Investigating the Popularity of Scientific Phenomena on Social Media Platforms

Sal Fu, Brian Lorenz, Jerry Xuan

Section 1: Motivation

Large numbers of scientific studies are published each day, many of which address pertinent gaps in human knowledge. Some results, however, become much more visible to the general populace than others. Why do some scientific results go viral, while others remain obscure? We wish to investigate what factors might influence the popularity of links about science on a social media platform and the insights that may lend to scientific communication.

We choose Reddit as our platform for analysis because it is the 8th most visited site around the world, with over 500 million users total. This particular platform is convenient because it has a specific forum, the "Science" subreddit, that is dedicated to sharing and discussing scientific results. This subreddit can also reach a broad viewership because all Reddit users are subscribed to it by default (at the time this was written, the "Science" subreddit has around 18 million subscribers).

On Reddit, users can post content and "Upvote" the posts that they like. Within a subreddit, posts with higher upvotes will end up on the first page of the subreddit (e.g. they are the first posts that users will see when they visit the subreddit). When a post in "Science" gathers enough votes, it will end up on the "Reddit" frontpage, where it will get an even large viewership (i.e., all Reddit users will see the post when they open the website).

For this work, we will use "Upvotes" as our metric of whether a scientific article is popular: i.e., our response variable. We plan to build a random forest model to understand which variables may contribute the most to whether a scientific article garners a large number of votes.

The structure of this document is as follows: In section 2, we describe how we acquire data from Reddit, and how we use that data to acquire more information about the posts. In section 3, we describe the process of manipulating various columns to make our model more robust. In section 4, we describe the process of model building and interpretation. In Section 5, we summarize our results and interpret them in the context of our broader question.

Section 2: Data Acquisition and Modification

Using the Python Reddit API Wrapper, we obtain data about individual posts in the "Science" subreddit for the past two years (from Nov 2015 to Nov 2017). We query the number of upvotes that a post has, as that will be our metric of popularity. We also query for fields that represent explanatory variables, variables which could potentially be used to predict a post's popularity. For instance, these include the title of the post, the time at which the post was created, the amount of "karma" the post's author has, the number of comments on the post, the scientific subfield of the post, and so on.

With the fields that we queried as a basis, we extract additional information from them. For example, as we acquired data from the API, we create an explanatory variable called the "journal h index", which tracks posts that link to "high-impact" scientific articles. For this, we reference the h index rankings from Scimago Country and Rank (scimagojr.com), and define posts with a high h index as those that link to journals which are in the top 10 of the h index rankings. When attempting to download the data, some explanatory variables were not in a readable format for a .csv file, so they required encoding. However, a few characters in the past two years of data broke that encoding. As a result, we acquired the data from November 2015-November 2017 in 9 chunks. The results from the API query are saved in the repository as `reddit_df_final[number].csv`, with `[number]` ranging from 1 to 9 in the order that they finished downloading (not necessarily chronological).

After we have generated this dataset, we work to extract additional information out of the post titles. First, we calculate the total length, mean word length, and max word length of each title. Then, we perform sentiment analysis on the titles to obtain a sentiment score for each title. All these became new explanatory variables for our data.

The title sentiment analysis works as follows: we obtain the corpus of words from SentiWordNet, which assigns nearly every word in the English language a positive or negative index with a value between 0 and 1. From each of our post titles, we remove "stopwords" like "about" and "the" (words which do not have positive or negative connotations). Taking words with an overall positive meaning to have a positive value, and treating negatively-connotating words analogously, we sum up the total indices of the remaining words in the title. We then divide that sum by the number of words in the title to acquire our desired variable, which can be thought of as the title's sentiment "density".

As the sentiment code takes multiple hours to run on our laptops, we have performed the analysis and saved the results in the repository as `reddit_df_final[number]_sent.csv`, with `[number]` matching the number index of the dataframe. In the interest of reproducibility, we have left a toggle setting below: if you set skip=FALSE, then the sentiment analysis will run before the rest of the code and recreate the `reddit_df_final[number]_sent.csv` files. If skip=TRUE (default), then the code runs by reading in the files that we have already performed the analysis on.

