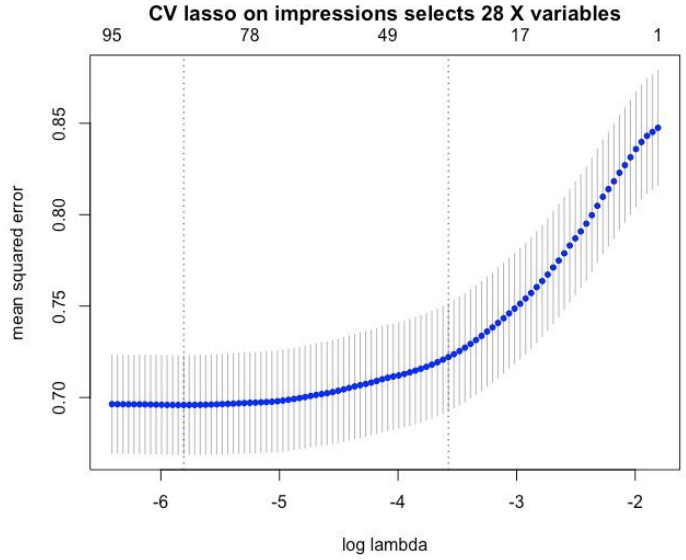


5. Modeling Impressions

Armed with a basic understanding of each of the explanatory variables, we set out to predict log impressions using cross-validated lasso regression, trees and random forest models. Our hypotheses were that each of the potential explanatory variables would have some impact on impressions.

5.1 CV Lasso

We ran a 100-fold cross validated lasso on log impressions, scaling all continuous variables. The lasso selected 28 explanatory variables and rejected 67 variables. It had an out-of-sample R^2 of just 0.15.



The 28 selected variables, their coefficients and the interpretations of how they affect impressions on the original scale are shown in the table below.[15] Many intuitions about local news are re-enforced.[16] In order of importance, including video, posting in January, linking to the 'weather' section of the KOMO website, posting more about crime, human interest, sports and writing longer headlines increases impressions.[17] Paying Facebook to boost the post doesn't make nearly as much difference as might be expected — just an additional 2,680 impressions, on average.[18] Posting in the dead of night unsurprisingly results in fewer impressions, as does posting in September of November, failing to link to a section of the KOMO website, failing to

[15] The interpretation is referenced against the intercept of 11.43, which corresponds to exp(11.43)=92,042 impressions, such that the impact of an extra word in the headline in the original scale is, for example: exp(11.43+1(0.0536))-exp(11.43)

[16] While none of us have lived in Seattle, crime, weather, human interest and sports are staples of local TV news and the saying "if it bleeds it leads" is as old as TV news.

[17] The effect of longer headlines increases linearly in this model, meaning adding nine more words would make it the most important factor, according to the model.

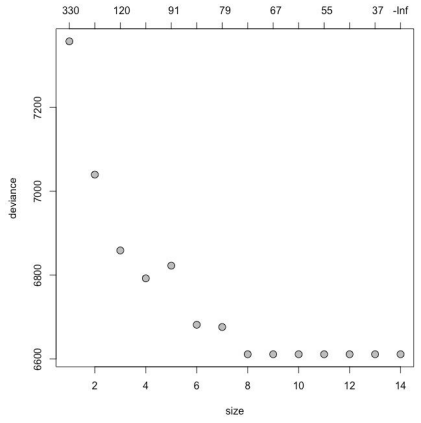
[18] It's worth noting that the coefficients vary slightly each time we ran the 100-fold CV lasso: paid, which has just 24 data points, had a coefficient as low as +0.02 in one model we ran.

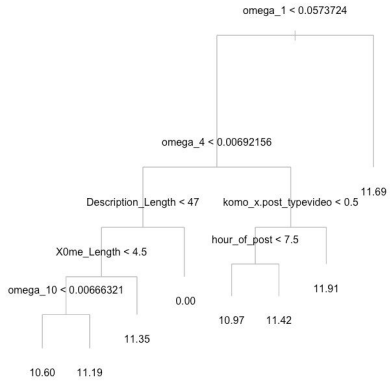
include a link of any kind, or writing about whimsy or celebrity. Interestingly, while the length of the headline seems to matter, the topics in it do not seem to, given the topics in the main message body.

Variable	Coefficient	Interpretation
Name length	0.0536	Increasing the length of the headline by one word adds 5,072 impressions, on average
Paid	0.1074	Paying Facebook to promote the post adds 10,440 impressions, on average
Posts last 3 hours	0.0523	An extra post within the last three hours is worth 4,937 impressions, on average
Message omega 1	0.1363	Including a standard deviation more loading on “crime” in the message adds 13,436 impressions, on average
Message omega 2	-0.0010	Including a standard deviation more loading on “social issues” in the message subtracts 95 impressions, on average
Message omega 4	0.0633	Including a standard deviation more loading on “sports” in the message adds 6015 impressions, on average
Message omega 5	-0.0258	Including a standard deviation more loading on “whimsy” in the message subtracts 2,346 impressions, on average
Message omega 6	0.0221	Including a standard deviation more loading on “traffic” in the message adds 2,060 impressions, on average
Message omega 7	0.0860	Including a standard deviation more loading on “human interest” in the message adds 8,262 impressions, on average
Message omega 8	-0.0216	Including a standard deviation more loading on “celebrity” in the message subtracts 1,968 impressions, on average
Message omega 9	0.0582	Including a standard deviation more loading on “schools” in the message adds 5,513 impressions, on average
Message omega 10	0.0013	Including a standard deviation more loading on “animals” in the message adds 116 impressions, on average
Name omega 7	0.0031	Including a standard deviation more loading on “outdoors” in the headline adds 286 impressions, on average
Description omega 3	-0.0166	Including a standard deviation more loading on “movies/sports” in the description subtracts 1,520 impressions, on average
Text posts	-0.0024	Posts with text and no link cost 221 impressions, on average
Video posts	0.4611	Posts containing video add 53,921 impressions, on average
2am	-0.1925	Posts published in the 2 a.m. hour get 16,121 less impressions, on average
3am	-0.2297	Posts published in the 3 a.m. hour get 18,890 less impressions, on average
4am	-0.3244	Posts published in the 4 a.m. hour get 25,499 less impressions, on average
5am	-0.2375	Posts published in the 5 a.m. hour get 19,458 less impressions, on average
7am	-0.0844	Posts published in the 7 a.m. hour get 7,451 less impressions, on average
January	0.2471	Posts published in January get 25,794 more impressions, on average
September	-0.0376	Posts published in September get 3,394fewer impressions, on average
November	-0.0790	Posts published in November get 6,990 fewer impressions, on average
Saturday	0.0033	Posts published on Saturdays get 301 more impressions, on average
No path 2	-0.0873	Posts without a path 2 get 7,691 fewer impressions, on average
No path 1	-0.0608	Posts without a path 1 get 5,429 fewer impressions, on average
‘Weather’ path 1	0.2308	Posts routed to the “weather” section of the KOMO website get an extra 23,898 impressions, on average
‘Tech’ path 2	-0.1483	Posts routed to the “tech” subsection of the KOMO website get 12,684 fewer impressions, on average

5.2 CART Model

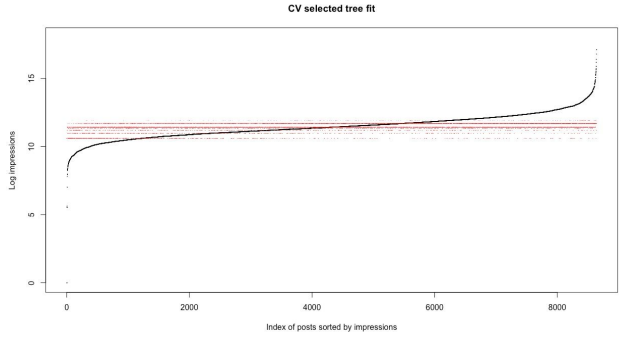
Next, we tried a tree model built using the CART algorithm to predict impressions. Setting the initial minimum node size to 1 and a minimum deviance requirement of 0.005 to proceed with a new split, then running a 100-fold cross validation resulted in a pruned tree with 8 nodes:[19]

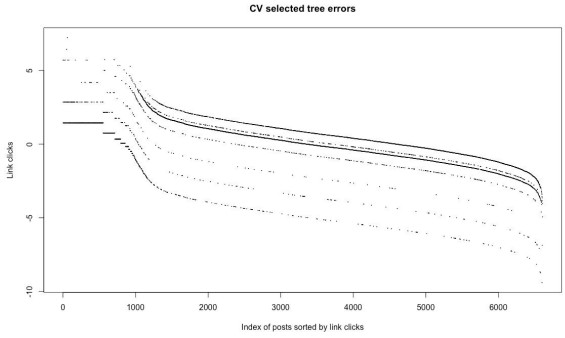




The model finds that a message text loading of less than 0.057 on crime is the most important split and results in a lower number of impressions, unless there is a higher loading on sport and a video. Posts with a higher loading on sports that lack a video do better if they are not posted in the dead of night. Posts with very low loadings on crime and sports with long sub-heads are predicted to perform worst of all — they are expected to have no impressions whatsoever. Posts with low loadings on crime and sports, with shorter sub-heads do better if they have long headlines. Among this set, those that have short headlines do better if they have animals in them. No other variables are chosen.

The plot below on the left shows the posts in rank order of impressions in black, with the model's predictions in red. We can see that the model significantly underpredicts how well the best posts do, and how badly the worst posts do. The plot on the right shows residuals, and demonstrates this bias toward the average more clearly:

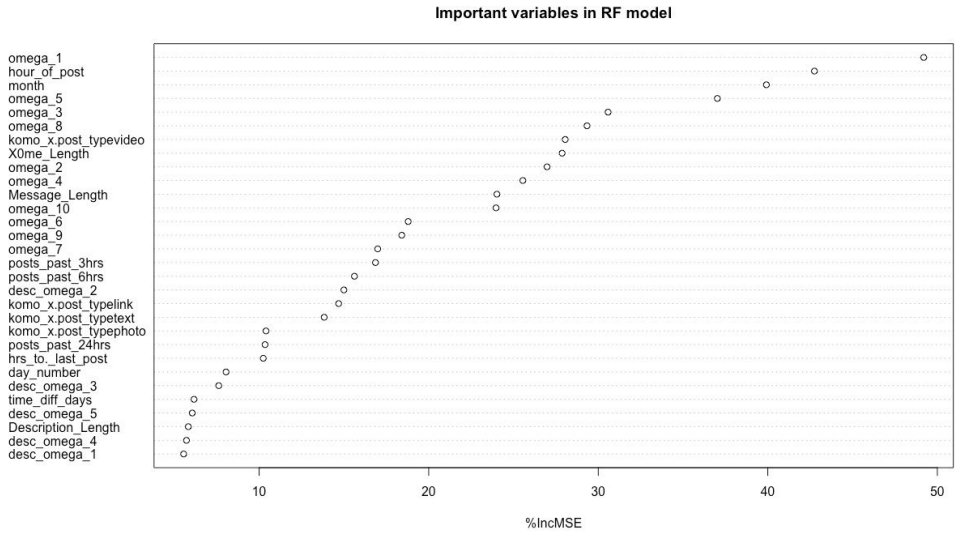




[19] Since the CART algorithm does not cope well with more than 32 variables at a time, we excluded the path, description omega and name omega variables from this analysis. We removed these as they seemed the least important variables, based on the prior lasso regression. The description and name omegas were also hard to interpret, and the path variable risked confounding with message omegas.

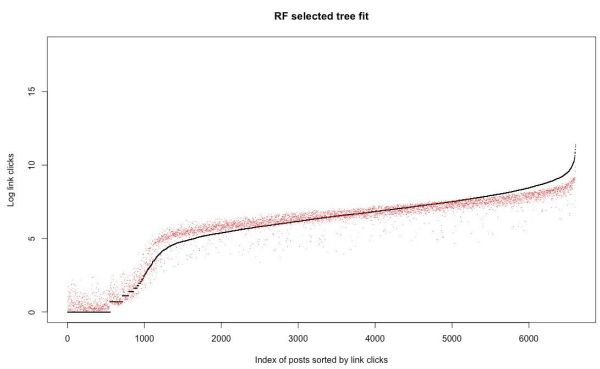
5.3 Random Forest Model

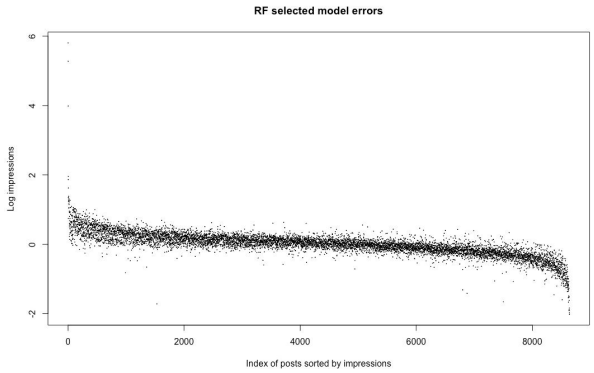
While the tree model is useful for interpretation, it can be improved upon when it comes to prediction with a random forest. A 500 tree forest yields the following important factors:



Many of the same things we saw in the lasso are repeated here. By order of importance, the loading on “crime” in the message, the time of day the post, the month, the loading on ‘whimsy,’ ‘social issues’ and ‘celebrity,’ the inclusion of video and the length of the headline are the top eight determinants of impressions. The random forest does not find paying Facebook to be a useful predictor of impressions. Unlike the lasso, it does find the loading on ‘weather’ important, though this likely reflects the fact that we left the “path” variables out of the random forest selection model, including the “weather” section of KOMO’s website, which probably had a confounding effect in the lasso.

The model fits much better than the CART model, especially across the middle of the range of performance, and has in-sample R^2 of 0.87, which may indicate an overfit. Out of bag R^2 is 0.28, still nearly twice that of the lasso model. Looking at the predicted impressions (in red, left) against the ranked actual observations (in black, left), and at the residuals (right), we see that the model still underpredicts how well the best posts will do, and overpredicts how well the worst posts will do:



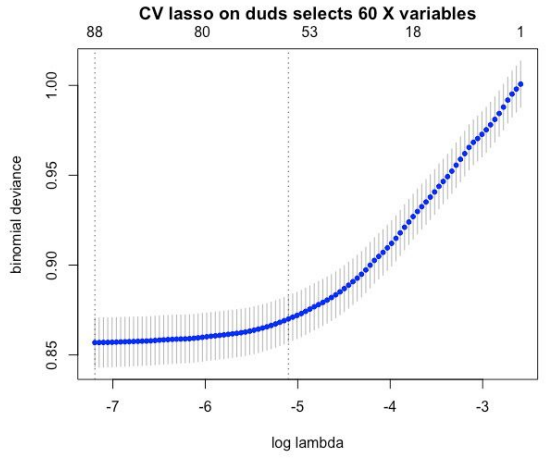


5.4 Categorical models to predict hits, duds

Since our general models fail to predict well at the extremes of the distribution, we hoped that a categorical model, which relaxes the requirement on a tight fit through the mean, might do a better job. Recall that the best 13.5% of posts account for 50% of all impressions ("hits") and that the worst 20% of posts account for just 4% of all impressions ("duds"). We built categorical models using cross validated lassos and random forests to attempt to predict both hits and duds. If KOMO knew when it had a hit, it could pay to promote it, or launch it at an ideal time. If KOMO knew when it had a dud, it could kill or rejig the post to improve it. However, the basic problem with all of our attempts to model hits and duds was that they attempted to enforce categorization on a fundamentally continuous vector of impressions. Still, we gave it a go, starting with an attempt to predict duds.[20]

5.4.1 Lasso to predict duds

A 100-fold cross validated binomial regression to predict duds has an error rate of (1,571+60)/8,639 = 0.19. It was quite complicated, selecting 60 X variables:



	Pred 0	Pred 1	error
True 0	6852	60	0.278 (false positive)
True 1	1571	156	0.187 (false negative)

Despite the large number of X variables, it does a terrible job categorizing duds if we categorize duds as anything predicted to be a dud at probability greater than 0.5, with a false positive rate of 0.28. If used, this would result in KOMO throwing 3 posts which were in fact not duds for every 7 that were — all while it accurately predicts less than 10% of the actual duds. It has a false negative rate of 0.19. This is presumably less important to KOMO — since it is currently publishing the duds, a false negative is not damaging, relative to the status quo.

