Understanding The Dataset

In the process of understanding the dataset, I figured identifying missing values would be a good place to start.

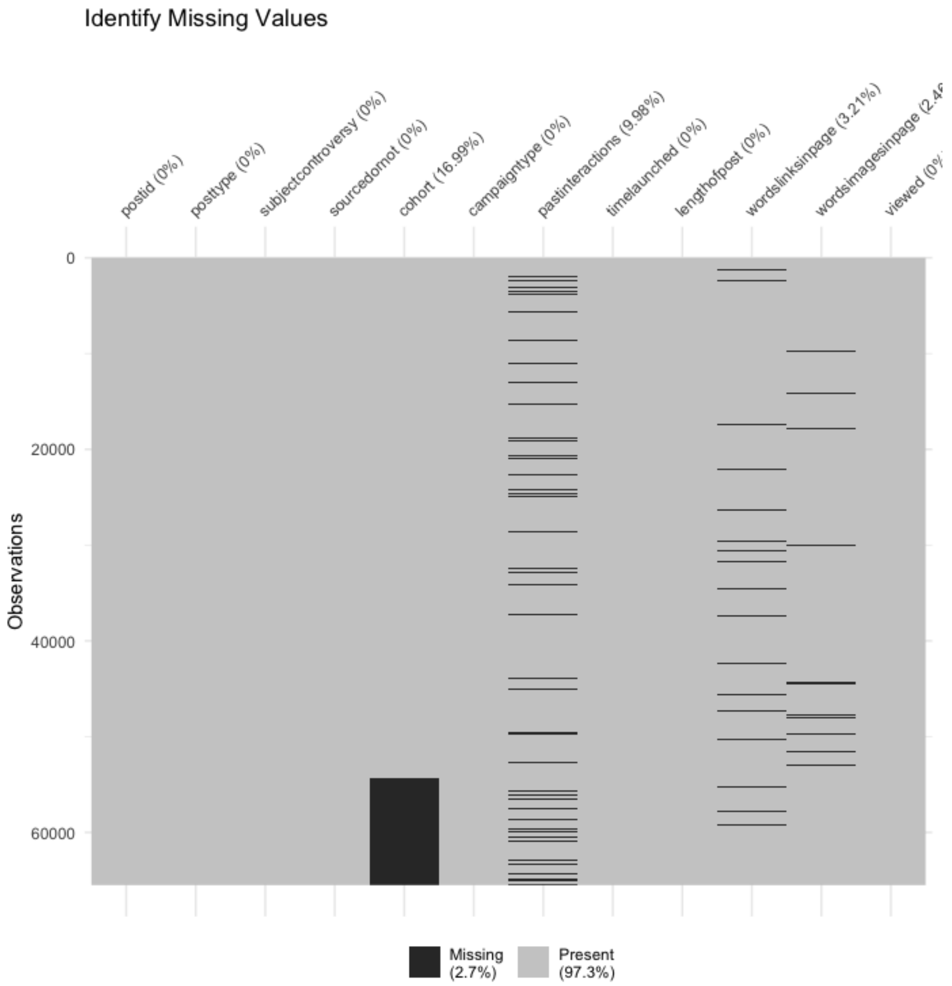


Chart 1 tells us that that all missing values are contained in 4 columns, with the bulk of it residing in the **cohort** variable. According to the data dictionary, the NAs under **cohort** would be referring to the unnamed cohort. The missing values in the other three columns will be dealt with in the next section.

Data Visualization

Next, the continuous variables were analyzed to get a feel for their distribution as well as to identify any patterns that may exist.