We wanted to get a sense for whether sentiment analysis worked as expected, so we looked at the titles with the most "positive" sentiment and the most "negative" sentiments. The top five positive titles are below:

```r
reddit_df %>%
  arrange(desc(title_sent)) %>%
  select(title,title_sent) %>%
  head(n=5L)
```

```
## # A tibble: 5 x 2
##                                                  title title_sent
##                                                  <chr>      <dbl>
## 1 Nonnutritive sweeteners and cardiometabolic health.      0.625
## 2   Babies know when they don<U+2019>t know something      0.625
## 3     Invasives are changing the color of woodpeckers      0.500
## 4    Facebook posts inspired by envy, UBC study finds      0.500
## 5    Happiness doesn't bring good health, study finds      0.475
```

The most positive two titles are: "Nonnutritive sweeteners and cardiometabolic health" and "Babies know when they don't know something". Some of these are reasonable? But others are: "Happiness doesn't bring good health, study finds," and "Facebook posts inspired by envy, UBC study finds." For some of these, we can see where the sentiment analysis failed. For instance, "Happiness doesn't bring good health" has positive words like "happiness," "good," and "health", which can make the average sentiment of the title positive even if the presence of the single negation "doesn't" alters the overall meaning of the statement.

Now, we print the five posts with the most negative title sentiments:

```r
reddit_df %>%
  arrange(title_sent) %>%
  select(title,title_sent) %>%
  head(n=5L)
```

```
## # A tibble: 5 x 2
##                                                          title title_sent
##                                                          <chr>      <dbl>
## 1                                          Induced Earthquakes    -0.6250
## 2                           Cats protect newborns against asthma    -0.5625
## 3                      "CRISPR \"Drives\" Out Fungal Resistance"    -0.5000
## 4 Intertumoral Heterogeneity within Medulloblastoma Subgroups    -0.5000
## 5                                              Sirtuin and LCAT    -0.5000
```

Sentiment analysis was a great exercise, but late,r when we computed variable importance for our random forest model, we will that its lack of robustness meant that it was not a meaningful predictor for our response variable.

Section 3: Variable Modification

After retrieving the data, we take multiple steps to wrangle it into a format that makes more sense given the variables we are trying to investigate. The data for "Time" that we retrieve from is in POSIX format, units of seconds elapsed from January 1st, 1970 in UTC time. We thus translate that metric of time into year, date, and hour.

After obtaining that information, we create a categorical variable for "time" to categorize the time of day that the post was created into "Day", "Morning", and "Night". This step was necessary because otherwise, the classification algorithm would not classify "23" and "0" as the same portion of the day (time wise) even though they should be. Furthermore, having 24 categories for hour would not be ideal for model building, as that gives too many options for splitting. We perform similar modifications to the "day" column, turning it into the time of month posted, and the "month" column, breaking that into the four seasons.

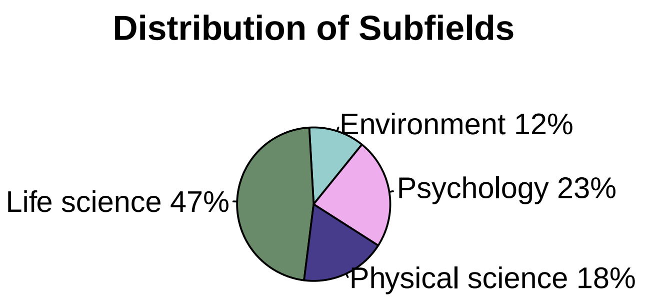
Next, we clean up the data in order to remove posts that do not share scientific content (e.g. AMAs, subreddit discussions) by filtering under the subfield column. We also remove all subfields that have less than 100 posts, since these were mostly errorneously named subfields that had only 1 post.