[20] While it would have been possible to predict both hits and duds at the same time in a multinomial model, we preferred to build separate binomial models to start with for ease of interpretation.

We tried retuning the rule to minimize false positives. Since we do not know the function that determines the economic value of a lost impression, we could not come up with an ideal rule. But setting the model to predict a dud at probability greater than 0.69 selects a single dud accurately with no false positives, with almost no change in the false negative rate, which increases only to 0.2. The false positives quickly ramp up as we relax the rule, but depending on KOMO's tolerance for false positives, and it's ability to quickly rewrite dud posts to improve them, a rule of somewhere between 0.5 and 0.69 seems optimal. Tweaking the rule won't change that this is a bad model, however!

5.4.2 Random Forest to predict duds

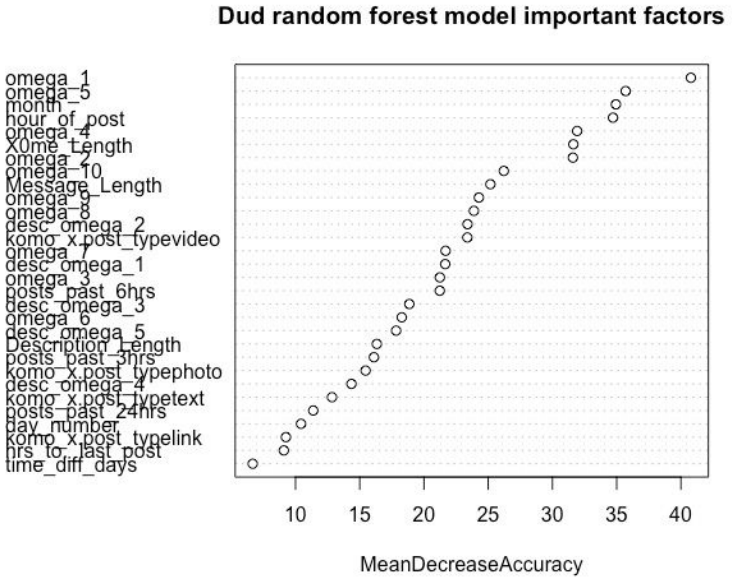
A 500 tree random forest predicts duds only slightly better, with an overall out of bag error rate of (1,277+151)/8,639 = 0.17 when we use the default setting of P > 0.5. This categorization rule results in a slightly lower false positive than the lasso, and the less important false negative is also improved, slightly:

	Pred 0	Pred 1	error
True 0	6761	151	0.2512 (false positive)
True 1	1277	450	0.159 (false negative)

Since the RF predicts larger categories better, we then weighted the duds higher in the algorithm to reflect the fact that they only made up 20% of the data points. This reduced overall performance to give an out of bag error rate of 0.19, but the false positive fell to 0.07. The increase in the false negative rate to 0.19 is probably acceptable, however. While this model predicts only 5% of duds, it does so at low cost and may still be a worthwhile incremental gain for KOMO if it can be implemented in a real time app to alert staff to duds before they post.

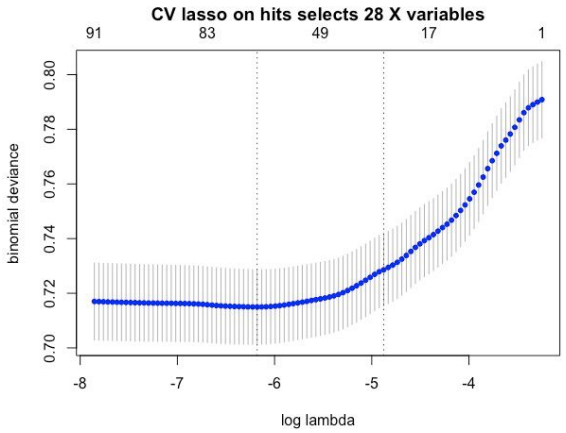
	Pred 0	Pred 1	error
True 0	6906	6	0.066 (false positive)
True 1	1642	85	0.192 (false negative)

The model finds most of the same factors important as the general prediction model previously developed, though the order of importance is slightly different:



5.4.3 Lasso to predict hits

Again, we were hampered here in fine tuning our models by the fact that we do not know the cost and revenue functions associated with hits. But if the goal with duds is to avoid false positives, the goal with hits surely leans in the opposite direction: to avoid false negatives. A 100-fold cross validated binomial lasso on duds selects 28 X variables, as the general model did in section 5.1. Using the default setting of P>0.5 to pick hits results in a model that only selects 4 hits, with an error rate of 1160/8639= 0.13.



	Pred 0	Pred 1	error
True 0	7475	0	0(false positive)
True 1	1160	4	0.134(false negative)

Clearly, we'd like to find more than 4 of the 1,164 viral hits there were in the year. While we can't pick the optimal trade off without knowing the cost and revenue functions, our intuition is that setting P>0.15 to pick hits looks like a better bet. This increases the overall error rate to

(469+2273)/8,639=0.32 and the false negative rate 0.77, but it seems acceptable to be wrong three times out of four if it helps KOMO accurately identify 695 viral hits a year (60% of all viral hits):

	Pred 0	Pred 1	error
True 0	5202	2273	0.766(false positive)
True 1	469	695	0.082(false negative)

5.4.4 Random Forests to predict hits

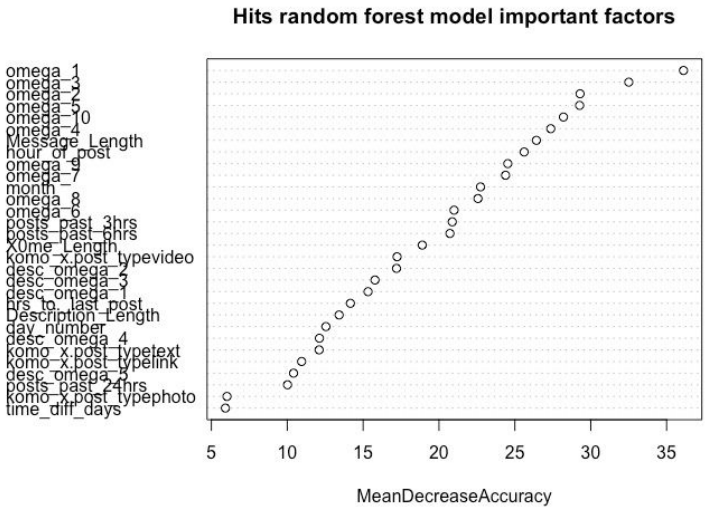
A 500 tree random forest predicts hits slightly better than the lasso if we use the default setting of P>0.5 to detect hits, with an identical overall error rate of (27+1,112)/8639 = 0.13. The false positive rate is higher than the lasso, though more hits are found. The false negative is similar to the lasso.

	Pred 0	Pred 1	error
True 0	7448	27	0.342 (false positive)
True 1	1112	52	0.130 (false negative)

As we did with duds, we re-weighted the hits in the random forest algorithm to reflect their lower incidence of 13.5%. This significantly reduced overall performance to give an error rate of 0.40, but the false negative fell to 0.07. The increase in the false positive rate to 0.79 may be acceptable, however. Though this model gets it wrong four times out five, it predicts 71% of all hits, a trade off that may be worthwhile.

	Pred 0	Pred 1	error
True 0	4387	3088	0.789(false positive)
True 1	337	827	0.071 (false negative)

Again, the model selects similar variables as important to the general model, in a slightly different order:

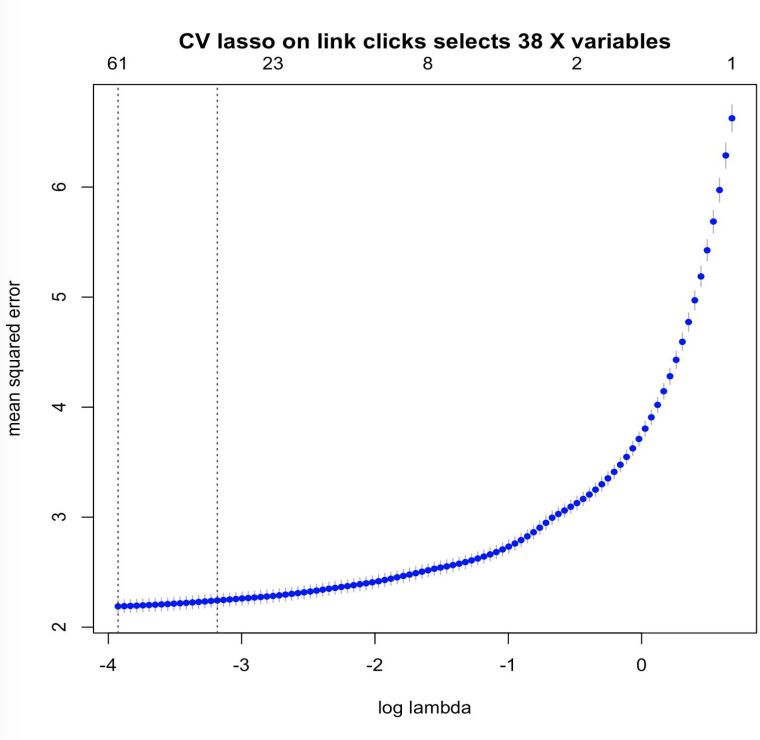

Hits random forest model important factors
omega_1
omega_3
omega_2
omega_5
omega_10
omega_4
Message_Length
hour_of_post
omega_9
omega_7
month
omega_8
omega_6
posts_past_3hrs
posts_past_6hrs
X0me_Length
komo_x.post_typevideo
desc_omega_2
desc_omega_3
desc_omega_1
hrs_to_last_post
Description_Length
day_number
desc_omega_4
komo_x.post_typetext
komo_x.post_typelink
desc_omega_5
posts_past_24hrs
komo_x.post_typephoto
time_diff_days
5
10
15
20
25
30
35
MeanDecreaseAccuracy

6. Modeling link clicks

Now we turned to link clicks, repeating the analysis we performed on log impressions on log link clicks for our subset of 6,606 posts.

6.1. CV Lasso

We ran a 100-fold cross validated lasso on log impressions, scaling all continuous variables. The lasso selected 38 explanatory variables. It had an out-of-sample R^2 of 0.66 — much better than we managed for log impressions.



The 38 selected variables, their coefficients and the interpretations of how they affect link clicks on the original scale are shown in the chart below.[21] While many of the variables follow what we saw when we looked at impressions, there is no place here for paying Facebook, which is surprising. And while videos increased impressions, here they reduce link clicks. This may be because users watch the video within Facebook's platform and do not click through to KOMO's website at the same rate as they do on text posts. Similarly, animal and weather stories decrease link clicks, despite increasing impressions (though the weather omega is likely confounded by the weather path variable).

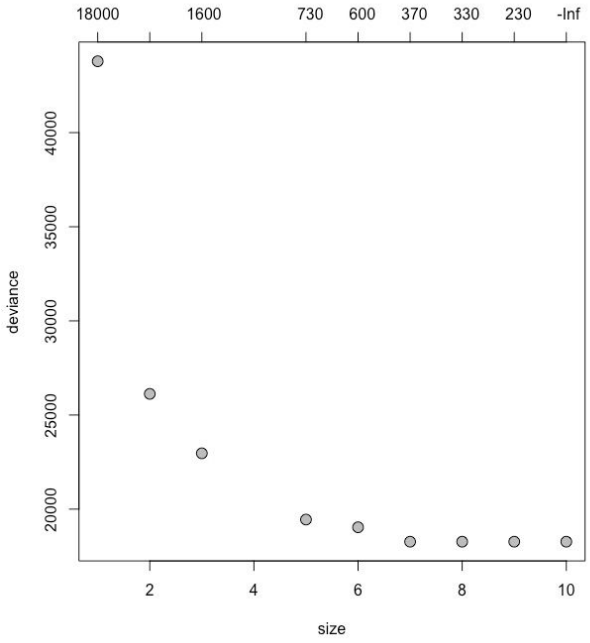
[21] The interpretation is referenced against an intercept of 6.03, which corresponds to exp(6.03)=415.72 link click such that the impact of an extra word in the headline (X0me_Length) is: exp(6.03+1(0.3482))-exp(6.03) or 173.15 additional clicks.

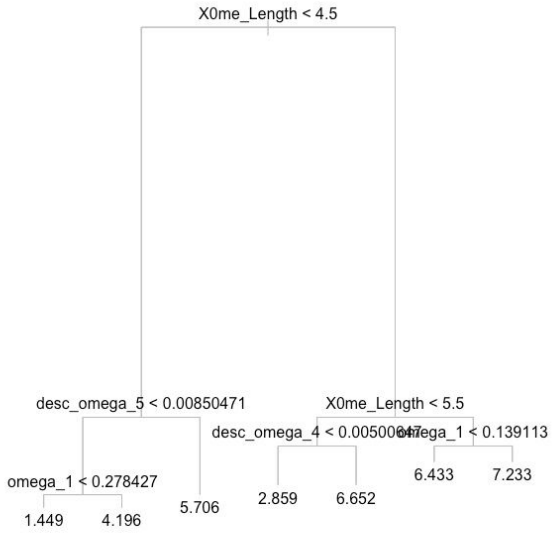
Variable	Coefficient	Interpretation
X0me_Length	0.3482	Adding a word to the headline adds 173 link clicks, on average
posts_past_3hrs	0.0146	An extra post in the last 3 hours adds 6 link clicks, on average
posts_past_6hrs	0.0049	An extra post in the last 6 hours adds 2 link clicks, on average
omega_1	0.2663	An extra standard deviation of loading on ‘crime’ adds 127 link clicks, on average
omega_2	-0.1315	An extra standard deviation of loading on ‘weather’ subtracts 51 link clicks, on average
omega_3	-0.1261	An extra standard deviation of loading on ‘social issues’ subtracts 49 link clicks, on average
omega_5	0.0982	An extra standard deviation of loading on ‘sports’ adds 42 link clicks, on average
omega_6	0.1338	An extra standard deviation of loading on ‘traffic’ adds 52 link clicks, on average
omega_7	0.0839	An extra standard deviation of loading on ‘human interest’ adds 36 link clicks, on average
omega_8	-0.0128	An extra standard deviation of loading on ‘celebrity’ subtracts 5 link clicks, on average
omega_9	0.1409	An extra standard deviation of loading on ‘schools’ adds 62 link clicks, on average
omega_10	-0.0895	An extra standard deviation of loading on ‘animals’ subtracts 36 link clicks, on average
X0me_omega_2	0.0639	An extra standard deviation of ‘crime’ in the headline adds 27 link clicks, on average
X0me_omega_5	-0.0177	An extra standard deviation of ‘industry’ in the headline subtracts 7 link clicks, on average
X0me_omega_7	0.0028	An extra standard deviation of ‘outdoors’ in the headline adds 1 link click, on average
X0me_omega_10	-0.0033	An extra standard deviation of the unclear Topic 10 in the headline subtracts 1 link click, on average
desc_omega_1	0.0202	An extra standard deviation of ‘crime’ in the sub headline adds 8 link clicks, on average
desc_omega_3	-0.0171	An extra standard deviation of ‘movies/sports’ in the sub headline subtracts 7 link clicks, on average
desc_omega_4	-0.0168	An extra standard deviation of ‘weather/environment’ in the sub headline subtracts 7 link clicks, on average
komo_x.post_typelink	0.1267	Posts without video result in an additional 56 link clicks
hour.f2	-0.1066	Posts in the 2am hour get 42 fewer link clicks, on average
hour.f5	-0.0274	Posts in the 5am hour get 11 fewer link clicks, on average
hour.f12	0.0895	Posts in the noon hour get an extra 39 link clicks, on average
hour.f19	0.0167	Posts in the 7pm hour get an extra 7 link clicks, on average
month.f1	0.2222	Posts in January got an extra 103 link clicks, on average
path_1news	0.0419	Posts routed to the “news” section of KOMO’s website got an extra 18 link clicks, on average
path_2local	0.1418	Posts routed to the “local” sub-section of KOMO’s website got 63 extra link clicks, on average
path_1KOMONews	-3.3431	Posts routed to the “KOMOnews” section of KOMO’s website got 401 link clicks less, on average
path_2photos	-1.3668	Posts routed to the “photos” subsection of KOMO’s website got 310 link clicks less, on average
path_1seattlerefined	0.4158	Posts routed to the “seattlerefined” section of KOMO’s website got an extra 214 link clicks, on average
path_2offbeat	0.0667	Posts routed to the “offbeat” subsection of KOMO’s website got 29 extra link clicks, on average

path_1weather	0.629	Posts routed to the “weather” section of KOMO’s website got an extra 364 link clicks, on average
path_2eat.drink	0.1193	Posts routed to the “eat.drink” subsection of KOMO’s website got 53 extra link clicks, on average
path_1home	0.4099	Posts routed to the “home” section of KOMO’s website got an extra 211 link clicks, on average
path_2tech	-0.2214	Posts routed to the “tech” subsection of KOMO’s website got 83 fewer link clicks, on average
path_2the.home	0.086	Posts routed to the “the.homel” subsection of KOMO’s website got 37 extra link clicks, on average
path_2travel	0.0091	Posts routed to the “travel” subsection of KOMO’s website got 4 extra link clicks, on average
path_1US	0.2115	Posts routed to the “US” section of KOMO’s website got an extra 98 link clicks, on average

