CSE574/CSE 474

Group [Elitecode Assignment 1]

1.1 Understanding APIs

1.1.1 How many API calls were required to collect the submissions?

30 API calls were required to collect the submissions.

1.1.2- Why did we set the submission limit at 1000?

Ans- We know that the default limit for retrieving things is set by reddit preferences, which is usually 25. The limit parameter can be used to change this. If we want to have as many Things as possible, we can set limit=None.

Since we can only get 1000 results from each listing due to an upstream limitation imposed by Reddit not PRAW. Any single reddit listing will display at most 1000 items.Therefore, we have set the submission limit at 1000. We have no way of going beyond this point.

1.1.3- How long, in minutes, would it take you to collect 1000 posts from 25 different subreddits? What about from 500 different subreddits?

Ans -We know that Reddit allows us to request up to 100 items at once. So, if our request is less than 100 items, PRAW can serve it in a single API call, but for larger requests, PRAW will split it into multiple API calls of 100 items each separated by a small 2 second delay to comply with API guidelines. As a result, requesting 250 items will necessitate three API calls and will take at least 2x2=4 seconds due to API delay.

Using this above logic, we solved for part a and part b as below

1) For 1000 posts from 25 different subreddits in total will be 25,000 reddit posts, which will require 250 calls. Thus time required: 249*2= 498 sec or 8.3 mins.

2) For 1000 posts from 500 different subreddits in total we are extracting 500,000 posts which will require 5000 calls. Thus, time required will be 4999*2 = 9998 secs or 166 mins.

1.2 Thinking about your sample

1.2.1 -Do you think these posts are representative of all the posts on that subreddit? (Yes or no, only)

Ans-No

1.2.2 Why or why not? That is, if you think so, why do you think there's not much sampling bias here? If not, what do you think might be different about these top posts than other posts?

Ans- We know that ,It is crucial to be aware of any potential biases that may exist in our data. If we are aware of these biases, we can account for them in the analysis to perform bias correction and gain a better understanding of the population that our data represents. For the below reasons there are potential reasons for the sampling bias here due to which the top posts are different than other posts.

a) Dangling references is the first type of gap we discovered. On Reddit, comments can only appear within the context of a submission's discussion and can only refer to other comments or submissions. In all cases, a submission must exist in order for a comment to refer to it, a unidirectional time relationship. These can be thought of as “known unknowns:” comments which refer to other comments or to a parent submission, where the referred-to comment or parent submission might not be picked in the 1000 posts.

b) There can be a case that objects that are never referenced in the API call are likely missing

c) The risks associated with missing data, risks associated with the uneven distribution of missing data over time, and risks associated with the uneven distribution of missing data across communities can lead to differences among the top posts than other posts and hence high chance of sampling bias.

Part 2.1 - Quick Descriptive Analysis

2.1.1 What are the names (subreddit_name_prefixed) of the 25 different subreddits that are in part2_data.csv?

Ans- 24 different subreddits that are in part2_data.csv

subreddit_name_prefixed
r/Jokes'
'r/news'
'r/science'
'r/WritingPrompts'
'r/Showerthoughts'
'r/worldnews'
'r/todayilearned'
'r/learnprogramming'
'r/announcements'
'r/funny'
'r/food'
'r/sports'
'r/gadgets'
'r/aww'
'r/mildlyinteresting'
'r/memes'
'r/technology'
'r/travel'
'r/books'
'r/gaming'

'r/cats'
'r/conspiracy'
'r/PoliticalHumor'
'r/hockey'

2.1.2 How many reddit authors (author_name) have a post in more than one unique subreddit in part2_data.csv (e.g. they have a top post in both r/news and r/hockey)?
Ans- 56