Now, we group the remaining subfields into four major categories so that our model can more easily divide among them. We made these selections by our impressions of the categories that these fields fall into, so this step is certainly subjective. Our funciton for conversion is below to display our categorizations.

```r
#Make a function that sorts fields into "Life Science,"
#"Psychology," "Physical Science," "Environment"
mk_cat_subfield <- function(field)
{
  ifelse(field == "Biology" | field =="Health" | field =="Cancer" |
           field =="Medicine" | field =="Animal Science" | field =="Paleontology" |
           field =="Epidemiology", "Life Science",

         ifelse(field =="Psychology" |
           field =="Neuroscience" | field =="Social Science" |
            field =="Anthropology", "Psychology",

           ifelse(field =="Geology" | field =="Astronomy" | field =="Physics" |
            field =="Chemistry" | field =="Nanoscience" | field =="Engineering" |
            field =="Computer Science", "Physical Science",

            "Environment")))
}
```

Here we make a pie plot of the distribution of articles into these four categories:

```r
subfield_freq = table(reddit_df$cat_subfield)
type_freq_perc = paste(round((subfield_freq/length(reddit_df$cat_subfield)),2)*100,
                       "%",sep="")
type_freq_lab = paste(c("Environment","Life science",
                        "Physical science", "Psychology"), type_freq_perc, sep = " ")
pie(subfield_freq,labels=type_freq_lab, init.angle = 51,edges = 10000,
    col = c("paleturquoise3","darkseagreen4","slateblue4","plum2"), main="Distribution of Subfields")
```

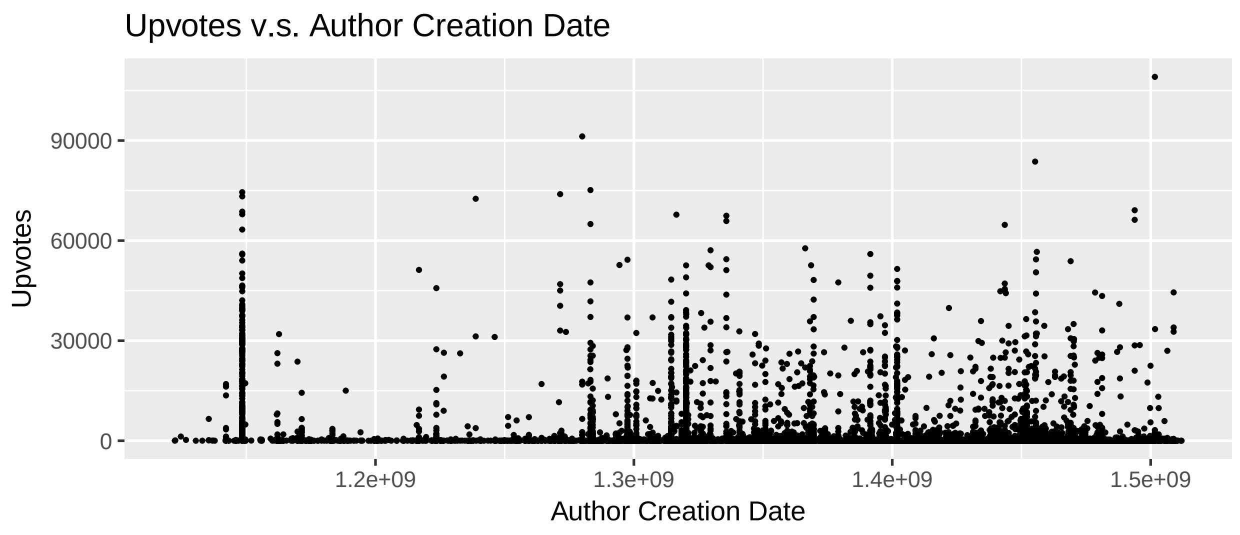


Next, we turn the categorical variables to factors, which the caret random forest model expects.

Then we plot select explanatory variables against our response variable to get a sense of any correlations between them.