6.2 CART Model

Next, we tried a tree model built using the CART algorithm to predict log link clicks. Setting the initial minimum node size to 1 and a minimum deviance requirement of 0.005 to proceed with a new split, then running a 100-fold cross validation resulted in a pruned tree with 7 nodes:[22]



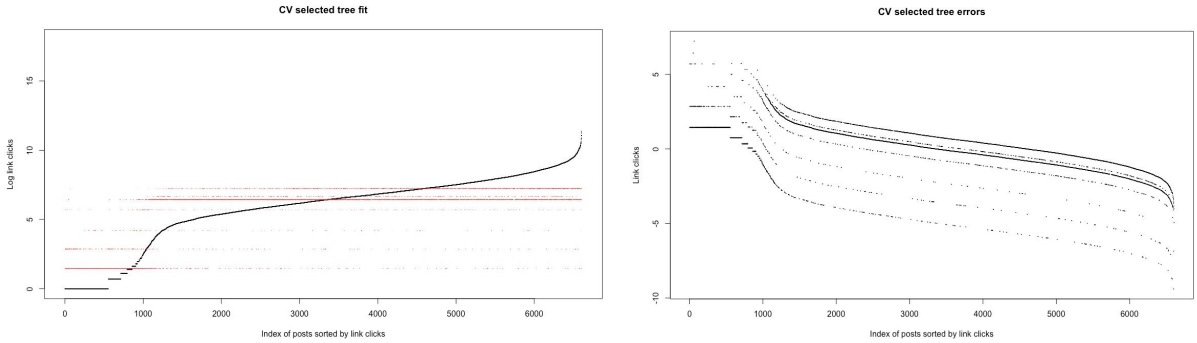


The model finds that the most important split is between posts with headlines of less than 4.5 words and those with more. Longer headlines in most cases result in more link clicks. The most link clicks come when there is a headline of greater than 5.5 words and a comparatively high loading on crime in the message. Posts with long headlines and a lower loading on crime also do well, though not as well as posts with medium length headlines and a lot of sports in the message. Medium length headlines with less sports do poorly. Among posts with short headlines, whimsy does best. Stories with short headlines and little whimsy do better

[22] As with impressions, we limited the tree and random forests analyses to 32 variables at a time.

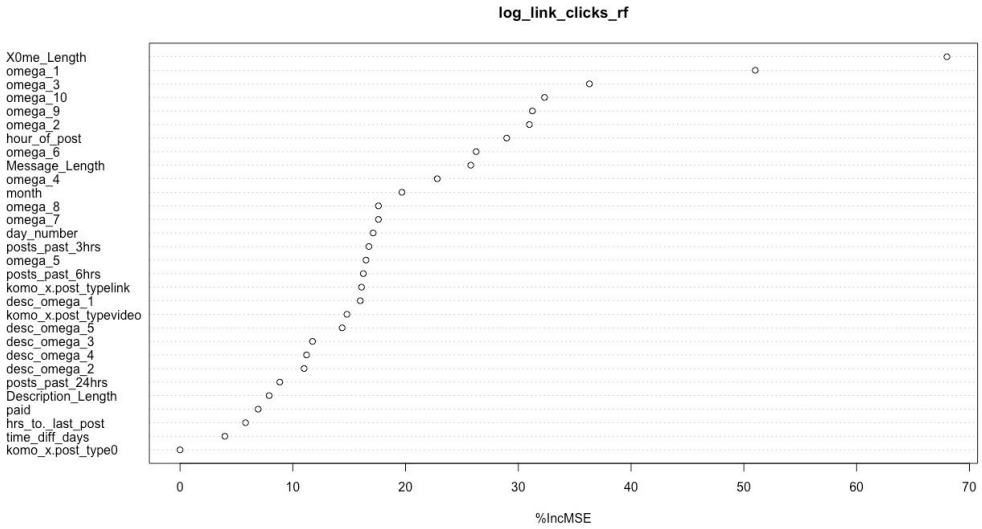
if they have a lot of crime in them. A short headline, little whimsy and little crime does worst of all. No other variables are chosen.

The plot below on the left shows the posts in rank order of link clicks in black, with the model's predictions in red. As with impressions, can see that the model significantly underpredicts how well the best posts do, and how badly the worst posts do. The plot on the right shows residuals, and demonstrates this bias toward the average more clearly:



6.3 Random Forest Model

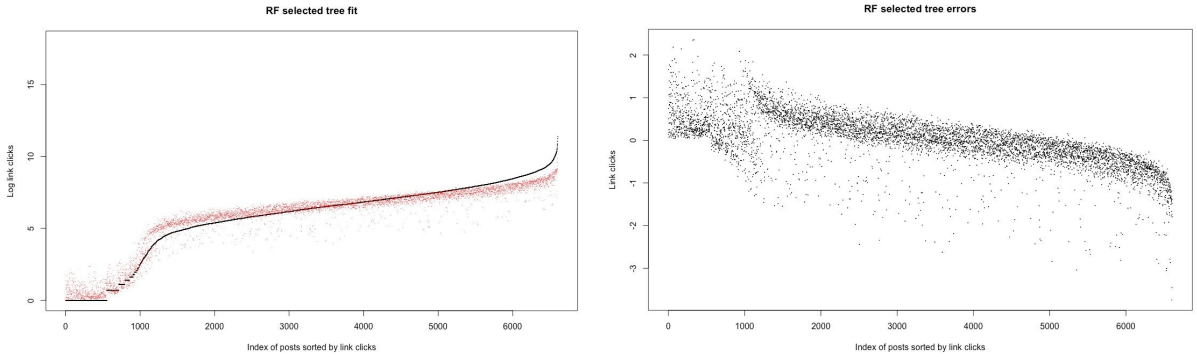
Again, we can improve the performance by switching to a random forest. A 500 tree forest finds the following factors important:



Headline length and the loadings on the various message topics are the most important variables when it comes to predicting link clicks, with 'crime' clearly the most important topic. Paying Facebook does feature but it is one of the least important variables.

The model fits much better than the CART model, especially across the middle of the range of performance, and has in-sample R^2 of 0.92, which again may indicate an overfit. Out of bag R^2 is 0.69, a hair better than the lasso model. Looking at the predicted link clicks (in red, left) against the ranked actual observations (in black, left), and at the residuals (right), we see

that, as with the impressions model the link clicks random forest model still underpredicts how well the best posts will do, and overpredicts how well the worst posts will do:

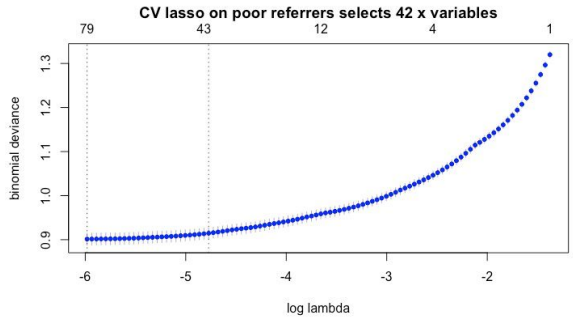


6.4 Categorical models to predict poor referrers

As we did with hits and duds, we wanted to try to build a categorical model to predict poor referrers, the 50% of posts that generate just 3.2% of link clicks. With the same caveats that applied to our attempts to build categorical models for impressions, we repeated the experiment with link clicks, using cross validated lasso regression and random forests.

6.4.1 Lasso to predict poor referrers

A 100-fold cross validated binomial regression to predict poor referrers has an error rate of 0.21 with the default rule of P>0.5. It selects 42 X variables:



	Pred 0	Pred 1	error
True 0	3911	239	0.157(false positive)
True 1	1170	1286	0.230(false negative)

In this case we want to minimize false positives, since we don't want to throw away good referrers by mistake. Once again, we can't choose exactly where to set P to maximize profitability without more information, but setting it up so that we only predict a poor referrer when P > 0.9 provides a conservative model with a false positive rate of just 0.03. True, we are only predicting one in five poor referrers, but there is little downside:

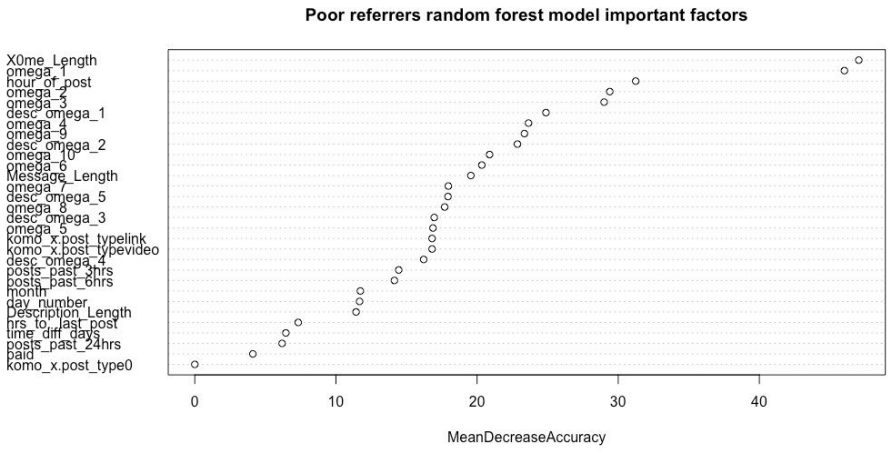
	Pred 0	Pred 1	error
True 0	4136	14	0.027(false positive)
True 1	1946	510	0.320(false negative)

6.4.2 Random Forest to predict poor referrers

A 500 tree random forest predicts poor referrers slightly better than the lasso if we use the default setting of P>0.5 to detect hits, with an identical overall error rate of 0.21, but the crucial false positive rate is higher than the lasso, though more hits are found. The false negative is similar to the lasso. Since this is a 50/50 sample of poor referrers versus good referrers, we cannot justify resampling. The random forest performs significantly worse than the tuned lasso regression.

	Pred 0	Pred 1	error
True 0	3885	265	0.225(false positive)
True 1	1127	1329	0.199(false negative)

Similar variables are important in this model to the lasso and the generalized CART model developed for link clicks:



7. Treatment effect of paying Facebook

Although we had a reasonable idea of many of the variables that drive impressions and link clicks at this point, we remained curious about the effect of paying Facebook, which we got conflicting signals about from our various models. Is paying Facebook worth it? We therefore performed a two-step lasso on log impressions to determine the treatment effect of payment.

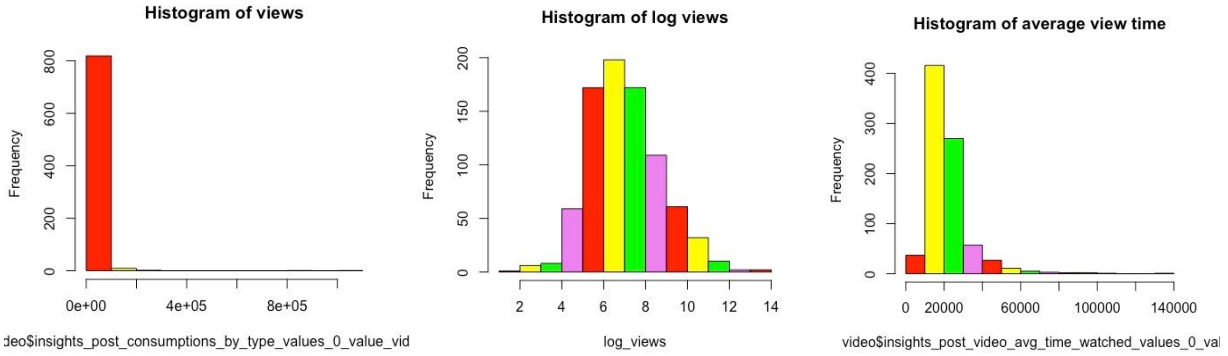
The first step lasso selected 42 of the other X values to predict payment, and had in sample R^2 of 0.88, confirming our suspicion that there was a large confounding effect. When we include dhat, the values of "paid" predicted by the other X variables, without penalty in a 100-fold cross-validated lasso that is otherwise identical to the one we ran in part 5.1, "paid" is no longer selected as significant.[23] With the very large caveat that we have only 24 paid datapoints to work with, we therefore find no evidence to reject the null hypothesis that paying Facebook makes no difference to impressions.

[23] Less punitive methods than cross validation allow "paid" to remain in the model, but we prefer the most conservative model and consistency with our regression in 5.1.

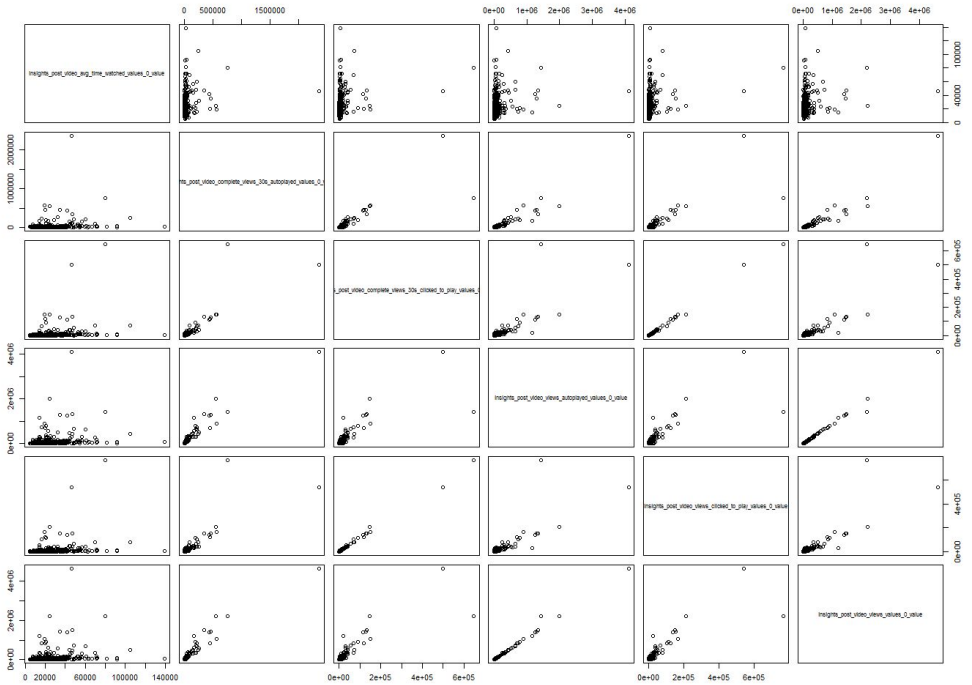
8. Initial Analysis of Potential Y Variables for Video

Now that we had a better understanding of what drives performance of posts in general, we wanted to delve deeper into the subset of 832 posts that contained video. We were interested in exploring any trade offs between increasing video views and increasing the average viewing time.