2.1.3 What is the mean number of upvotes (ups) for posts in r/Jokes?
Ans-41057.7813440321

2.1.4 What is the variance of the number of upvotes in r/news?
Ans- 600707867.6203133

2.1.5 What is the standard deviation of the number of upvotes received across the entire dataset?
Ans- 43101.54614748169

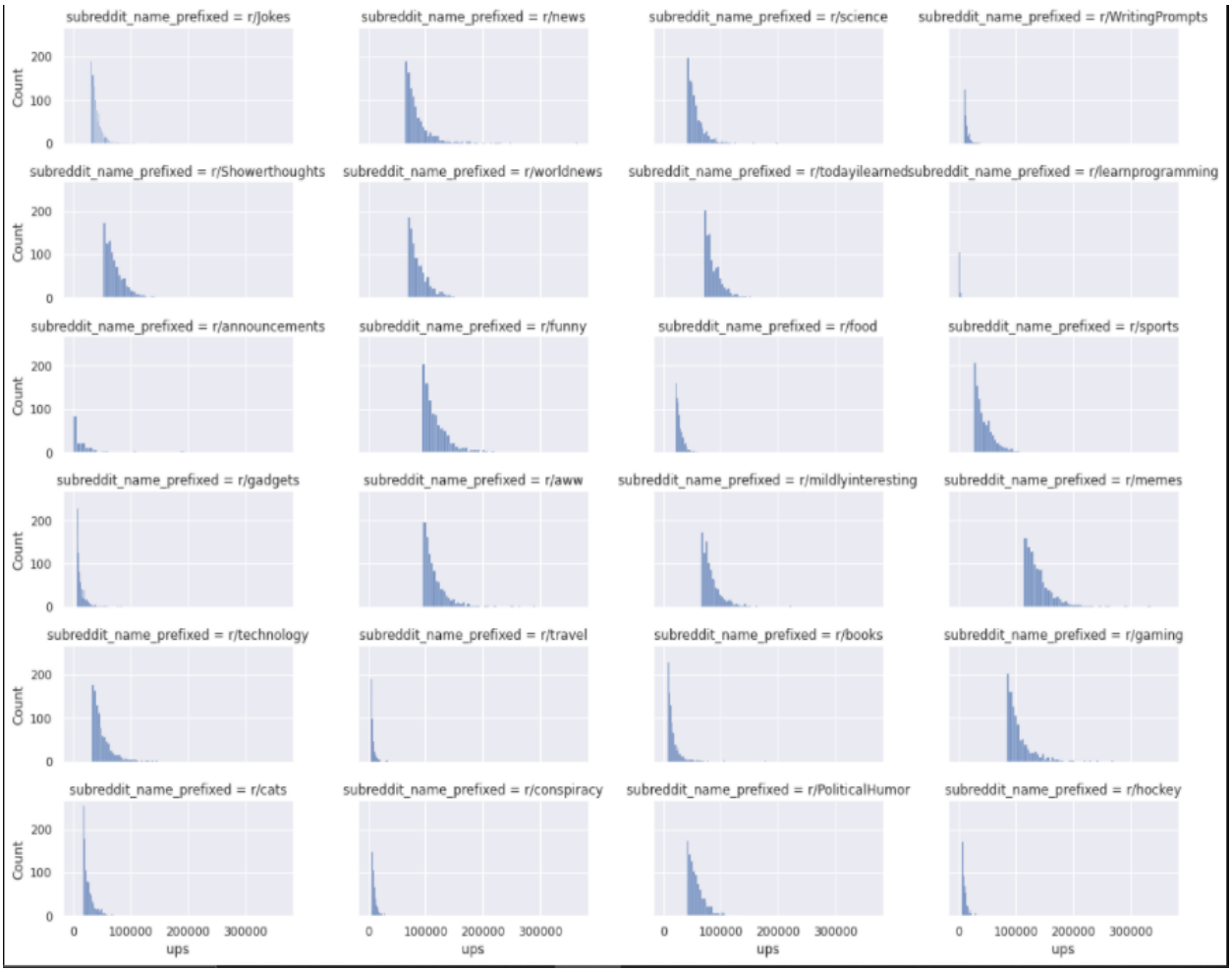
2.1.6 (No code for this) Mathematically, what is the relationship between the standard deviation of the number of upvotes and the variance of upvotes?
Ans – Variance of Upvotes = (standard deviation of the number of upvotes)^2