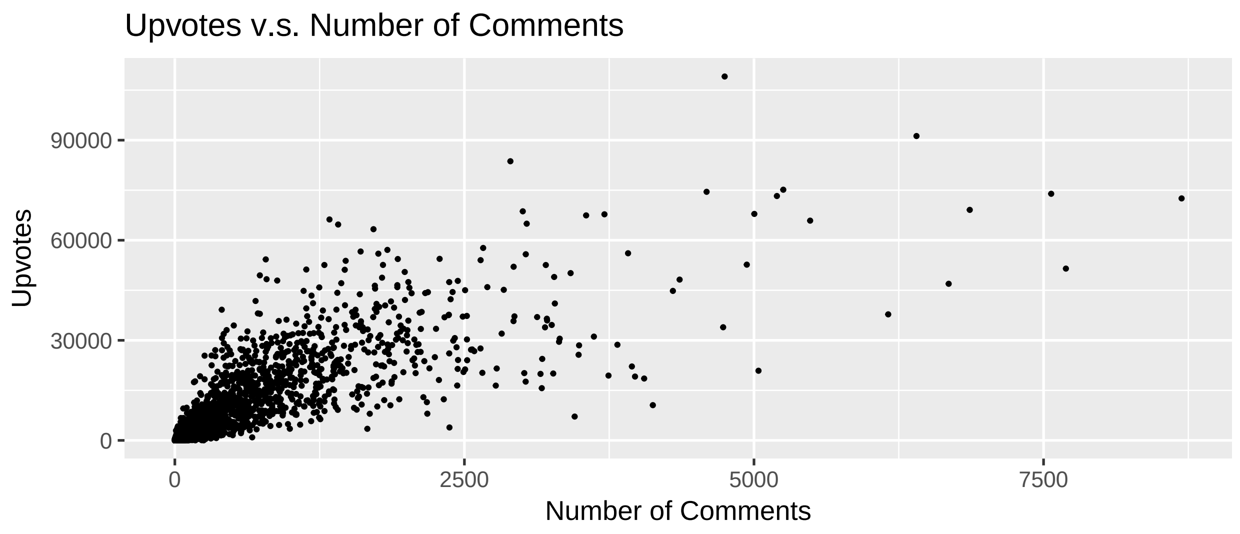
Author created date vs upvotes. We see a few vertical lines of dots (e.g. the one near 1.15e+09), which we looked into and realized that they are single authors that are prolific on Reddit! Since these outliers could bias our model's prediction vis-a-vis author_created_date, we later remove that column from the dataframe.

```
acd_plot <- ggplot(data=reddit_df, aes(x=author_created_date, y=upvotes))+
  geom_point(size=0.5) +
  xlab("Author Creation Date")+
  ylab("Upvotes")+
  ggtitle("Upvotes v.s. Author Creation Date")
ggsave(acd_plot, filename='./acd_plot.png', width=8, height=5)

acd_plot
```



Number of comments vs upvotes. This is an example of a variable that shows strong positive correlation with our response variable. Indeed, we could have used number of comments instead of upvotes as our response variable and achieved a very similar model. To create a fair and meaningful prediction model, we thus remove num_comments and other highly correlated variables (link_karma,comment_karma, and gilded).

```
nc_plot <- ggplot(data=reddit_df, aes(x=num_comments, y=upvotes))+
  geom_point(size=0.5) +
  xlab("Number of Comments")+
  ylab("Upvotes")+
  ggtitle("Upvotes v.s. Number of Comments")
ggsave(nc_plot, filename='./nc_plot.png', width=8, height=5)

nc_plot
```



The other columns removed below are due to conversions to categorical varialbes (i.e. post_hour is now irrelevant since we have cat_post_hour, breaking that variable into three times of day).