Like our previous performance variables, video views benefits from a log transformation, though we can manage without a transformation of average view time:



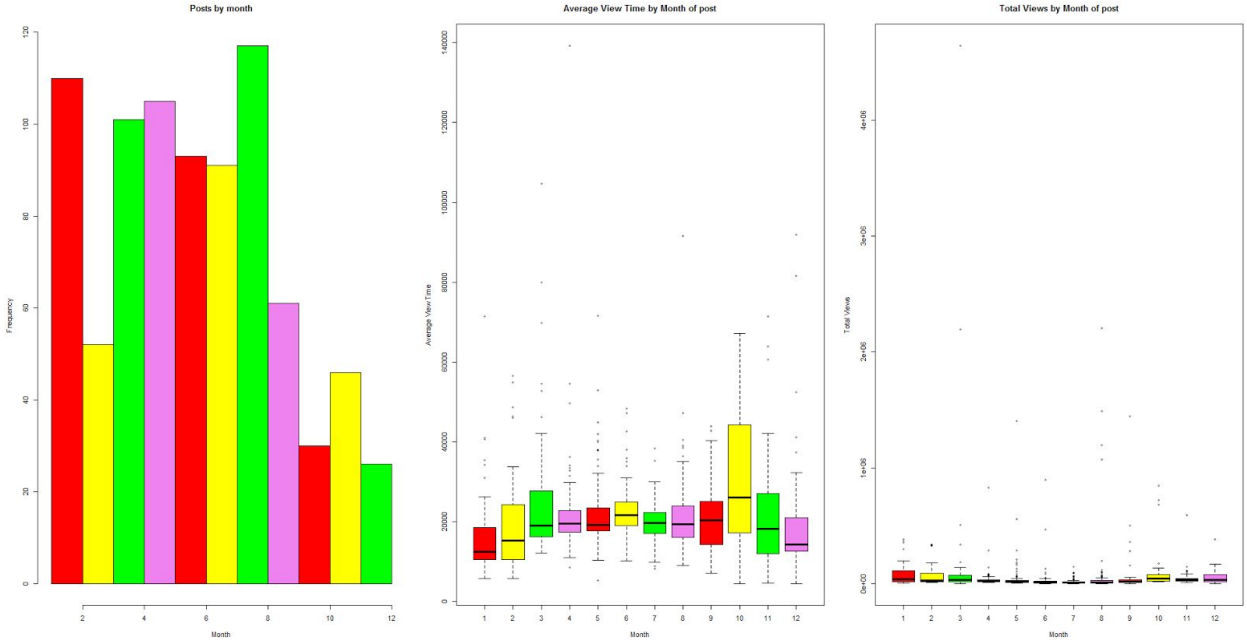
While the total number of views (bottom row) is highly correlated with other video performance metrics, average view time (top row) has no correlation with rest of the potential Y variables we considered:



9. Initial Analysis of X variables for Videos

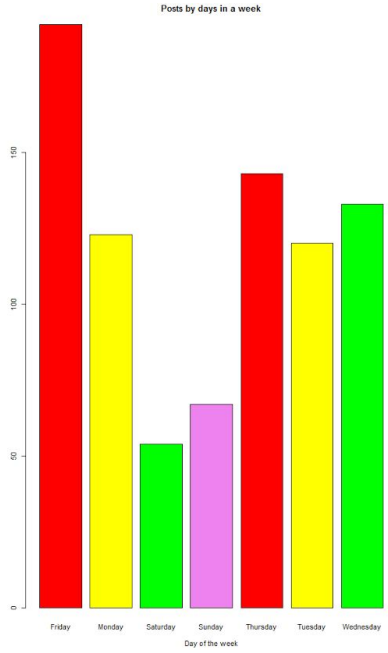
We repeated the exercise we undertook in part 4, but for video views and average viewing time. We wanted to see if charting revealed any association with our potential X variables.[24]

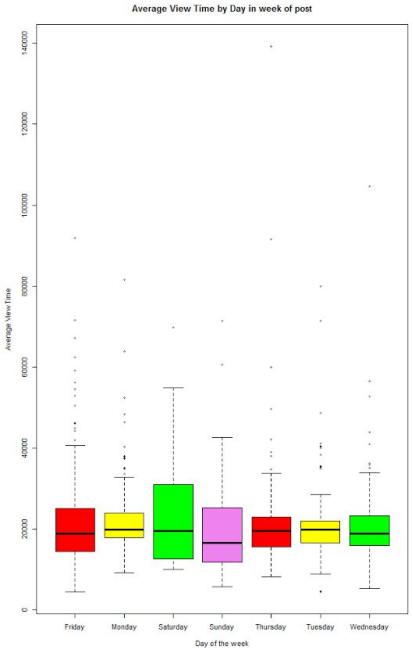
Months: KOMO-TV varied significantly in how frequently it posted from month to month: January, March, April and July all had lots of posts, but February, October and December had few. There was some variation from month to month in average view time and in views and there is one extreme outlier.

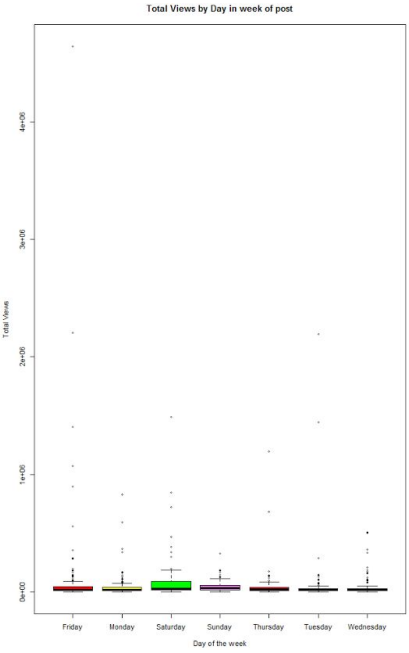


Days: Fridays and Thursdays saw the heaviest posting of videos, and Saturday the lightest. The day of the week seems to have little impact on viewing time or views.

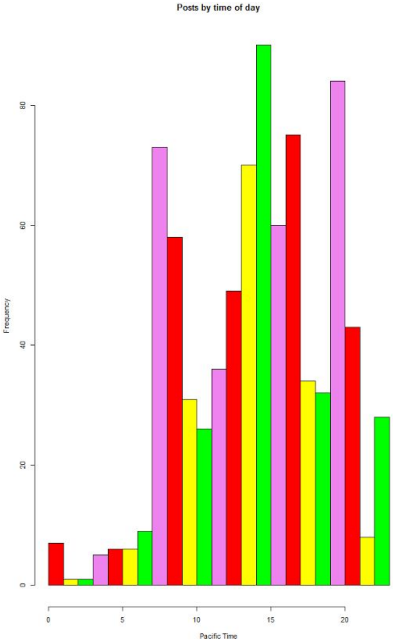
[24] In each of these analyses, the leftmost plot shows the distribution of the x variable, the center shows the x variable plotted against average view time and the rightmost shows the x variable plotted against total views.

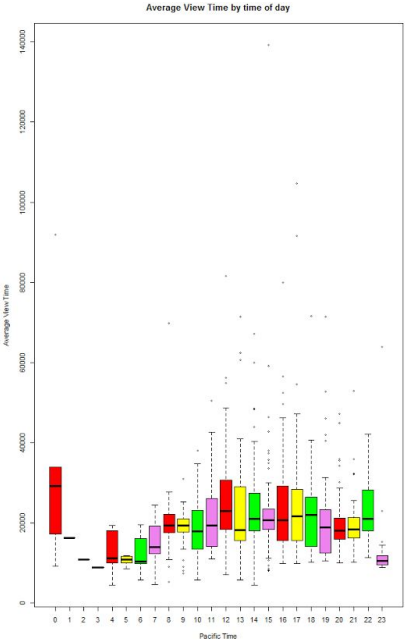


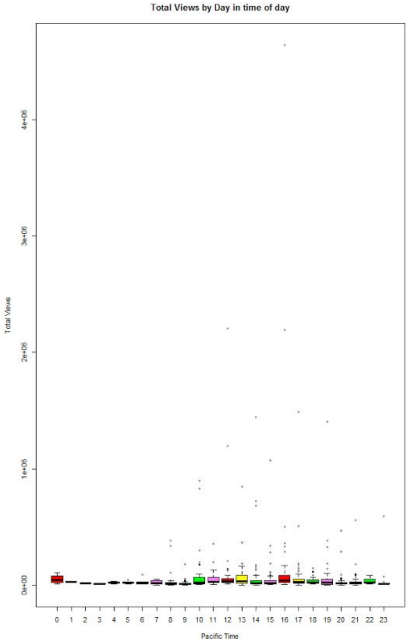




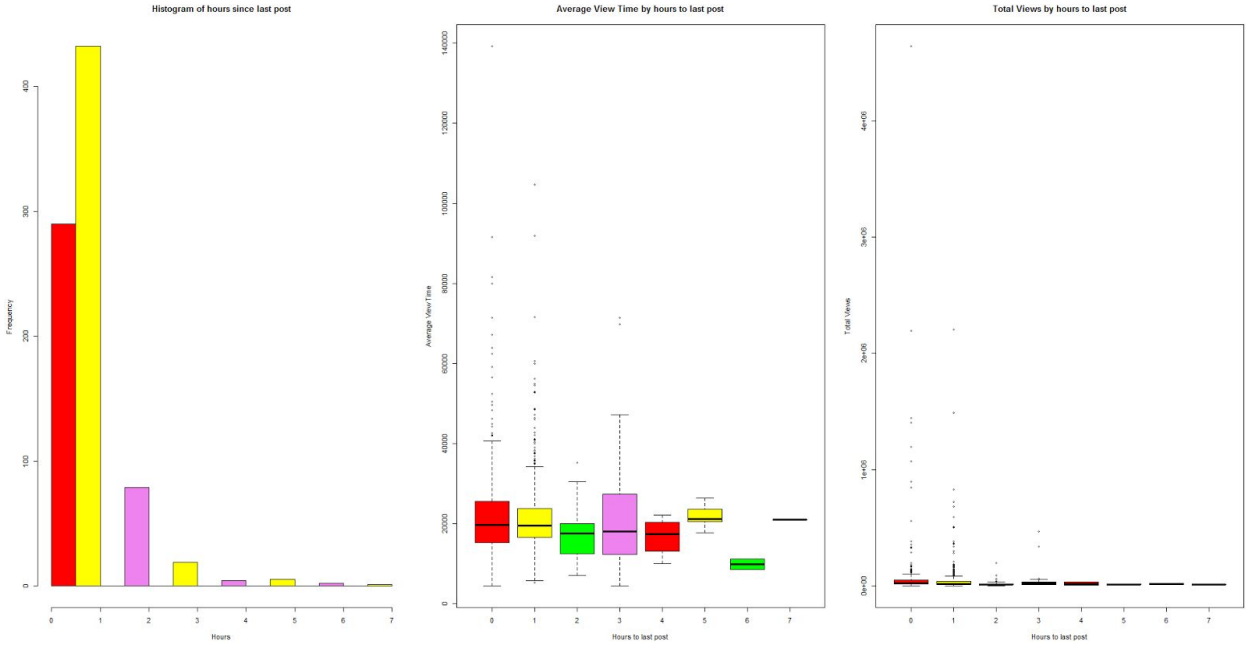
Hour: There are video posting peaks at 8am, 3pm and 8pm. Average viewing time seems to peak around midnight. Views are hard to read.



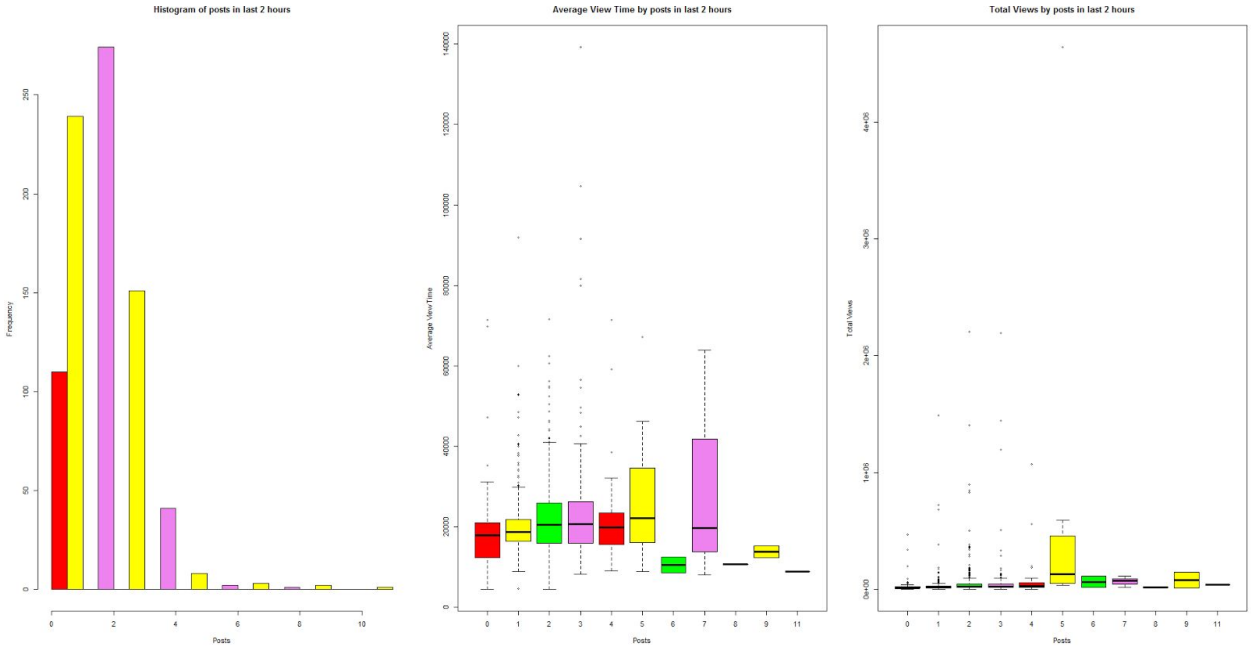




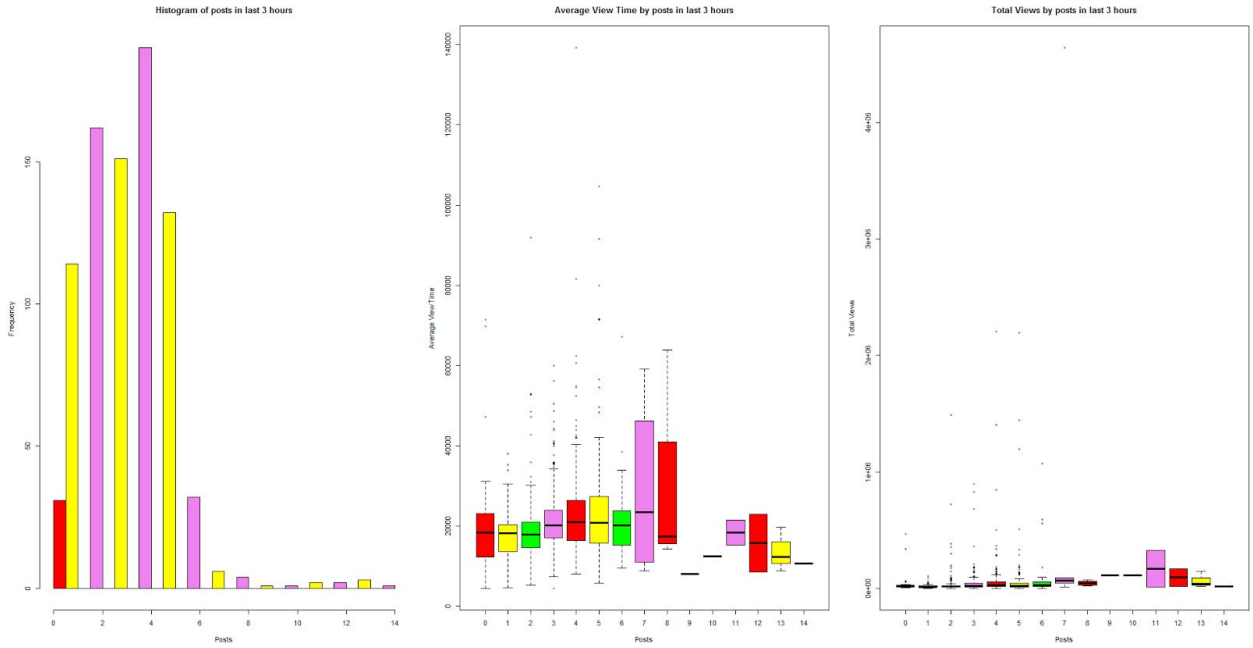
Hours to last post: Most videos are posted within an hour of another post, creating a heavy left skew to the distribution. We don't see any clear association here:



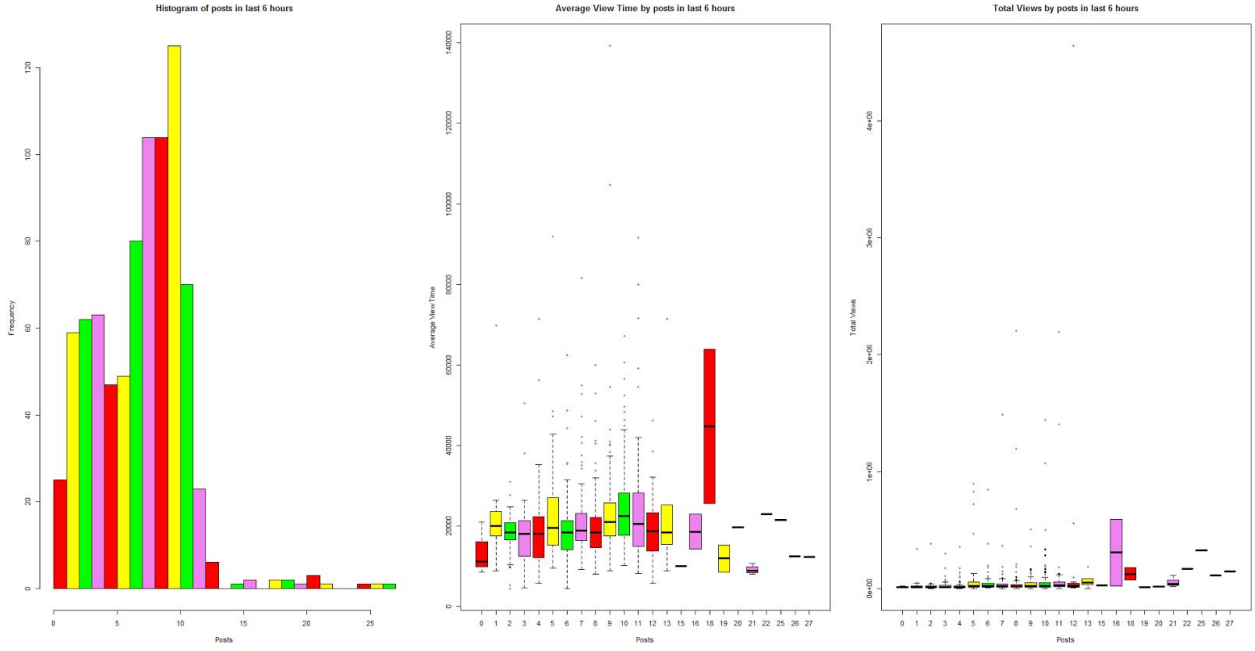
Posts in past 2 hours: Again, we see a left skew and no clear association:



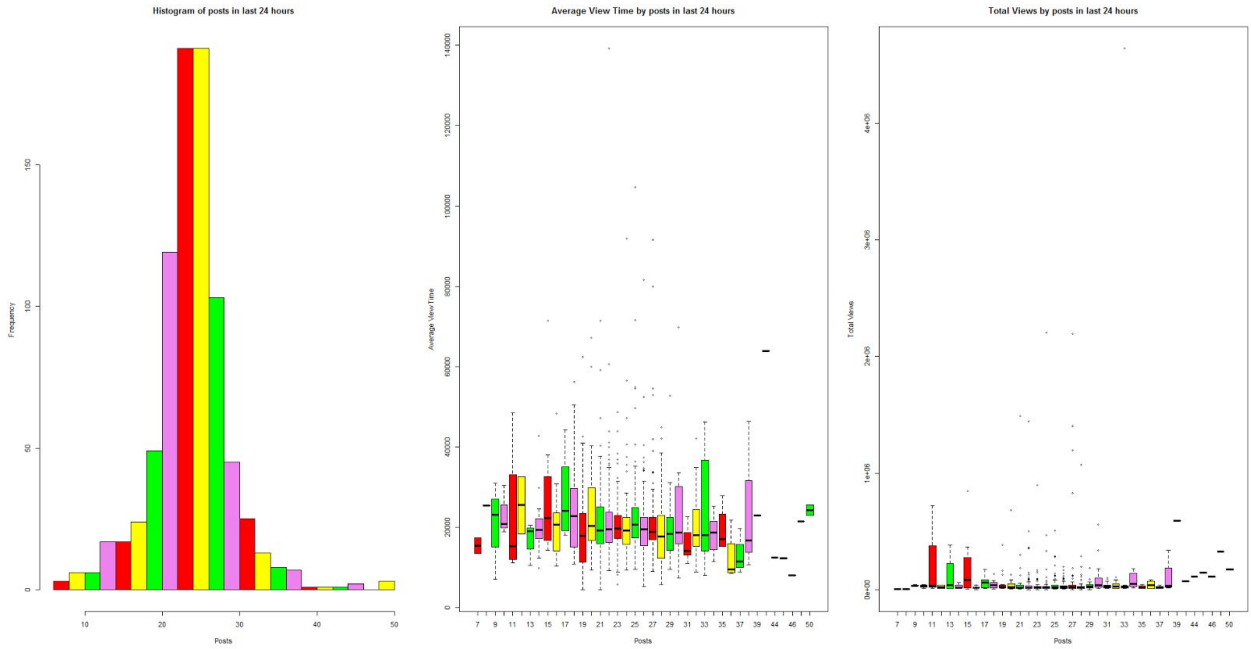
Posts in past 3 hours: This is a more normal distribution. Average viewing time appear to grow up to 7 posts over the past three hours, then decline. Views peak around 11 posts:



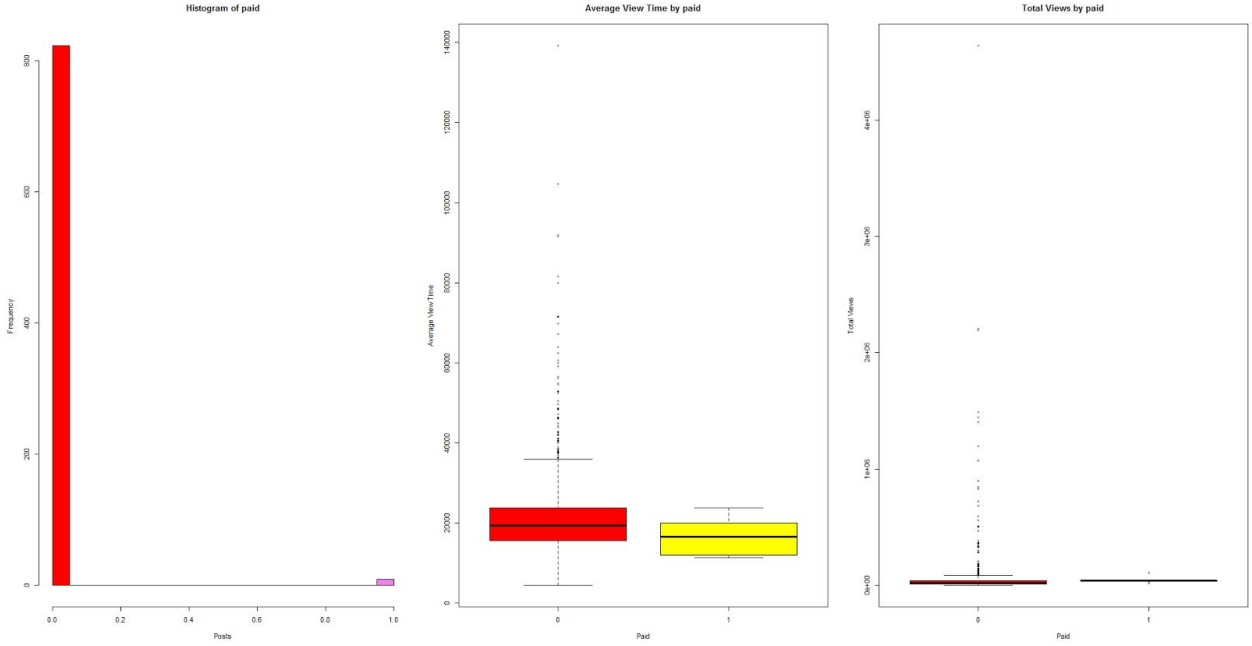
Posts in past 6 hours: Again we see average viewing time appear to grow up until around 10 posts:



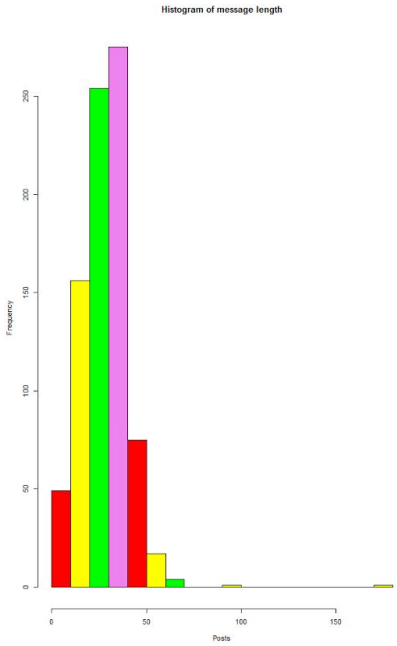
Posts in past 24 hours: There is no clear association:

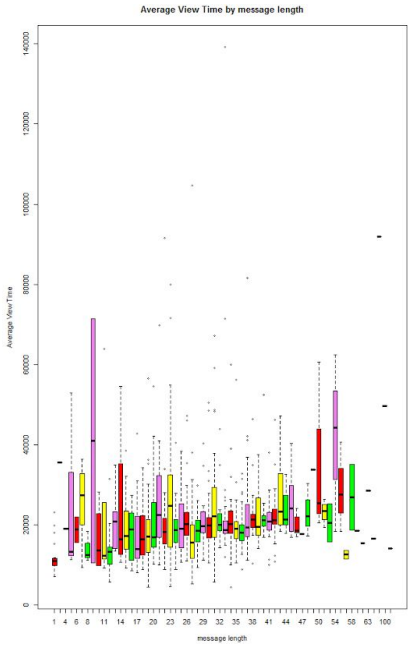


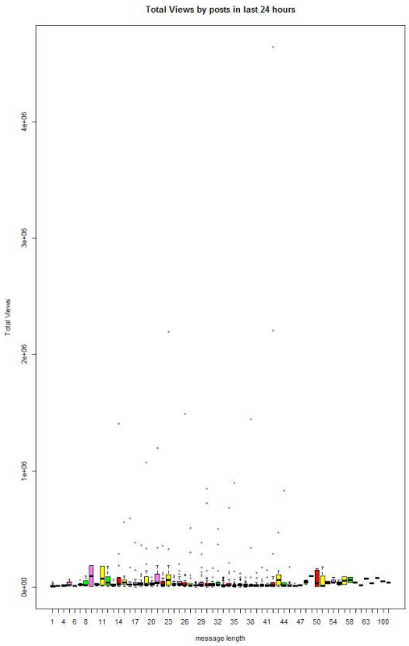
Paid: Again, there are only 12 videos that KOMO-TV paid Facebook to boost. While we previously found no statistically significant evidence that paying helped increase impressions, we do observe from the middle chart that paid videos are associated with a shorter average viewing time:



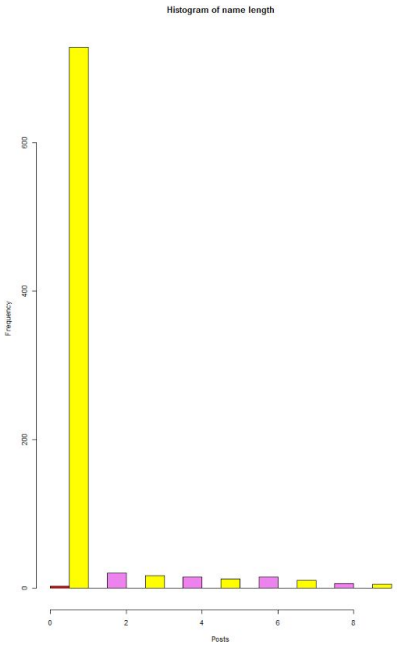
Message length: Longer messages appear to associate with longer average viewing times:

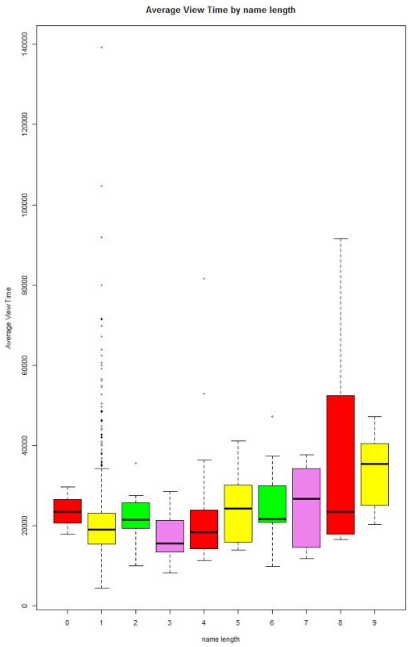


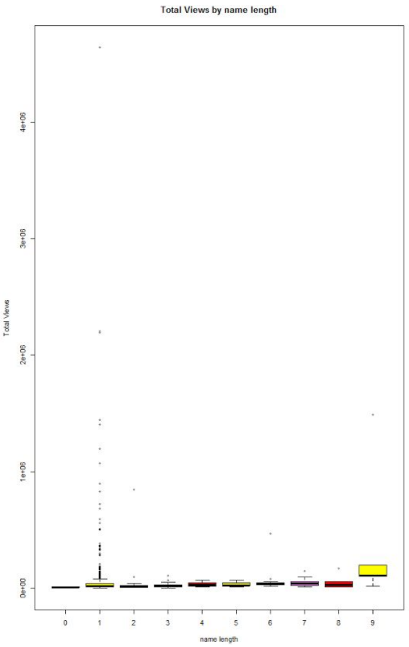




Name length: long headlines seem to have the same association:

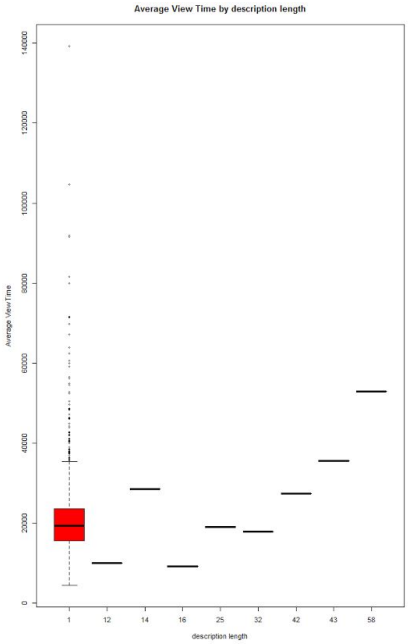


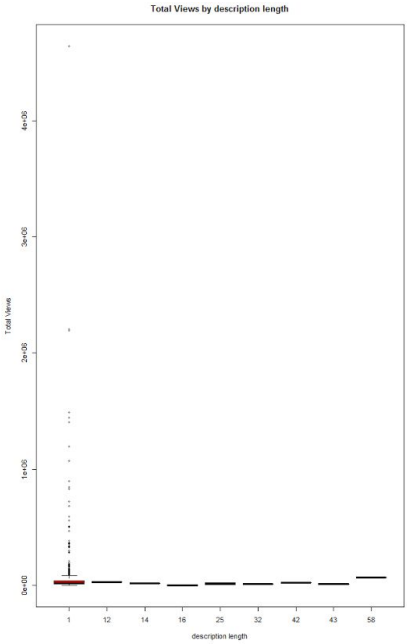




Description length: has remarkably few data points above 0, though those greater than zero again appear to associate with increased viewing time:


Histogram of description length
Frequency
Posts
0
200
400
600
800
0
10
20
30
40
50
60


Average View Time by description length
Average View Time
description length
0
20000
40000
60000
80000
100000
120000
140000
1
12
14
16
25
32
42
43
58