2.1.7 Which subreddit had the third highest median number of upvotes?
Ans- r/aww

2.1.8 What is the conditional probability of an author having a top post in r/news, given that they have a top post in r/worldnews?
Ans- 0.1009

Part 2.1 - Histograms

2.2.1 - Submit your histogram image in your assignment



2.2.2 - Based on your histogram, which subreddit would you say is the least popular? (Note, there is more than one reasonable answer here. We are looking mostly for how you justify your response using the histogram)

Ans- From the histogram we can see that ‘r/learnprogramming’ is one of the least popular subreddits.

Plotting and using the empirical CDF

2.2.3 - Approximately (within 1-2 percentage points) what percent of top posts for each of the three subreddits plotted below have less than 100,000 upvotes? (Give answers for each subreddit)

Ans- r/news--- 0.8410462776659959

r/worldnews--- 0.7907444668008048

r/science ----0.9848637739656912

2.2.4 - Approximately (within 1-2 percentage points) what is the probability that a post on each of the three subreddits plotted below has more than 70,000 upvotes? (Give answers for each subreddit)

Ans-

r/news -----0.7334004024144869

r/worldnews -- 0.9678068410462777

r/science --0.12613521695257315Temporal Trends

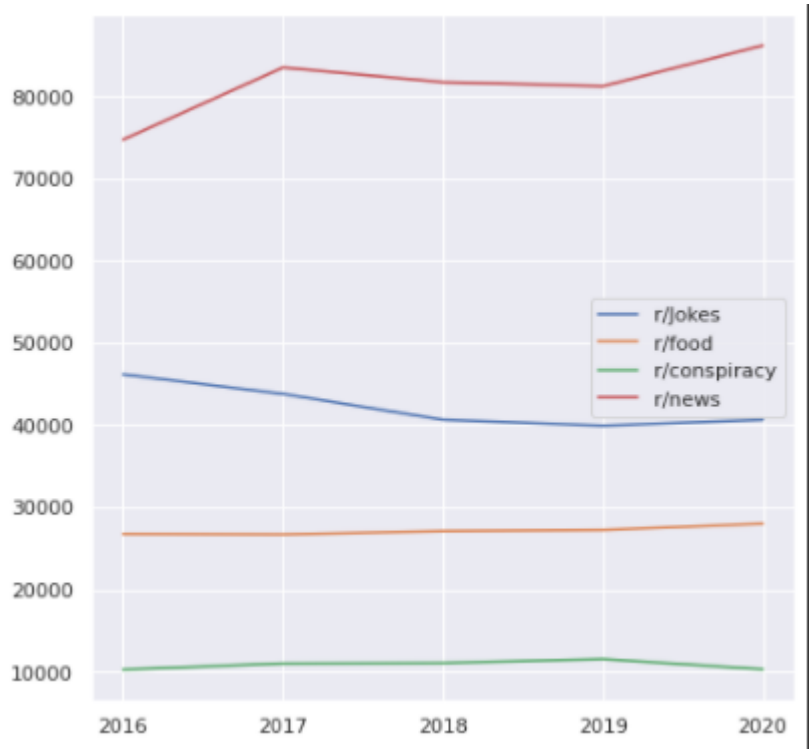
2.2.5 - How many posts in the dataset were sent in 2010?

Ans- 35

2.2.6 - In your report, provide a table (a screenshot of a pandas dataframe is fine) that shows the average number of upvotes for r/memes each year from 2015 to 2020. The table should be sorted by year (i.e. 2015, then 2016, etc.). Note again, if a year does not have data, there should be zeros in this table!

	year	subreddit_name_prefixed	ups
156	2015	r/memes	0.000000
177	2016	r/memes	0.000000
192	2017	r/memes	0.000000
235	2018	r/memes	131206.000000
255	2019	r/memes	135859.126984
272	2020	r/memes	141141.427305

2.2.7 - Plot a line graph of the temporal trend of mean upvotes from 2016-2020 for the following subreddits: r/Jokes, r/food,r/conspiracy, and r/news . You can plot them individually, or use the faceting approach from above. Write your code for this in the cell below; copy the resulting plot to your PDF report. Hint: Doing part 2.2.8 will be easiest if you make sure that the plot for each subreddit has its own y-axis!