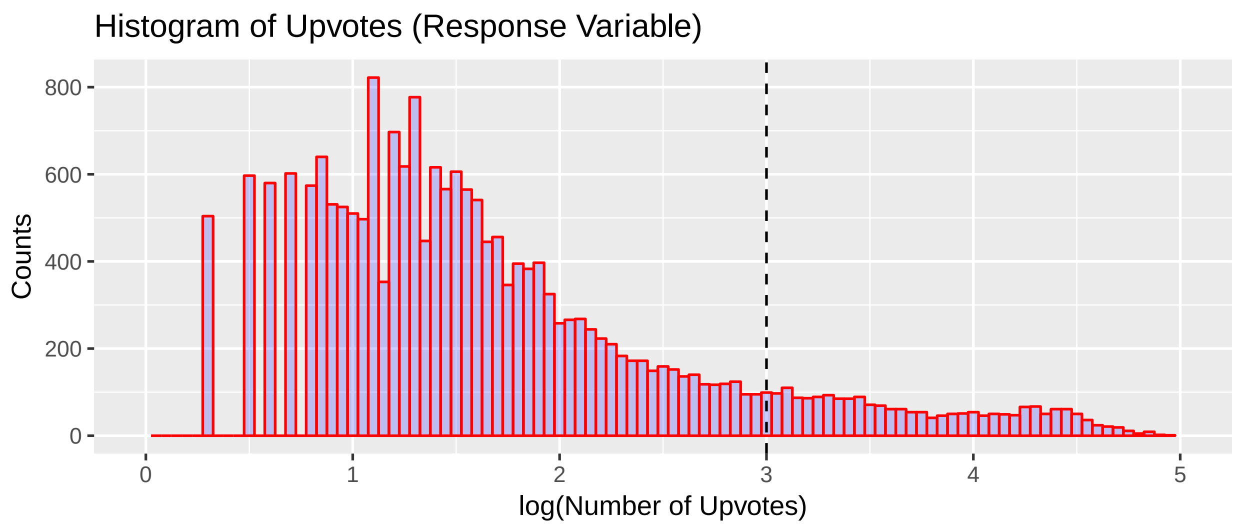
```
#To prepare for model building, we remove unnecessary columns
reddit_df <- reddit_df %>%
  select(-X1,-author,-id,-created_utc,-domain,-url,-author_flair,-post_day,
         -post_hour,-post_month,-num_comments,-gilded,-link_karma,
         -comment_karma,-title,-X1_1,-author_created_date)
```

Now we remove any rows that have missing values (NAs). We also print the proportion of rows removed below to ensure that we are not removing a large fraction of the data. It is comforting to see that less than 0.03% of our observations contained any NAs, so it is unlikely that we are introducing significant bias by removing these.

```
## [1] "Proportion of rows removed:  0.000235971494643447"
```

Here we plot a histogram of the number of upvotes, showing that the vast majority of our data has "low" (<1000) upvotes. The huge range (from 0 to 100000) and high level of skewedness of upvotes motivate us to separate our response into a binary variable. Before we do so,we examine two more plots:

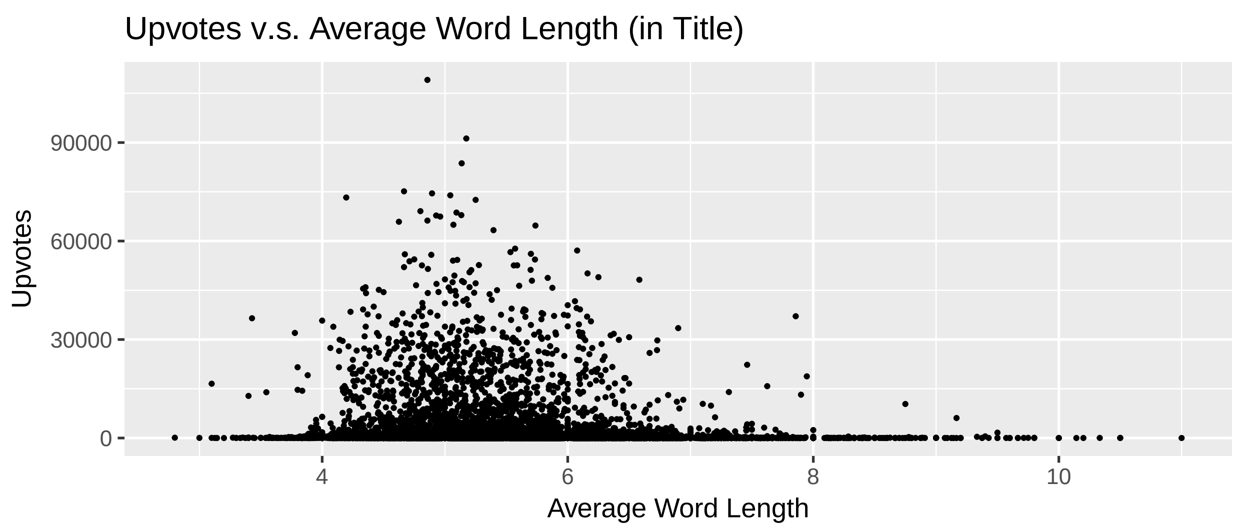
```
hist_upvotes <- qplot(log10(reddit_df$upvotes+1),
      geom="histogram",
      binwidth = 0.05,
      main = 'Histogram of Upvotes (Response Variable)',
      xlab = "log(Number of Upvotes)",
      ylab = "Counts",
      fill=I("blue"),
      col=I("red"),
      alpha=I(.2),
      xlim=c(0,5)) +
  geom_vline(xintercept=3,show_guide=TRUE,linetype="dashed")

hist_upvotes
```



```
#ggsave(hist_upvotes, filename='./hist_upvotes.png', width=8, height=5)
```