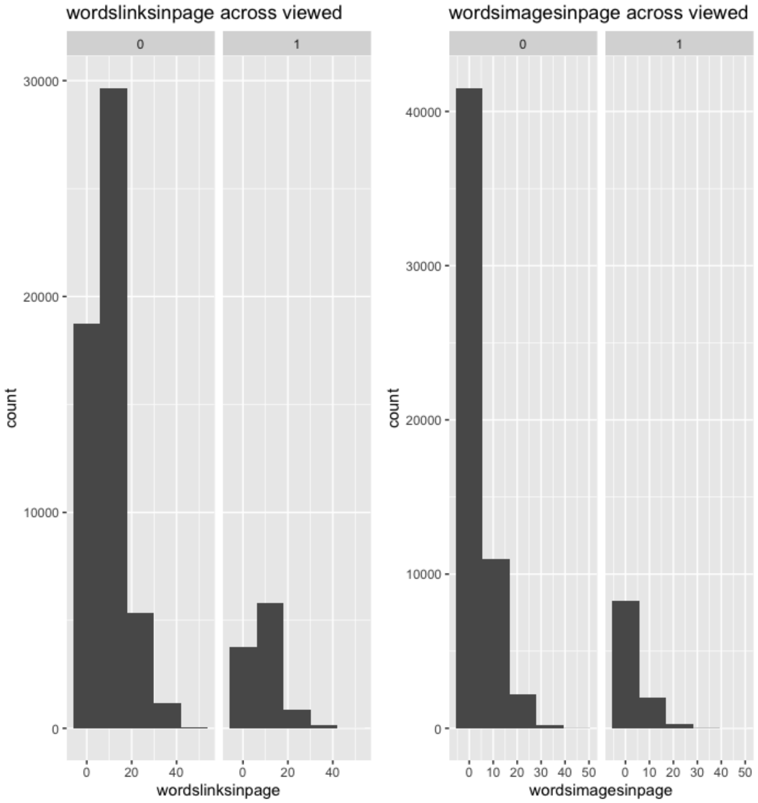


Chart 2

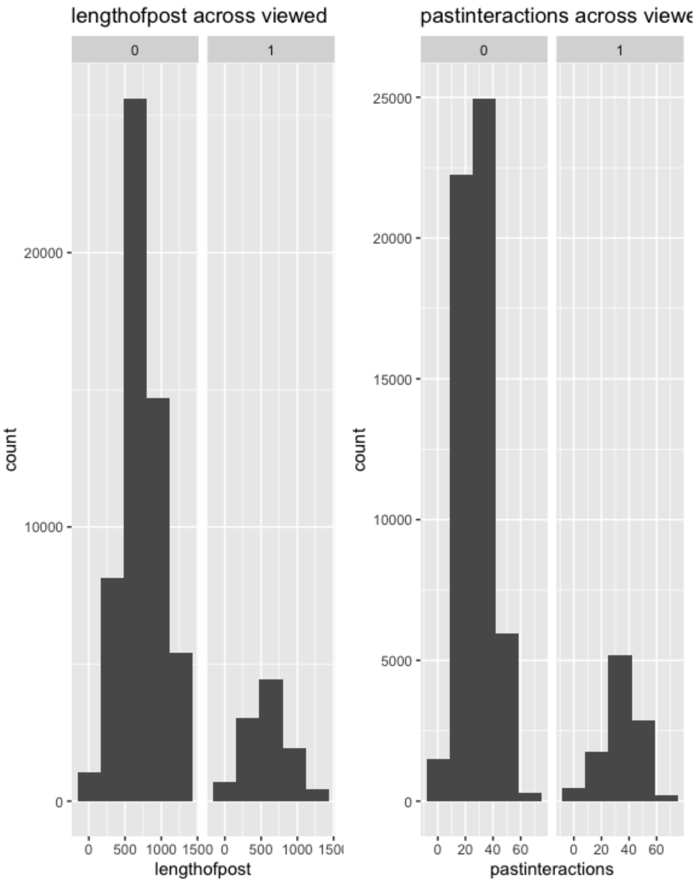


Chart 3

Chart 2 & 3 is a collection of histograms showing the distribution of 2 continuous variables against the dichotomous variable of whether or not a post was viewed (0 = not viewed; 1 = viewed). The mean of the distributions in **wordsimagesinpage** and **wordslinksinpage** is quite low, suggesting that the bulk of the data has a low numerical value, meaning most posts have few words associated with images and links. **lengthofpost** and **pastinteractions** seem to have roughly even distributions across posts that were viewed and those that were not.

Plotting categorical variables is a bit different than their continuous counterparts. Since these variables are discrete, frequency distributions are a more appropriate method of visualizing these variables.