2.2.8 - Using what you have plotted, make an argument for which of the four subreddits is the most "up and coming" - i.e. the one that seems to be getting more popular over time. NOTE: There is more than one reasonable answer here. We are looking for how you justify your answer using the (plotted) data.

Ans- r/news even though is already popular among the 4 still shows an upward trend thus can be called the most “up and coming” subreddit.

Part 2.3 - Data Cleaning and some final regression-oriented data exploration

2.3.1- There are two continuous variables that are very clearly not going to be useful for our analysis. Identify them, and explain why they are not useful (note: you do NOT need to know why these variables take on the values they do in our data. You just need to know why we don't want to use them!)

Ans- subreddit_id is not useful because we are using subreddit_name_prefixed, and num_reports is also not useful because it contains null values.

2.3.2- There are two (supposedly) binary variables that are very clearly not going to be useful for our analysis. Identify them, and explain why they are not useful (note: you do NOT need to know why these variables take on the values they do in our data. You just need to know why we don't want to use them!)

The two binary variables that are very clearly not going to be useful for our analysis are 'is_cross_postable' and 'media_only' because these only contain 'False' and thus do not add any information.

2.3.3 - Explain why we it is not useful to use both subreddit_id and subreddit_name_prefixed in any predictive analysis of per-post upvotes.

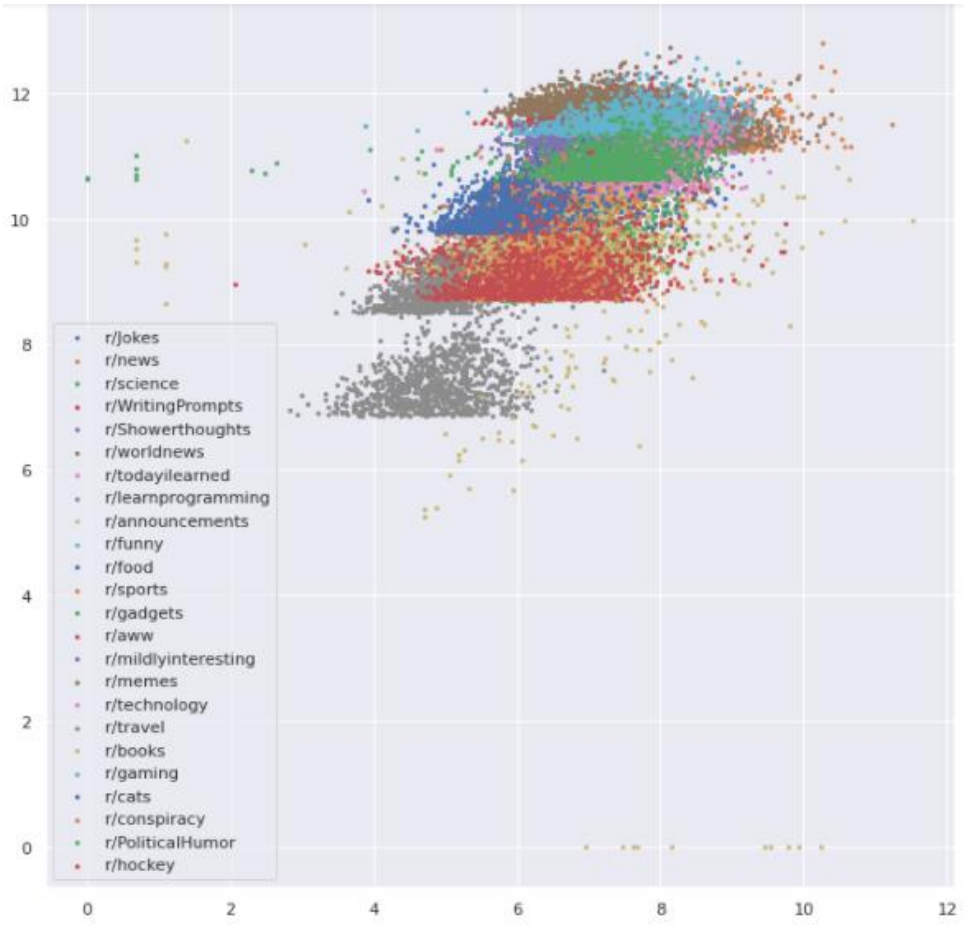
Ans -It is not useful to use both subreddit_id and subreddit_name_prefixed in any predictive analysis of per-post upvotes because both variables are used to identify a subreddit and hence using both is not required