Here, we plot the mean title word length vs. number of upvotes. There appears to be a sweet spot around 5. It appears that lower mean title word lengths give higher upvotes. There are, however, more posts near the low range too, so we cannot make that conclusion just yet.

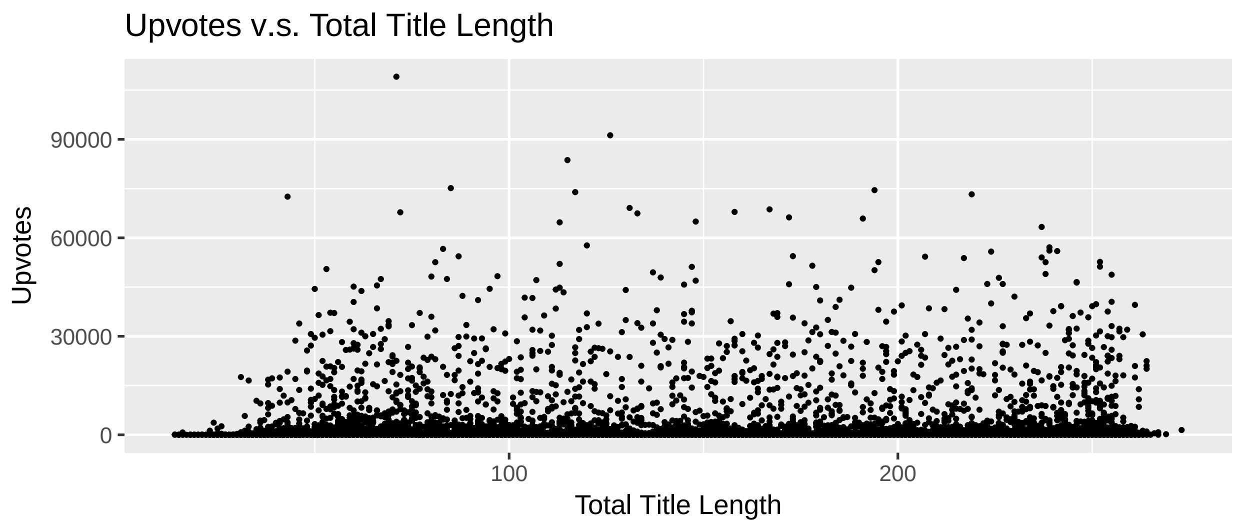
```
mtl_plot <- ggplot(data=reddit_df, aes(x=mean_title_length, y=upvotes))+
  geom_point(size=0.5) +
  xlab("Average Word Length")+
  ylab("Upvotes")+
  ggtitle("Upvotes v.s. Average Word Length (in Title)")

mtl_plot
```



```
#ggsave(mtl_plot, filename='./mtl_plot.png', width=8, height=5)
```

Here, we plot the total title length vs. number of upvotes. No strong correlation could be seen easily by eye.

```
tl_plot <- ggplot(data=reddit_df, aes(x=title_len, y=upvotes))+
  geom_point(size=0.5) +
  xlab("Total Title Length")+
  ylab("Upvotes")+
  ggtitle("Upvotes v.s. Total Title Length")
ggsave(tl_plot, filename='./tl_plot.png', width=8, height=5)

tl_plot
```