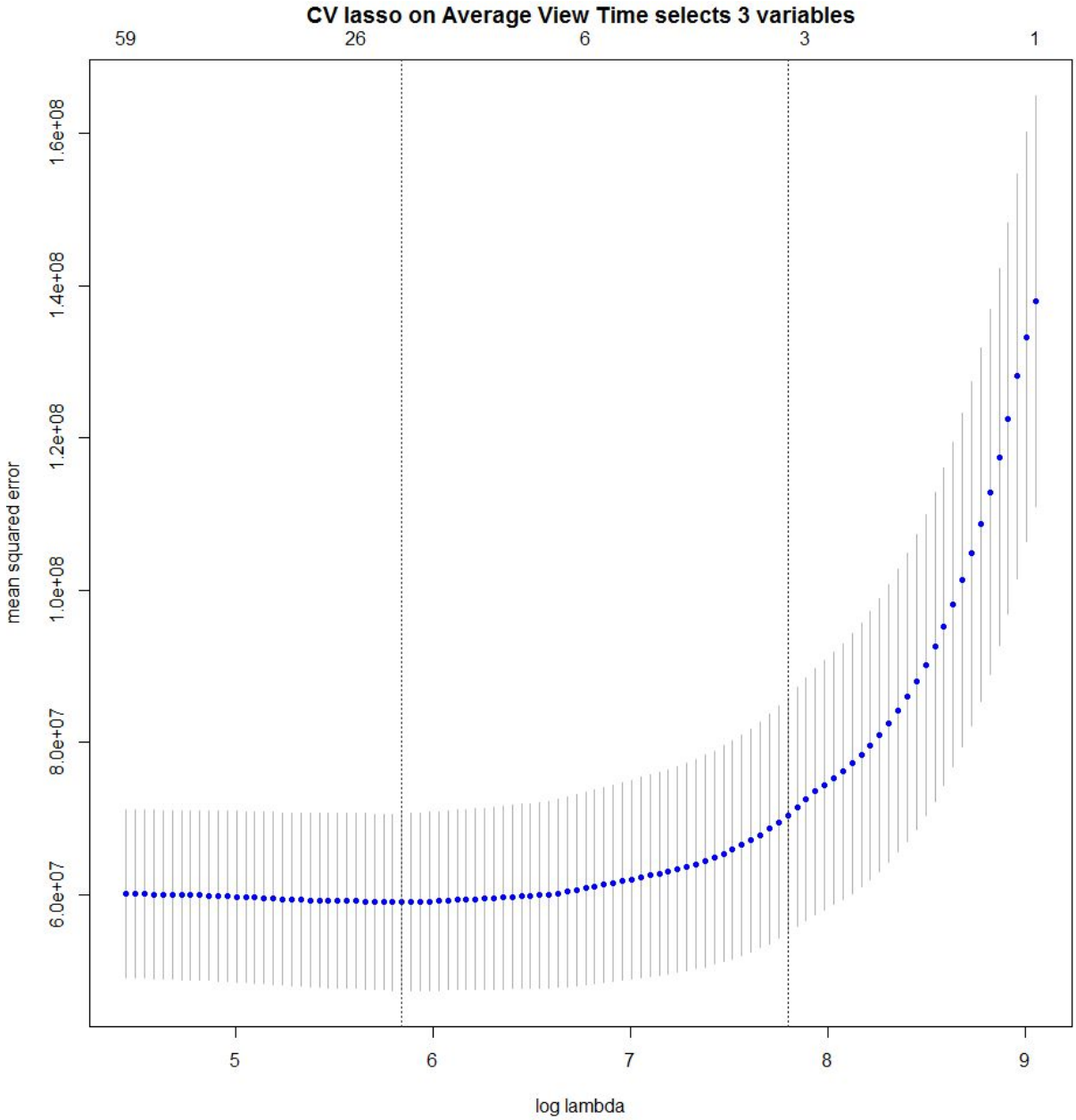

Total Views by description length
Total Views
description length
0e+00
1e+06
2e+06
3e+06
4e+06
1
12
14
16
25
32
42
43
58

10. Modeling Average View Time

We set out to predict average view time using cross-validated lasso regression, trees and random forest models. Our hypotheses were that each of the potential explanatory variables would have some impact on average view time.

10.1 CV Lasso

We ran a 100-fold cross validated lasso on average view time, scaling all continuous variables. The lasso selected 3 explanatory variables and rejected 80 variables. It had an out-of-sample R^2 of 0.57.

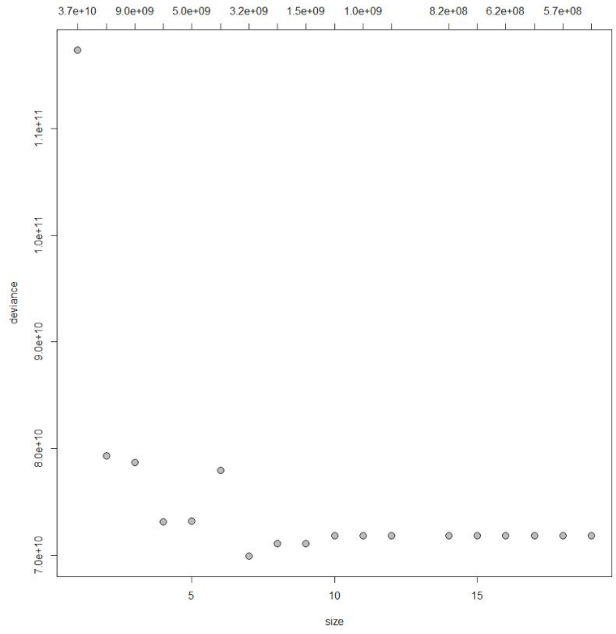


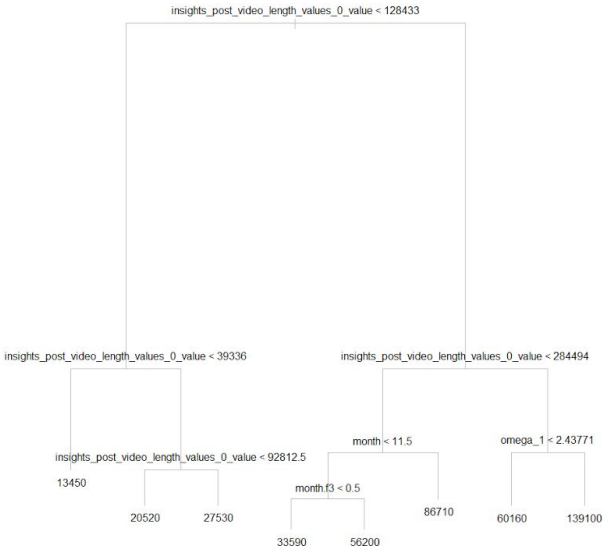
The two selected variables (the third is the intercept), their coefficients and the interpretations of how they affect average view time on the original scale are shown in the table below. It's unsurprising that the prime predictors are the length of the video and the loading on crime, as "if it bleeds it leads" is a timeworn local news mantra.

Variable	Coefficient	Interpretation
Video length	0.1024	Increasing the length of the video by 1 millisecond adds 0.1024 milliseconds average view time, on average
Message omega 1	541.76	Including a standard deviation more loading on "crime" in the message adds 541.76 milliseconds on average view time, on average

10.2 CART Model

Next, we tried a tree model built using the CART algorithm to predict average view time. Setting the initial minimum node size to 1 and a minimum deviance requirement of 0.005 to proceed with a new split, then running a 100-fold cross validation resulted in a pruned tree with 8 nodes:[25]



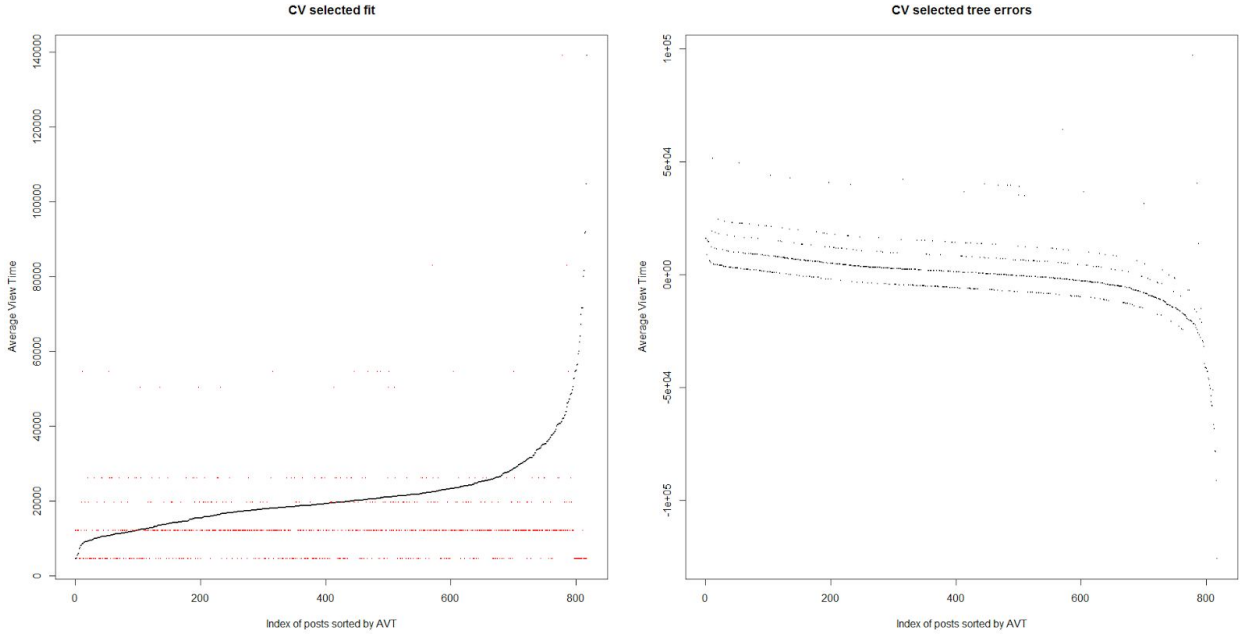


The model is dominated by video length and finds that a video of less than 128,443 milliseconds (2 minutes, 8 seconds) is the most important split and results in a lower average view time. Posts with a higher loading on crime and longer video do better. Posts with the shortest videos (less than 39,336 milliseconds) are predicted to perform worst of all. December is a bad month. No other variables are chosen.

The plot below on the left shows the videos in rank order of average view time in black, with the model's predictions in red. As with our previous CART models can see that the model significantly underpredicts how well the best videos do, though poor videos seem reasonably

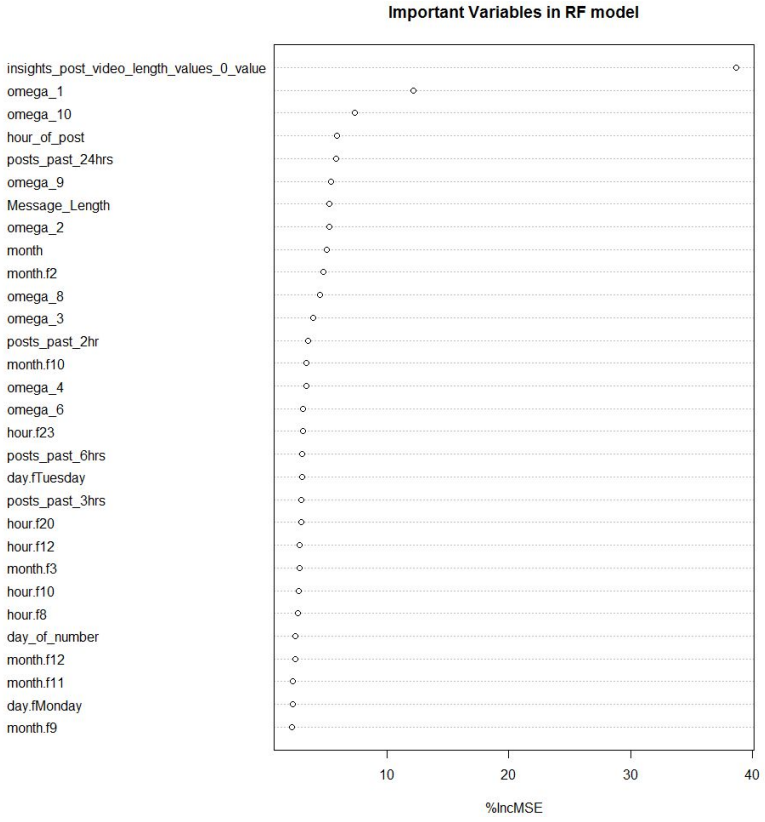
[25] As with the wider analysis, we again limited ourselves to the same set of 32 variables at a time, this time including video length, to cope with the 32 variable limit in CART.

well predicted. The plot on the right shows residuals, and demonstrates this bias toward the average more clearly:



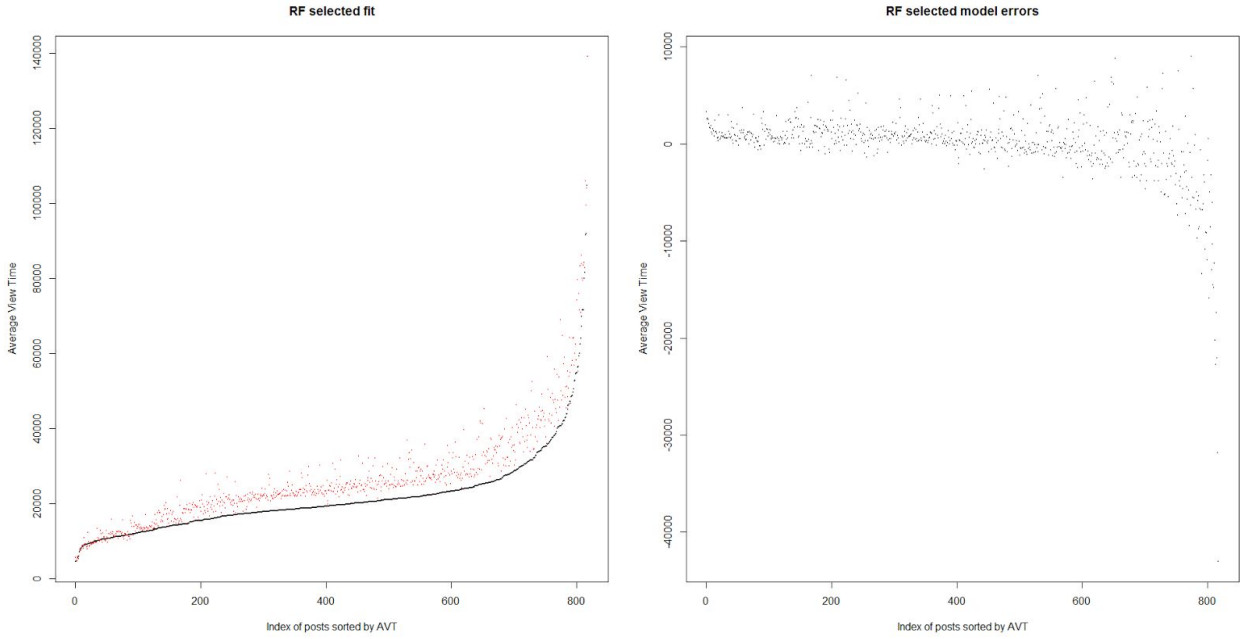
10.3 Random Forest Model

Again, we can improve on the tree model with a random forest. A 500 tree forest yields the following important factors:



The length of the video is by far the most important predictor. The loading on “crime” in the message, the loading on ‘animals’, the time of day the post, posts within 24 hours, the loading on ‘schools’, message length, and the loading on ‘weather’ are the other top eight predictors of length of view. The random forest does not find paying Facebook to be a useful predictor of average view time.

The model fits better than the CART model, especially across the middle of the range of performance, and has in-sample R^2 of 0.87 (almost 1.5 times better), which may indicate an overfit. Out of bag R^2 is 0.51, similar to the lasso model. Looking at the predicted view times (in red, left) against the ranked actual observations (in black, left), and at the residuals (right), we see that the model tends to overpredict how well most posts will do across the middle of the range, but that the performance of the very best videos is underpredicted:

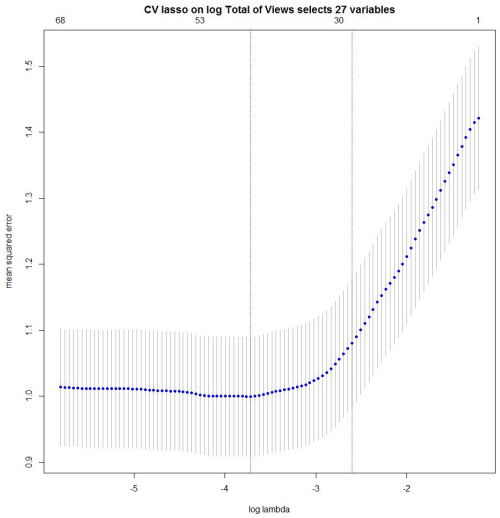


11. Modeling Total Views

Now we turned to video views, repeating the analysis we performed with average view time to log total of views for video posts.