2.3.4 - Explain why it is not useful to use permalink in any predictive analysis of per-post upvotes.

Ans – It is not useful to use permalink in any predictive analysis of per-post upvotes because the link of a post does not determine the number of upvotes it is going to receive

Univariate relationships with the outcome

2.3.5 - Plot the relationship between num_comments and upvotes as a scatterplot with log-scaled axes, with the posts from different subreddits as different color points. Paste this plot into your PDF writeup



2.3.6 - Describe, briefly (a sentence) the relationship between num_comments and upvotes.

Ans- num_comments and upvotes are directly dependent on each other. Thus, post with more upvotes tend to have more comments and vice versa.

2.3.7 - Which of these has the strongest positive correlation with ups?

Ans - 'score' has the correlation of one, if we disregard it then 'num_crossposts' has strongest positive correlation with ups

2.3.8 - Which of these has the weakest positive correlation with ups?

Ans- 'is_video' has the weakest positive correlation with ups

Questions to check understanding

3.1.1 - Report your error on the test data, in RMSE. State what this metric means for the expected error in terms of the number of upvotes (not log upvotes!) you should expect to be off on any given prediction

Ans- We achieved RMSE of 0.32. This means we can expect our predictions to off from the actual values by this amount in either direction

3.1.2 - What did the whole one-hot encoding thing on subreddit_name_prefixed actually do?

Ans- It resulted in the conversion of categorical features into numbers which can be processed by the machine learning models.

It is a common way of preprocessing categorical features for machine learning models. This type of encoding creates a new binary feature for each possible category and assigns a value of 1 to the feature of each sample that corresponds to its original category.

3.1.3 - What does the argument drop = "first" do for us when we are doing that to subreddit_name_prefixed?

Ans- We only need n-1 one hot vectors to represent a categorical feature with n values. So the 'drop="first"' argument removes one redundant column.

3.1.3 - Why did we need to add one to the outcome variable before using log?

Ans- We need to add one to the outcome variable before using log because many features have 0 as their values and log of 0 is undefined. To process those we add 1 before using log.

It also helps us keep the values which were >=0 as >=0

3.1.4 - What does the StandardScaler do? Why do we want to do that?