Now we turn the continuous upvote response variable into a binary response. We choose 1000 upvotes as our cut-off, because that is the approximate threshold where a post will end up on the front page of the Science subreddit, from which it could reach the front page of the entire website.

```
#Funciton to create a new column for categorical upvotes
#Ranges: 1000+ high      0+ low
mk_cat_upvotes <- function(upv)
```

```r
{
  ifelse(upv>1000, "High", "Low")
}

#Apply the function and remove the old column
reddit_df <- reddit_df %>%
  mutate(cat_upvotes = mk_cat_upvotes(upvotes)) %>%
  select(-upvotes)

#Turn the column into a factor
reddit_df$cat_upvotes <- factor(reddit_df$cat_upvotes)
```

Section 4: Model Building and Assessment

After wrangling the data, we are ready to build our Random Forest model. We partition the data into test (20%) and training (80%) sets, setting aside the test data for later use to appropriately assess our model's accuracy. We did not apply this test data until after finalizing and tuning our model, so it will be an independant assessment of accuracy.

```r
#Set the seed for reproducibility
set.seed(47)

#Split the data into test and training, with 80% going to training
inTrain <- createDataPartition(y = reddit_df$cat_upvotes, p=0.80, list=FALSE)
reddit.train <- reddit_df[inTrain,]
reddit.test <- reddit_df[-c(inTrain),]
```

By looking at the number of observations in each category below, we see that predicting everything as "low" gives a fairly high accuracy. In other words, for the model, there is little consequence in masclassifying the posts with "high" upvotes. To avoid this problem in building our model, we sample from our training set to have an equal number of observations with "high" and "low" upvotes. In doing so, we force the model to find signal to separate the two populations rather than predicting everything as low.

```
## [1] "Number of high upvotes:  1724"
```

```
## [1] "Number of low upvotes:  15224"
```

```r
### Sample data such that there is a equal number of observations for each category.
set.seed(4747)

#Count the number of high upvotes
num_high_upvote <- sum(reddit.train['cat_upvotes'] == "High")

reddit.train.high <- subset(reddit.train, reddit.train['cat_upvotes'] == "High")
reddit.train.low <- subset(reddit.train, reddit.train['cat_upvotes'] == "Low")

#Select an equal proportion of high and low upvotes
inTrain.low <- createDataPartition(y = reddit.train.low$cat_upvotes,
                                   p=num_high_upvote/dim(reddit.train.low)[1],
                                   list=FALSE)
reddit.train.low <- reddit.train.low[inTrain.low,]

### Combine the 2 dataframes into the final training data
reddit.train.final <- rbind(reddit.train.low, reddit.train.high)
```

Below we see the final number of training observations to be used by the model and verify that there are an equal proportion of "high" and "low" upvotes.

```
## # A tibble: 2 x 2
##   cat_upvotes `n()`
##        <fctr> <int>
## 1        High  1724
## 2         Low  1724
```

Now we grow the random forest of 1000 trees on this training set, using the out of bag (OOB) error rate to tune the `mtry` parameter (number of variables at each split). We let mtry be an odd number from 3-11, where these values were chosen to hopefull find some signal (mtry=1 may not be particularly useful since it never directly compares two variables at a given split) and doesn't consider too much of the data at every split (we have 14 explanatory variables).

```
#Build the random forest model. This chunk takes ~3 minutes to run
set.seed(4747)

rf.reddit <- train(cat_upvotes ~., data=reddit.train.final, method="rf",
                      trControl = trainControl(method="oob"),
                      ntree=1000, tuneGrid = data.frame(mtry=c(3,5,7,9,11)),
                      importance = TRUE,na.action = na.exclude)
```

The value of mtry does not appear to have a large impact on our accuracy as seen below, but mtry=3 gives the greatest oob prediction rate, so we build our final model with that value.

```
## Random Forest
##
## 3448 samples
##   12 predictor
##    2 classes: 'High', 'Low'
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    3    0.6345708  0.2691415
##    5    0.6209397  0.2418794
##    7    0.6218097  0.2436195
##    9    0.6142691  0.2285383
##   11    0.6174594  0.2349188
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 3.
```

A more in-depth look at our final model is below, where we see that it predicted more observations as "high" than "low," leading to a lower error rate for "high" upvoted posts than "low" upvoted.

```
##
## Call:
##  randomForest(x = x, y = y, ntree = 1000, mtry = param$mtry, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 36.8%
## Confusion matrix:
```

```
##      High  Low class.error
## High 1146  578   0.3352668
## Low   691 1033   0.4008121
```