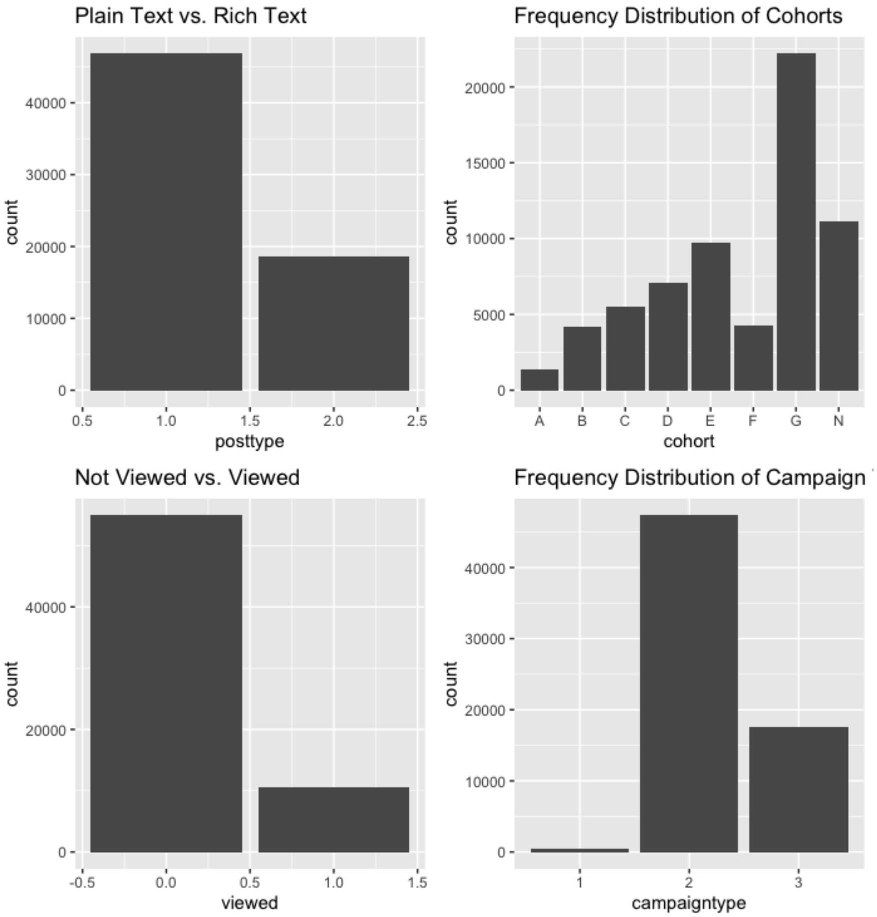


Chart 4

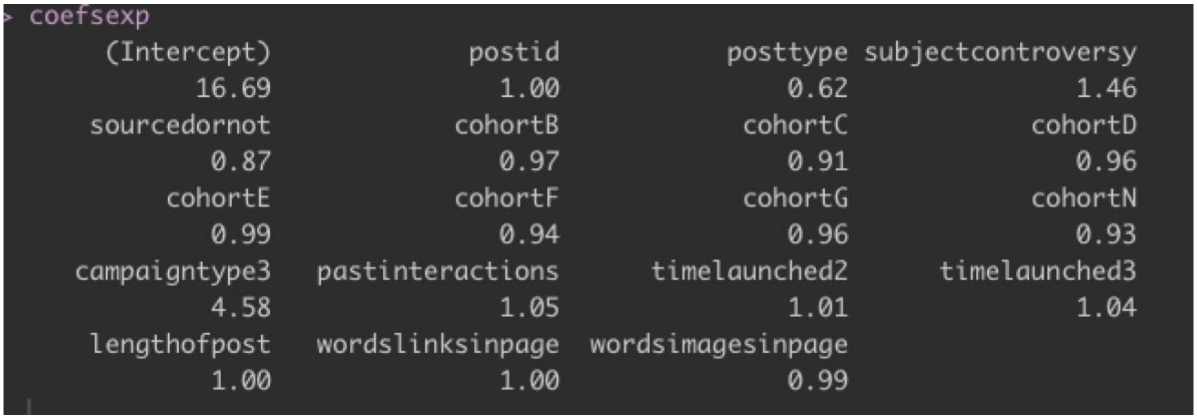
In Chart 4, it is clear to us that most posts are plain text, and also that most posts are not viewed. It is not yet known whether these two metrics overlap. Out of the different cohorts of consumer groups, the most popular cohort is G - fast food, takeout food, comedy central, hotels, and photography. It seems that campaigntype 1 is almost negligible in comparison to **campaigntype** 2 & 3. The disproportionate frequency may affect the model, which is why **campaigntype** 1 was dropped from the feature.