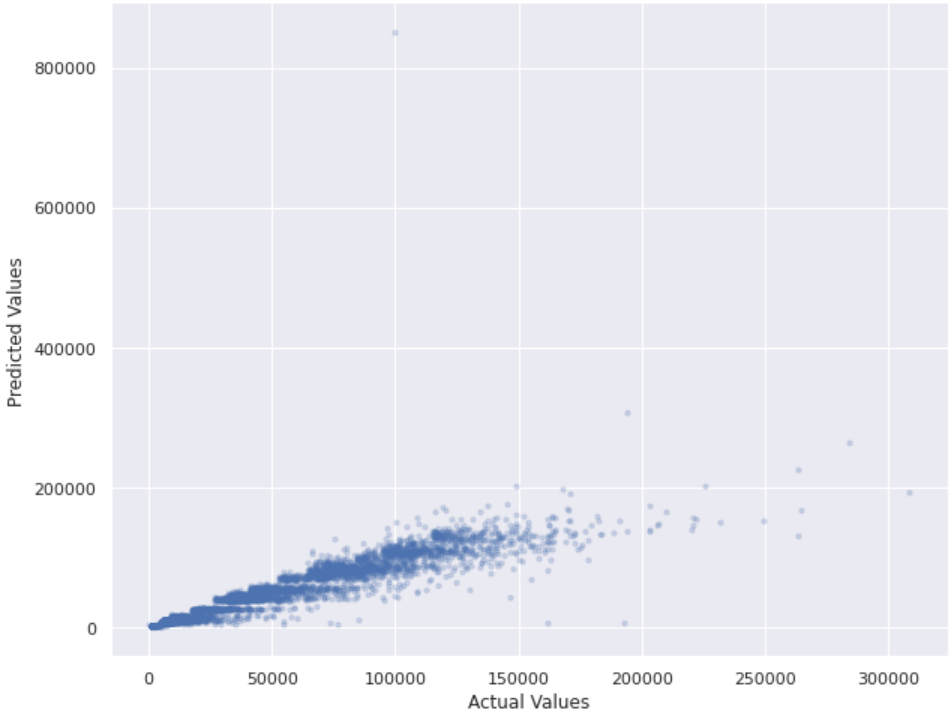
a) The StandardScaler function standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance is calculated by dividing all of the values by the standard deviation. StandardScaler produces a distribution with a standard deviation of one.

b) When the characteristics of the input dataset differ significantly between their ranges, or simply when they are measured in different units of measure, StandardScaler comes into picture. The mean is removed, and the data is scaled to the unit variance using StandardScaler. Outliers, on the other hand, have an impact on the calculation of the empirical mean and standard deviation, which narrows the range of characteristic values.

These differences in the initial features can cause problems for many machine learning models. For instance, for models based on the calculation of distance, if one of the features has a wide range of values, the distance will be governed by that particular characteristic.

The notion behind the StandardScaler is that variables that are measured at different scales do not contribute equally to the fit of the model and the learning function of the model could end up creating a bias. So, to solve this problem, we need to standardize the data ($\mu = 0$, $\sigma = 1$) that is typically used before we integrate it into the machine learning model.

3.1.5 - Provide a scatterplot that compares the true values in y_test to the absolute value of the difference between y_test and your predictions. The axes should be on the original scale (i.e. not the log scale you're predicting on.



3.1.6 - What does this plot suggest about how well your model fits the data as the true number of upvotes changes?

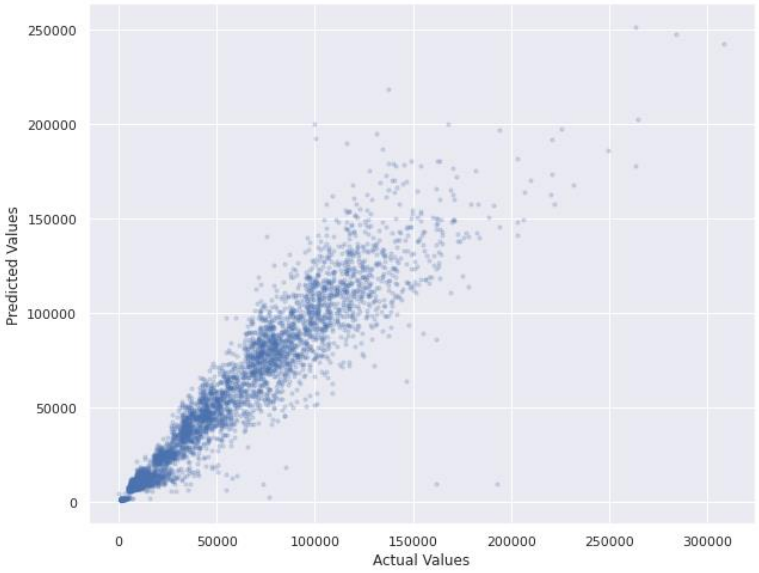
Ans -The plot suggests that our model does not fits well, and model predictions do not increase as the true upvotes increase

3.1.7 - What is the new RMSE with the logged independent variables?