Using this best model, we predict our test data to assess our model's accuracy.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction High  Low
##       High  280 1441
##       Low   151 2364
##
##                Accuracy : 0.6242
##                  95% CI : (0.6094, 0.6388)
##     No Information Rate : 0.8983
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1164
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6497
##             Specificity : 0.6213
##          Pos Pred Value : 0.1627
##          Neg Pred Value : 0.9400
##              Prevalence : 0.1017
##          Detection Rate : 0.0661
##    Detection Prevalence : 0.4063
##       Balanced Accuracy : 0.6355
##
##        'Positive' Class : High
##
```

The confidence interval for the accuracy of the model is between 0.6094 and 0.6388, which is greater than 50%: this suggests that there *is* some information in this model, i.e., that there is a basis on which to distinguish popular posts from less popular posts. With that in mind, we print the variable importance from the model, which indicates which explanatory variables most directly impact the response. These values are computed by examining the decrease in node purity when permuting the values of a particular variable across the dataset. If permuting the values of a variable causes node to be significantly more impure, then that variable is likely very important in distinguisihing the response.

```
## rf variable importance
##
##                              Importance
## title_len                       32.4462
## imageyes                        22.9815
## mean_title_length               18.0403
## journal_h_indexlow              14.9710
## author_flair_binaryyes           5.6484
## cat_post_hournight               4.7274
## post_year2017                    3.4060
## cat_subfieldPsychology           3.2528
## cat_post_monthSummer             2.8350
## cat_subfieldLife Science         2.3745
## cat_subfieldPhysical Science     2.1147
## cat_post_hourmorning             1.1758
## max_title_length               0.9420
## post_year2016                  0.9390
## cat_post_monthWinter          -0.3165
## cat_post_monthSpring          -1.1771
## cat_post_daymid               -1.7099
## cat_post_daylate              -1.9566
## title_sent                    -2.4560
```

Of the variables that we use in our model, the output above suggests that the four most important variables contributing to the popularity of a post are:

1. Length of the title of the post (longer better)
2. Whether there is a thumbnail accompanying the post (thumbnail better)
3. The average length of words in the title (shorter better)
4. The *h* index of content linked to in the post (lower better)

Beyond this, there is a larger dropoff before the next variable, so it is safe to conclude that these are the most important four variables accroding to our model.

The first one might be because long post titles that summarize the content of the article allow readers to get their information without having to click through the article, causing them to upvote. Thumbnails give an article more legitimacy and catch the reader's attention on the webpage, so the second one also makes sense. The third and the fourth could be interpreted together, suggesting that readers are looking for a simpler presentation of the content. Having a rigorous, academic seems to negatively impact upvotes, so linking to a popular science summary of such an article might garner more attention. Also, the shorter average word length suggests that readers dislike long scientific jargon and would rather have the article presented in more digestable terms.

Beyond these top four, we see that the subfield that the article is about is mostly irrelevant. This seems to suggest that the presentation of the article - rather than the content - attracts readers.

Section 5: Conclusions

In this project, we attempt to look for variables that may affect whether a science post on Reddit gathers a large number of upvotes. From the results of the previous section, we find that the most important variables are post title length, whether there is a thumbnail, the average length of a word in the title, whether the post links to a high-impact scientific journal, and whether the the post was published at nighttime PST. It seems that whether the post.

Overall, these results suggest a kind of "fast-food" model of scientific media consumption, in which users want to get the gist of the post. The variables suggest that the posts that resonate with users are the ones that convey the point of the post quickly. Posts with subfields that are more relevant or relateable may also contribute to the popularity of a post.

Data generalizability: something about how Reddit isn't representative of the population, but the results generally make sense given our experiences with scientific communication. seems wishy washy though because the extent of our project doesn't let us go into that much more detail… segue into next small section

What held us back: not being able to get more information from our posts, especially from the specifics of post title or even post content (because we did have access to the URL) –> sentiment analysis fell short, tackling the other problem would've required more time than we probably had

Concluding remarks: something that links abck to the motivating question… wax eloquence here