Since the variable we are trying to predict is binomial, it is appropriate to use logistic regression to perform this analysis. Before running the model, the data needs to be reformatted to ensure smooth operation. The missing data in the continuous variables were replaced with the means of their respective columns. This was a quick and easy way of ensuring that the variables do not have any missing values while also ensuring their distributions largely remain the same. The missing values in the cohort variable were replaced with ‘N’ to denote NA. Using ‘N’ would allow the model to treat the NAs as a level. Variables with more than two categories were converted to factors since the logistic regression model takes in numerical values.

Running the Logistic Regression Model

All variables were included in the initial model. **cohort**, **timelaunched**, **wordslinksinpage**, are “statistically insignificant” meaning that it has a negligible impact on this model, which may be attributed to any number of reasons including high levels of correlation with other variables/levels. It may also mean that the signal (effect) to noise (variation) ratio is too large. The AIC serves as an indicator to the model’s performance. It penalizes unnecessary coefficients; therefore, lower values of AIC are desirable. It can be used to compare the performance of multiple models, but must not be considered as the only metric for performance much like the p-value.

To make more sense of this model, an odds ratio table was created:



```
> coefsexp
       (Intercept)             postid           posttype subjectcontroversy
             16.69               1.00               0.62               1.46
      sourcedornot            cohortB            cohortC            cohortD
              0.87               0.97               0.91               0.96
           cohortE            cohortF            cohortG            cohortN
              0.99               0.94               0.96               0.93
     campaigntype3   pastinteractions      timelaunched2      timelaunched3
              4.58               1.05               1.01               1.04
      lengthofpost   wordslinksinpage  wordsimagesinpage
              1.00               1.00               0.99
```

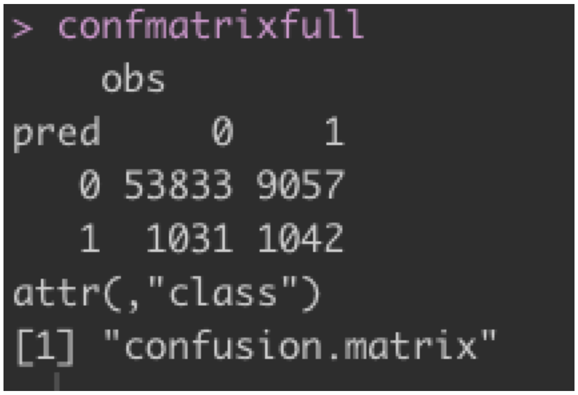
Table 1

Going off of Table 1, it seems that posts with Rich Text are less likely to be viewed compared to posts with Plain Text. Posts seem to be viewed more if they are more controversial. Interestingly, posts with sources perform worse than posts without. It seems that posts perform better when they have previously interacted with members of its respective cohort. As the number of words related to an image in the post increases, it is less likely that a post will be viewed. As a note of caution, it is important to identify **lengthofpost** and **wordslinksinpage** as interacting variables,

To simplify the model even further, a stepwise function was used to cut down predictor variables resulting in a model with similar performance and accuracy to the original model but with less complexity. It behaves like an Occam’s razor for regression models. The tuned model showed signs of improvement, with a slightly lower AIC score (Model1: 49021, Model2: 49006). The odds ratios of each variable were largely similar, as the StepAIC’s main function is to remove insignificant variables, not alter significant variables.

Calculating Optimal Payoff

Using the predictive power of the original model, a new column was created in the dataset predicting whether a post will be viewed or not. To measure the accuracy of the predictions made by the model, a confusion matrix was built:



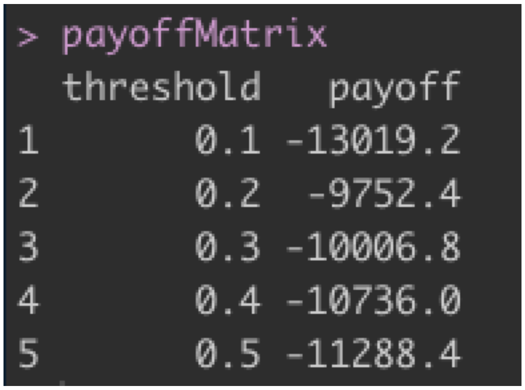
```
> confmatrixfull
   obs
pred     0    1
   0 53833 9057
   1  1031 1042
attr(,"class")
[1] "confusion.matrix"
```

Table 2

As Table 2 shows, there split between True Positives (which means that the model predicted that the post will be viewed and it was) is higher than False Positives (Also known as Type 1 Error - which means that the model predicted that the post will be viewed and it was not). On the other side of things, True Negatives (the model predicted that the post will not be viewed and it was not) far outweighed the number of False Negatives (Also known as Type 2 Error - the model predicted that the post will not be viewed and it was. The accuracy of the original model is 84.48%. The Confusion matrix for the new model is similar to the matrix of the original model, and its accuracy is also 84.48%.