11.1 CV Lasso

We ran a 100-fold cross validated lasso on log total number of views, scaling all continuous variables. The lasso selected 27 explanatory variables and rejected 56 variables. It had an out-of-sample R^2 of 0.3.



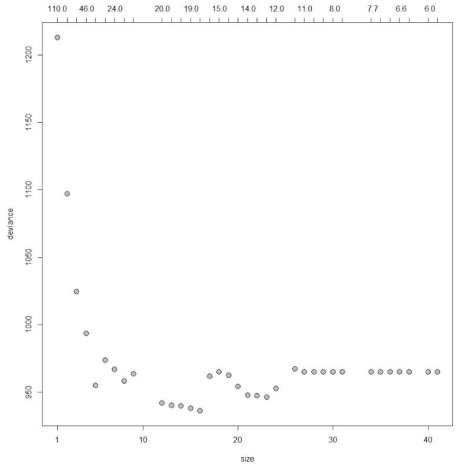
The 26 selected variables (excluding the intercept), their coefficients and the interpretations of how they affect total views on the original scale are shown in the chart below. Longer headlines and almost all the omega variables except celebrity increase views, as does posting more frequently, posting longer videos and posting in January and in March. Saturdays do better than other days of the week. We were surprised that longer videos attracted more views, as well as longer viewing time.

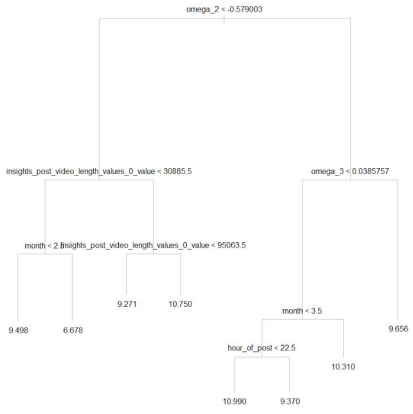
Variable	Coefficient	Interpretation
X0me_Length	0.0221	Adding a word to the headline adds 543 views, on average
post_past_2hrs	0.0671	An extra post in past 2 hours adds 1,688 views, on average
post_pas_3hrs	0.0446	An extra post in past 3 hours' adds 1,109 views on average
post_past_6hrs	6.608194e-13	An extra post in past 6 hours' adds almost no additional views, on average
omega_1	0.1628	An extra standard deviation of loading on' crims' adds 4,299 views, on average
omega_2	0.0206	An extra standard deviation of loading on 'weather' adds 506 views, on average
omega_3	-0.0515	An extra standard deviation of loading on 'social issues' subtracts 1,220 views, on average
omega_4	0.0383	An extra standard deviation of loading on 'sports' addts 949 views, on average

omega_6	0.0628	An extra standard deviation of loading on ‘traffic’ adds 1,576 views, on average
omega_7	0.1838	An extra standard deviation of loading on ‘human interest’ adds 4,907 views, on average
omega_8	-0.0346	An extra standard deviation of loading on ‘celebrity’ subtracts 827 views, on average
omega_9	0.0937	An extra standard deviation of loading on ‘schools’ adds 2,388 views, on average
insights_post_video_length_values_0_value	4.91915e-07	An extra millisecond in the video length adds 0.012 views, on average
X0me_omega_7	0.0111	An extra standard deviation of ‘outdoors’ in the headline adds 271 views, on average
hour.f9	-0.0080	Posts in the 9am hour get 193 fewer views, on average
hour.f12	0.1688	Posts in the noon hour get an extra 4,471 views, on average
hour.f13	0.0441	Posts in the 1pm hour get an extra 1,096 views, on average
hour_f16	0.1834	Posts in the 4pm hour get an extra 4,895 views, on average
hour.f23	-0.0039	Posts in the 11pm hour get an reduced 94 views, on average
month.f1	0.2684	Posts in January get an extra 7,487 views, on average
month.f3	0.0603	Posts in March get an extra 1,511 views, on average
month.f6	-0.3155	Posts in June get 6,580 fewer views, on average
month.f7	-0.4365	Posts in July get an reduced 8,601 views, on average
month.f8	-0.3582	Posts in August get an reduced 7,321 views, on average
month.f9	-0.2310	Posts in September get an reduced 5,016 views, on average
day.fSaturday	0.2707	Posts on Saturday got an extra 7,560 views, on average

11.2 CART Model

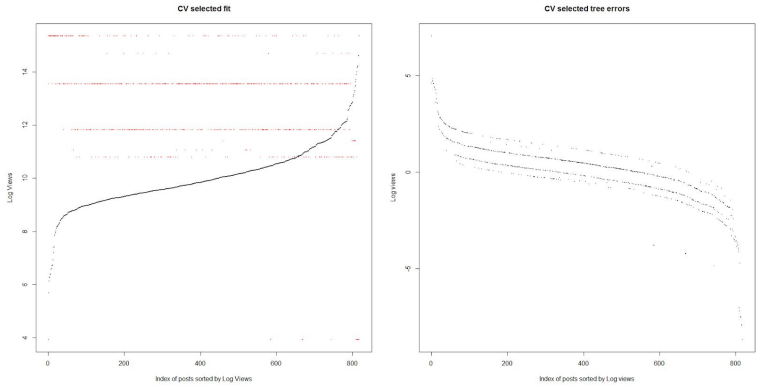
Next, we tried a tree model built using the CART algorithm to predict log total views. Setting the initial minimum node size to 1 and a minimum deviance requirement of 0.005 to proceed with a new split, then running a 100-fold cross validation resulted in a pruned tree with 8 nodes:





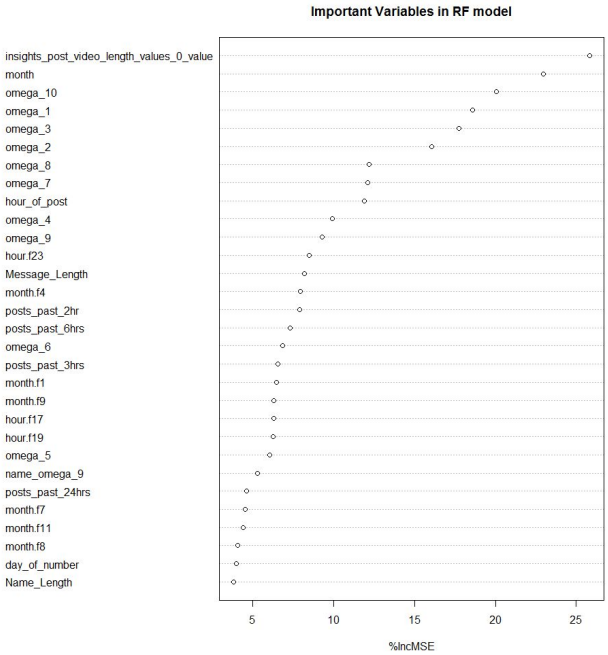
The model finds that a video post with lower loading of ‘weather’ (less than -0.58) is the most important split and results in a lower number of views, unless it includes a long video with a length of over 95,063 milliseconds. Posts with more loading of ‘weather’(more than -0.58) and posted after March do better than posts earlier in the year that are posted very late at night, but posts with more loading of ‘weather’ that are posted in the first three months of the year and earlier than 10:30 p.m. are predicted to perform best of all. No other variables are chosen.

The plot below on the left shows the posts in rank order of log total views in black, with the model’s predictions in red. The plot on the right shows residuals, and demonstrates a bias toward the average that we previously saw in our other tree models:

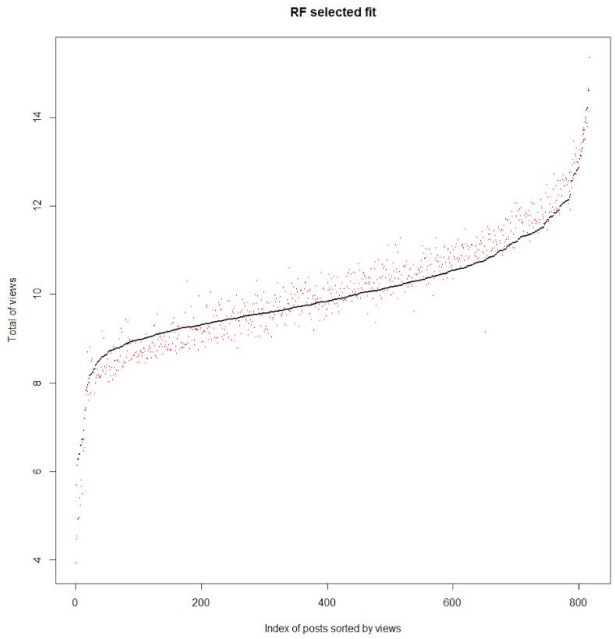


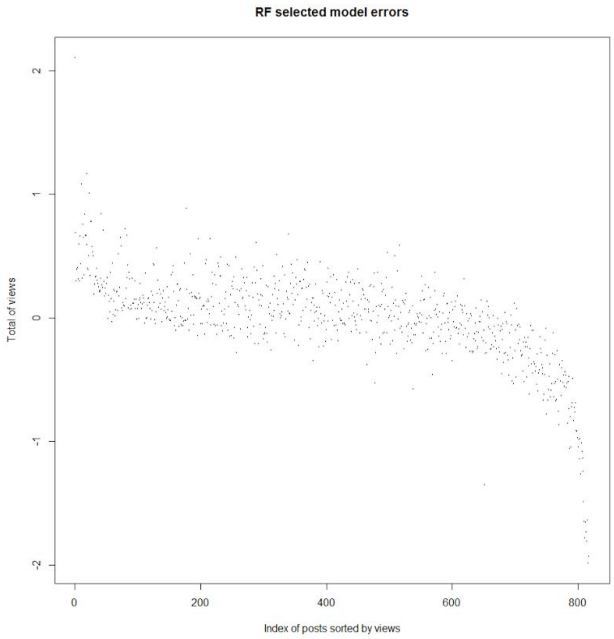
11.3 Random Forest Model

A 500 tree forest yields the following important factors. Length again reveals itself to be a key predictor of views, followed by the time of year at the content. Animal videos seem important, though the Lasso model did not say so.



The model fits much better than the CART model, especially across the end of the range of performance, and has in-sample R^2 of 0.85, which might suggest an over-fit. Out of bag R^2 is 0.43. Looking at the predicted impressions (in red, left) against the ranked actual observations (in black, left), and at the residuals (right), we see that the model still underpredicts how well the best posts will do and overpredicts how well the worst posts will do: :





12. Next Steps

While we feel that, given the constraints of time and the dataset, we have made significant progress in predicting and understanding the drivers of performance, several obvious avenues of potential further study remain open:

- An obvious flaw in our models is the inability to predict and incorporate “trending” topics. We can conceive of a model that uses the performance and content of posts in the hours immediately preceding the current post to help predict the performance of the current post. A constraint in the current dataset that would make such a model hard to build is that all performance variables are given for the lifetime of the post, whereas a real-time model would have to use real-time performance data at the time of posting. Another alternative would be to add an additional dataset containing, for example, Seattle trends data to our current dataset.
- A simpler next step would be to add interaction, squared and cubed effects to each of our Lasso models. While a full set of integration effects would blow up the number of potential x variables to the point where $p > n$, understanding the interaction between topic and time, for example, could both improve model performance and help develop business insights.
- Our dataset does not include demographic information. Adding demographics might allow for a network analysis, improve prediction and help us to develop additional insights.
- We also have many unexplored performance metrics in the data, especially in the video data that could be interesting, including, in particular, negative engagement.

13. Appendix

Sample untransformed file structure, as supplied:

{"message": "Do you think the Confederate flag should be banned in schools?",

"link": "http://www.komonews.com/news/national/Students-suspended-for-clothing-displaying-Confederate-flag-328092611.html",

"link_to_post": "http://facebook.com/114431401958534_940695755998757",

"name": "Students suspended for clothing displaying Confederate flag",

"insights":

{"post_video_retention_graph_clicked_to_play": {"title": "Lifetime Percentage of viewers clicked-to-play at each interval (video post)",

"name": "post_video_retention_graph_clicked_to_play",

"description": "Lifetime: Clicked-to-play views of your video at each moment as a percentage of all views, including views shorter than 3 seconds. (video post) (Total Count)",

"id": "114431401958534_940695755998757/insights/post_video_retention_graph_clicked_to_play/lifetime",

"period": "lifetime",

"values": [{"value": {}}]},

"post_negative_feedback_by_type_unique": {"title": "Lifetime Negative feedback",

"name": "post_negative_feedback_by_type_unique",

"description": "Lifetime: The number of people who have given negative feedback to your post, by type. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_negative_feedback_by_type_unique/lifetime",

"period": "lifetime",

"values": [{"value": {"hide_clicks": 2}}]},

"post_video_complete_views_paid_unique": {"title": "Lifetime Paid views to 95%",

"name": "post_video_complete_views_paid_unique",

"description": "Lifetime: Number of times your video was viewed to 95% of its length after paid promotion. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_video_complete_views_paid_unique/lifetime",

"period": "lifetime", "values": [{"value": 0}]},

"post_video_complete_views_organic": {"title": "Lifetime Organic views to 95%",

"name": "post_video_complete_views_organic",

"description": "Lifetime: Number of times your video was viewed to 95% of its length without any paid promotion. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_video_complete_views_organic/lifetime",

"period": "lifetime", "values": [{"value": 0}]},

"post_video_views_unique": {"title": "Lifetime Unique Video Views",

"name": "post_video_views_unique",

"description": "Lifetime: Number of unique people who viewed your video for more than 3 seconds. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_video_views_unique/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_video_views_organic_unique": {"title": "Lifetime Organic Video Views",

"name": "post_video_views_organic_unique",

```
                "description": "Lifetime: Number of times your video was viewed for more than 3 seconds without any paid promotion. (Unique Users)",
                "id": "114431401958534_940695755998757/insights/post_video_views_organic_unique/lifetime",
                "period": "lifetime",
                "values": [{"value": 0}]},
        "post_consumptions_by_type_unique": {"title": "Lifetime Post consumers by type",
                "name": "post_consumptions_by_type_unique",
                "description": "Lifetime: The number of people who clicked anywhere in your post, by type. (Unique Users)",
                "id": "114431401958534_940695755998757/insights/post_consumptions_by_type_unique/lifetime",
                "period": "lifetime",
                "values": [{"value": {"other clicks": 91, "link clicks": 19}}]},
        "post_impressions_viral_unique": {"title": "Lifetime Post viral reach",
                "name": "post_impressions_viral_unique",
                "description": "Lifetime: The number of people who saw your Page post in a story from a friend. (Unique Users)",
                "id": "114431401958534_940695755998757/insights/post_impressions_viral_unique/lifetime",
                "period": "lifetime",
                "values": [{"value": 22}]},
        "post_video_complete_views_30s_paid": {"title": "Lifetime Paid 30-Second Views",
                "name": "post_video_complete_views_30s_paid",
                "description": "Lifetime: Number of times your video was viewed for 30 seconds or viewed to the end, whichever came first, after a paid promotion.
(Total Count)",
                "id": "114431401958534_940695755998757/insights/post_video_complete_views_30s_paid/lifetime",
                "period": "lifetime",
                "values": [{"value": 0}]},
        "post_negative_feedback_unique": {"title": "Lifetime Negative feedback",
                "name": "post_negative_feedback_unique",
                "description": "Lifetime: The number of people who have given negative feedback to your post. (Unique Users)",
                "id": "114431401958534_940695755998757/insights/post_negative_feedback_unique/lifetime",
                "period": "lifetime",
                "values": [{"value": 2}]},
        "post_storytellers_by_action_type": {"title": "Lifetime Talking About This (Post) by action type",
                "name": "post_storytellers_by_action_type",
                "description": "Lifetime: The number of unique people who created a story about your Page post by interacting with it. (Unique Users)",
                "id": "114431401958534_940695755998757/insights/post_storytellers_by_action_type/lifetime",
                "period": "lifetime",
                "values": [{"value": {"comment": 31, "like": 6, "share": 1}}]},
        "post_story_adds": {"title": "Lifetime Post Stories",
                "name": "post_story_adds",
                "description": "Lifetime: The number of stories generated about your Page post. (Total Count)",
                "id": "114431401958534_940695755998757/insights/post_story_adds/lifetime",
                "period": "lifetime",
```

```
                "values": [{"value": 42}]},
        "post_storytellers": {"title": "Lifetime Talking About This (Post)",
                "name": "post_storytellers",
                "description": "Lifetime: The number of unique people who created a story by interacting with your Page post. (Unique Users)",
                "id": "114431401958534_940695755998757/insights/post_storytellers/lifetime",
                "period": "lifetime",
                "values": [{"value": 38}]},
        "post_video_avg_time_watched": {"title": "Lifetime Average time video viewed",
                "name": "post_video_avg_time_watched",
                "description": "Lifetime: Average time video viewed (Total Count)",
                "id": "114431401958534_940695755998757/insights/post_video_avg_time_watched/lifetime",
                "period": "lifetime", "values": [{"value": 0}]},
                "post_video_length": {"title": "Lifetime Video length",
                "name": "post_video_length",
                "description": "Lifetime: Length of a video post (Total Count)",
                "id": "114431401958534_940695755998757/insights/post_video_length/lifetime",
                "period": "lifetime",
                "values": [{"value": 0}]},
        "post_story_adds_by_action_type_unique": {"title": "Lifetime Talking About This (Post) by action type",
                "name": "post_story_adds_by_action_type_unique",
                "description": "Lifetime: The number of unique people who created a story about your Page post by interacting with it. (Unique Users)",
                "id": "114431401958534_940695755998757/insights/post_story_adds_by_action_type_unique/lifetime",
                "period": "lifetime",
                "values": [{"value": {"comment": 31, "like": 6, "share": 1}}]},
        "post_negative_feedback": {"title": "Lifetime Negative Feedback from Users",
                "name": "post_negative_feedback",
                "description": "Lifetime: The number of times people have given negative feedback to your post. (Total Count)",
                "id": "114431401958534_940695755998757/insights/post_negative_feedback/lifetime",
                "period": "lifetime",
                "values": [{"value": 2}]},
        "post_video_retention_graph_autoplayed": {"title": "Lifetime Percentage of auto-played viewers at each interval (video post)",
                "name": "post_video_retention_graph_autoplayed",
                "description": "Lifetime: Auto-played views of your video at each moment as a percentage of all views, including views shorter than 3 seconds. (video post) (Total Count)",
                "id": "114431401958534_940695755998757/insights/post_video_retention_graph_autoplayed/lifetime",
                "period": "lifetime",
                "values": [{"value": {}}]},
        "post_video_complete_views_30s": {"title": "Lifetime Total 30-Second Views",
                "name": "post_video_complete_views_30s",
```

"description": "Lifetime: Total number of times your video was viewed for 30 seconds or viewed to the end, whichever came first. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_video_complete_views_30s/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_video_views_paid": {"title": "Lifetime Paid Video Views",

"name": "post_video_views_paid",

"description": "Lifetime: Number of times your video was viewed more than 3 seconds after paid promotion. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_video_views_paid/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_impressions_by_paid_non_paid_unique": {"title": "Lifetime Post impressions by paid and non-paid",

"name": "post_impressions_by_paid_non_paid_unique",

"description": "Lifetime: The number of impressions of your Page post broken down by paid and non-paid. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_impressions_by_paid_non_paid_unique/lifetime",

"period": "lifetime",

"values": [{"value": {"paid": 0, "total": 3583, "unpaid": 3583}}]},

"post_impressions_paid": {"title": "Lifetime Post Paid Impressions",

"name": "post_impressions_paid",

"description": "Lifetime: The number of impressions of your Page post in an Ad or Sponsored Story. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_impressions_paid/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_impressions_fan_paid_unique": {"title": "Lifetime Paid reach of a post by people who like your Page",

"name": "post_impressions_fan_paid_unique",

"description": "Lifetime: The number of people who like your Page and who saw your Page post in an ad or sponsored story. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_impressions_fan_paid_unique/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_engaged_users": {"title": "Lifetime Engaged Users",

"name": "post_engaged_users",

"description": "Lifetime: The number of people who clicked anywhere in your posts. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_engaged_users/lifetime",

"period": "lifetime",

"values": [{"value": 117}]},

"post_video_views_clicked_to_play": {"title": "Lifetime Clicked-to-Play Video Views",

"name": "post_video_views_clicked_to_play",

"description": "Lifetime: Number of times people clicked to play your video and viewed it more than 3 seconds. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_video_views_clicked_to_play/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_impressions_fan": {"title": "Lifetime Post Impressions by people who have liked your Page",

"name": "post_impressions_fan",

"description": "Lifetime: The number of impressions of your Page post to people who have liked your Page. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_impressions_fan/lifetime",

"period": "lifetime",

"values": [{"value": 4037}]},

"post_story_adds_unique": {"title": "Lifetime Talking About This (Post)",

"name": "post_story_adds_unique",

"description": "Lifetime: The number of unique people who created a story by interacting with your Page post. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_story_adds_unique/lifetime",

"period": "lifetime",

"values": [{"value": 38}]},

"post_impressions_by_story_type": {"title": "Lifetime Post Viral Impressions by story type",

"name": "post_impressions_by_story_type",

"description": "Lifetime: The number of times people saw this post via stories published by their friends. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_impressions_by_story_type/lifetime",

"period": "lifetime",

"values": [{"value": {"other": 30}}]}

"post_engaged_fan": {"title": "Lifetime People who have liked your Page and engaged with your post",

"name": "post_engaged_fan",

"description": "Lifetime: The number of people who have liked your Page and clicked anywhere in your posts. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_engaged_fan/lifetime","

"period": "lifetime",

"values": [{"value": 117}]},

"post_consumptions_by_type": {"title": "Lifetime Post Consumptions by type",

"name": "post_consumptions_by_type",

"description": "Lifetime: The number of clicks anywhere in your post, by type. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_consumptions_by_type/lifetime",

"period": "lifetime",

"values": [{"value": {"other clicks": 151, "link clicks": 19}}]},

"post_impressions_organic_unique": {"title": "Lifetime Post organic reach",

"name": "post_impressions_organic_unique",

"description": "Lifetime: The number of people who saw your Page post in news feed or ticker, or on your Page's timeline. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_impressions_organic_unique/lifetime",

"period": "lifetime", "values": [{"value": 3572}]},

"post_video_complete_views_30s_organic": {"title": "Lifetime Organic 30-Second Views",

"name": "post_video_complete_views_30s_organic",

"description": "Lifetime: Number of times your video was viewed for 30 seconds or viewed to the end, whichever came first, without a paid promotion. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_video_complete_views_30s_organic/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_impressions_paid_unique": {"title": "Lifetime Post Paid Reach",

"name": "post_impressions_paid_unique",

"description": "Lifetime: The number of people your advertised Page post was served to. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_impressions_paid_unique/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_impressions_unique": {"title": "Lifetime Post Total Reach",

"name": "post_impressions_unique",

"description": "Lifetime: The total number of people your Page post was served to. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_impressions_unique/lifetime",

"period": "lifetime",

"values": [{"value": 3583}]},

"post_video_retention_graph": {"title": "Lifetime Percentage of viewers at each interval (video post)",

“name": "post_video_retention_graph",

"description": "Lifetime: Views of your video at each moment as a percentage of all views, including views shorter than 3 seconds. (video post) (Total Count)",

"id": "114431401958534_940695755998757/insights/post_video_retention_graph/lifetime",

"period": "lifetime",

"values": [{"value": {}}]},

"post_video_complete_views_organic_unique": {"title": "Lifetime Organic views to 95%",

"name": "post_video_complete_views_organic_unique",

"description": "Lifetime: Number of times your video was viewed to 95% of its length without any paid promotion. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_video_complete_views_organic_unique/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_video_views_organic": {"title": "Lifetime Organic Video Views",

"name": "post_video_views_organic",

"description": "Lifetime: Number of times your video was viewed for more than 3 seconds without any paid promotion. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_video_views_organic/lifetime",

"period": "lifetime", "values": [{"value": 0}]},

"post_impressions_by_paid_non_paid": {"title": "Lifetime Post impressions by paid and non-paid",

"name": "post_impressions_by_paid_non_paid",

"description": "Lifetime: The number of impressions of your Page post broken down by paid and non-paid. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_impressions_by_paid_non_paid/lifetime",

"period": "lifetime",

"values": [{"value": {"paid": 0, "total": 4072, "unpaid": 4072}}]},

"post_negative_feedback_by_type": {"title": "Lifetime Negative Feedback from Users by Type",

"name": "post_negative_feedback_by_type",

"description": "Lifetime: The number of times people have given negative feedback to your post, by type. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_negative_feedback_by_type/lifetime",

"period": "lifetime",

"values": [{"value": {"hide_clicks": 2}}]},

"post_impressions_fan_paid": {"title": "Lifetime Post Paid Impressions by people who have liked your Page",
"name": "post_impressions_fan_paid",

"description": "Lifetime: The number of paid impressions of your Page post to people who have liked your Page. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_impressions_fan_paid/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_video_complete_views_30s_clicked_to_play": {"title": "Lifetime Clicked-to-Play 30-Second Views",

"name": "post_video_complete_views_30s_clicked_to_play",

"description": "Lifetime: Number of times people clicked to play your video and viewed it for 30 seconds or to the end, whichever came first. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_video_complete_views_30s_clicked_to_play/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_impressions_fan_unique": {"title": "Lifetime Post reach by people who like your Page",

"name": "post_impressions_fan_unique",

"description": "Lifetime: The number of people who saw your Page post because they've liked your Page (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_impressions_fan_unique/lifetime",

"period": "lifetime",

"values": [{"value": 3566}]},

"post_stories": {"title": "Lifetime Post Stories",

"name": "post_stories",

"description": "Lifetime: The number of stories generated about your Page post. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_stories/lifetime",

"period": "lifetime",

"values": [{"value": 42}]},

"post_impressions": {"title": "Lifetime Post Total Impressions",

"name": "post_impressions",

"description": "Lifetime: The number of impressions of your Page post. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_impressions/lifetime",

"period": "lifetime",

"values": [{"value": 4072}]},

"post_consumptions_unique": {"title": "Lifetime Post Consumers",

"name": "post_consumptions_unique",

"description": "Lifetime: The number of people who clicked anywhere in your post. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_consumptions_unique/lifetime",

"period": "lifetime",

"values": [{"value": 107}]},

"post_video_views": {"title": "Lifetime Total Video Views",

"name": "post_video_views",

"description": "Lifetime: Total number of times your video was viewed for more than 3 seconds. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_video_views/lifetime",

"period": "lifetime", "values": [{"value": 0}]},

"post_video_views_paid_unique": {"title": "Lifetime Paid Video Views",

"name": "post_video_views_paid_unique",

"description": "Lifetime: Number of times your video was viewed more than 3 seconds after paid promotion. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_video_views_paid_unique/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_video_complete_views_30s_unique": {"title": "Lifetime Unique 30-Second Views", "name": "post_video_complete_views_30s_unique",

"description": "Lifetime: Number of unique people who viewed your video for 30 seconds or to the end, whichever came first. (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_video_complete_views_30s_unique/lifetime",

"period": "lifetime",

"values": [{"value": 0}]},

"post_fan_reach": {"title": "Lifetime Post reach by people who like your Page",

"name": "post_fan_reach",

"description": "Lifetime: The number of people who saw your Page post because they've liked your Page (Unique Users)",

"id": "114431401958534_940695755998757/insights/post_fan_reach/lifetime","

"period": "lifetime",

"values": [{"value": 3566}]},

"post_stories_by_action_type": {"title": "Lifetime Post Stories by action type",

"name": "post_stories_by_action_type",

"description": "Lifetime: The number of stories created about your Page post, by action type. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_stories_by_action_type/lifetime",

"period": "lifetime",

"values": [{"value": {"comment": 35, "like": 6, "share": 1}}]},

"post_consumptions": {"title": "Lifetime Post Consumptions",

"name": "post_consumptions",

"description": "Lifetime: The number of clicks anywhere in your post. (Total Count)",

"id": "114431401958534_940695755998757/insights/post_consumptions/lifetime",

"period": "lifetime",

"values": [{"value": 170}]},

"post_story_adds_by_action_type": {"title": "Lifetime Post Stories by action type",

"name": "post_story_adds_by_action_type",

```
                "description": "Lifetime: The number of stories created about your Page post, by action type. (Total Count)",
                "id": "114431401958534_940695755998757/insights/post_story_adds_by_action_type/lifetime",
                "period": "lifetime",
                "values": [{"value": {"comment": 35, "like": 6, "share": 1}}]},
        "post_video_views_autoplayed": {"title": "Lifetime Auto-Played Video Views",
                "name": "post_video_views_autoplayed",
                "description": "Lifetime: Number of times your video started automatically playing and people viewed it for more than 3 seconds. (Total Count)",
                "id": "114431401958534_940695755998757/insights/post_video_views_autoplayed/lifetime",
                "period": "lifetime",
                "values": [{"value": 0}]},
        "post_impressions_organic": {"title": "Lifetime Post Organic Impressions",
                "name": "post_impressions_organic",
                "description": "Lifetime: The number of impressions of your post in News Feed or ticker or on your Page's Timeline. (Total Count)",
                "id": "114431401958534_940695755998757/insights/post_impressions_organic/lifetime",
                "period": "lifetime",
                "values": [{"value": 4042}]},
        "post_impressions_by_story_type_unique": {"title": "Lifetime Post viral reach by story type",
                "name": "post_impressions_by_story_type_unique",
                "description": "Lifetime: The number of people who saw your Page post in a story from a friend, by story type. (Unique Users)",
                "id": "114431401958534_940695755998757/insights/post_impressions_by_story_type_unique/lifetime",
                "period": "lifetime",
                "values": [{"value": {"other": 22}}]},
        "post_impressions_viral": {"title": "Lifetime Post Viral Impressions",
                "name": "post_impressions_viral",
                "description": "Lifetime: The number of impressions of your Page post in a story generated by a friend. (Total Count)",
                "id": "114431401958534_940695755998757/insights/post_impressions_viral/lifetime",
                "period": "lifetime",
                "values": [{"value": 30}]},
        "post_video_complete_views_paid": {"title": "Lifetime Paid views to 95%",
                "name": "post_video_complete_views_paid",
                "description": "Lifetime: Number of times your video was viewed to 95% of its length after paid promotion. (Total Count)",
                "id": "114431401958534_940695755998757/insights/post_video_complete_views_paid/lifetime",
                "period": "lifetime",
                "values": [{"value": 0}]},
        "post_video_complete_views_30s_autoplayed": {"title": "Lifetime Auto-Played 30-Second Views",
                "name": "post_video_complete_views_30s_autoplayed",
                "description": "Lifetime: Number of times your video started automatically playing and people viewed it for 30 seconds or to the end, whichever came
first. (Total Count)",
                "id": "114431401958534_940695755998757/insights/post_video_complete_views_30s_autoplayed/lifetime",
                "period": "lifetime",
```

"values": [{"value": 0}]}},

"description": "About 20 western Virginia high school students were suspended Thursday after holding a rally to protest a new policy banning vehicles with Confederate flag symbols from the school parking lot and refusing to take off clothing displaying the symbol.",

"caption": "komonews.com",

"id": "114431401958534_940695755998757",

"picture":
"https://external.xx.fbcdn.net/safe_image.php?d=AQBqd75RrvZ_AE0R&w=130&h=130&url=https%3A%2F%2Fscontent.xx.fbcdn.net%2Fhphotos-xft1%2Fv%2Ft1.0-9%2F12036903_940695302665469_1677160858596008416_n.jpg%3Foh%3Dd671adcc1ed17e30f4014523117a848c%26oe%3D56A4D844&cfs=1&sx=220&sy=0&sw=440&sh=440",

"created_time": "2015-09-17T20:50:00+0000"}