To quantify the costs associated with predicting Type 1 and Type 2 errors, a Payoff Matrix was built with the help of the confusion matrix. It costs the Firm $0.4 for each Type 1 error, and a loss of $1.2 for each Type 2 error. This intuitively makes sense since underestimating the number of posts that would be viewed is more damaging to the Firm, than overestimating it.

payoff = -0.4 * False positive - 1.2 * false negative



```
> payoffMatrix
  threshold   payoff
1       0.1 -13019.2
2       0.2  -9752.4
3       0.3 -10006.8
4       0.4 -10736.0
5       0.5 -11288.4
```

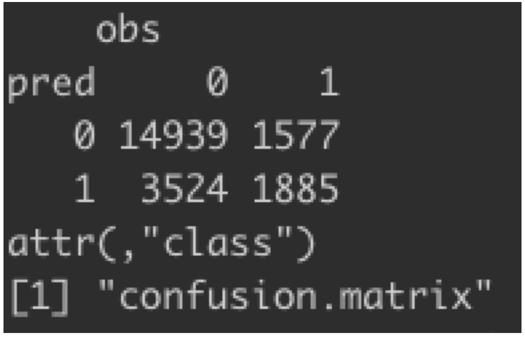
Table 3

Since the company does not make a profit off of either Type 1 or Type 2 errors, all payoffs are negative as it costs the Firm money. The highest payoff is the most desirable outcome in the context of this problem, which corresponds to a threshold of 0.2. This threshold will be useful when cross validating the model.

Cross Validating the Regression Model

It is important to ensure that the model being built is not overfitted. Overfitting essentially means that the model does a great job modeling a specific dataset but no other dataset, which renders it useless since it has memorized specific patterns and structures pertaining to one dataset. To avoid this, the dataset is split into a training set and a test set. Approximately 2/3rds of the data is randomly assigned to the training set and the rest is assigned to the test set. The training set model was created using the formula from the StepAIC regression model built earlier.

The lower AIC score of the model run on the training dataset suggests that it is more effective than the StepAIC model. The relatively smaller size of the Training dataset may have caused this decrease in AIC. Using this model, the predictions in the test dataset were created, along with the confusion matrix and the accuracy of the predictions.



```
    obs
pred     0    1
   0 14939 1577
   1  3524 1885
attr(,"class")
[1] "confusion.matrix"
```

Table 4

Table 4 shows that the new model has been optimized for reducing the number of false negatives. On the other hand, the number of False Positives remains relatively high compared to True Positives. The accuracy of this model is 76.49%. To cross validate this model, the model was run once again with the threshold chosen from the payoff matrix (0.2). The accuracy of this model was 76.65%. We can conclude that the results for the training set are comparable the cross validated model, which confirms the stability of this model.

Limitations of This Model

The cohort variable contained a section of NA values which were replaced with the letter “N” to represent a cohort with insufficient information. The model may have performed better if the “N” cohort was dropped out of the variable. Similarly, campaigntype1 had very few observations, which may have made it a bad anchor in the model. If it was removed, the model may have been altered significantly.

The variable PostID was included in the model, which may have impacted the model's performance. PostID is a unique identification code, which, although statistically significant, does not logically make sense to include in the model.

Conclusion

The models were built to predict what influences post interactions. However, the dataset itself is a stark reminder of the difficulty in achieving and maintaining consistent interaction within cohorts. In order to improve future performance, it would be optimal to focus on plain text posts, with a sustainably high level of controversy, which seems organic (does not contain a source, i.e. does not come across as overtly professional) and is as concise as possible. To maximize reach, it would be ideal to focus on cohorts with larger populations, such as the cohort containing fast food, takeout food, comedy central, hotels, photography which would mean that the target audience could be amateur and professional photographers, food bloggers, tourists, comedy fans, and a broader, more generalized audience surrounding takeout food and fast food.