Ans- 0.2967378402306904

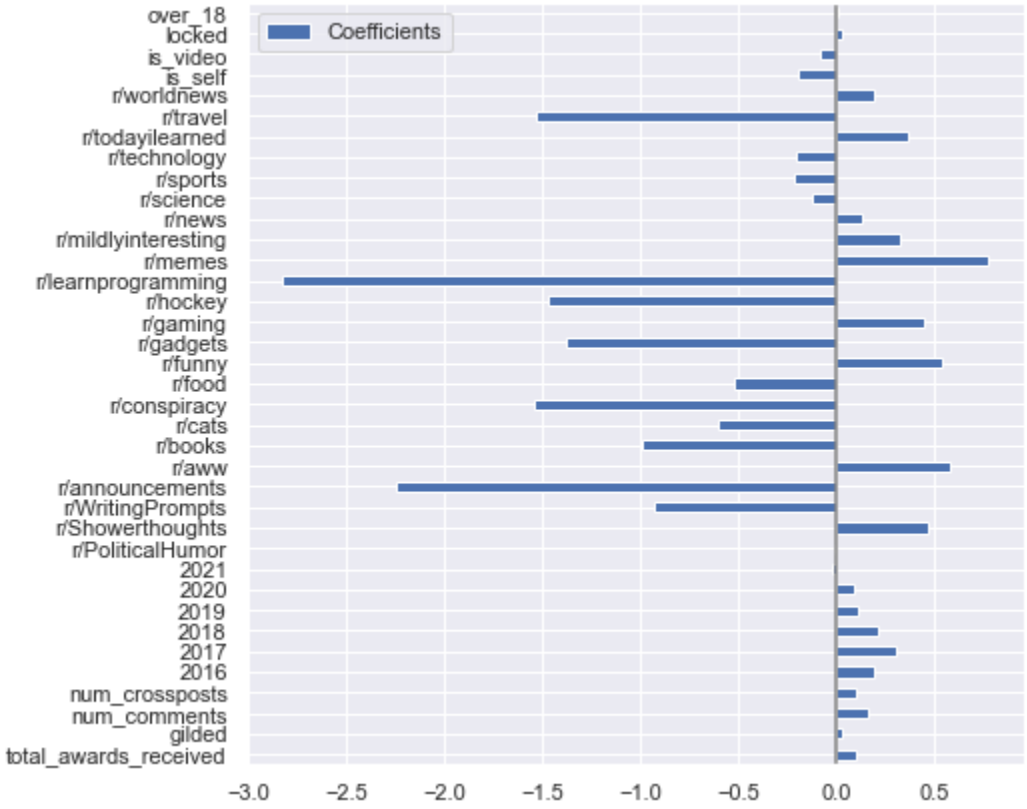
3.1.8 - How did this compare to the old RMSE? Why do you think that is? Hint: It may help to re-plot the same figure as you did in 3.1.5, but with the new model, in order to answer this question**.**

The newer model is better than the previous one as the RMSE value is less. The log scaling helps in reducing the non-linearity between features, which results in better RMSE value and since log reduces the dynamic range of the variable while not skewing the scale.



Part 3.2 - Exploration of regression coefficients

Below is the plot for our objective analysis



3.2.1 - What is the strongest positive predictor of upvotes? How many more log(upvotes+1) does a one standard deviation increase in the feature correspond to?

Ans- The strongest positive predictor is : r/memes. One standard deviation increase in the feature correspond to 0.779551 more log(upvotes+1)

3.2.2 - What is the strongest negative predictor of upvotes? How many fewer log(upvotes+1) does a one standard deviation increase in the feature correspond to?

Ans- The strongest negative predictor is : r/learnprogramming. One standard deviation increase in the feature correspond to 2.83246257 fewer log(upvotes+1)

Part 3.3 - 574 Only - Attempting to Improve Your Predictions

3.3.1 - Describe at least two changes you made – at least one to the feature set, and at least one different model – to try to improve prediction. Explain *why* you think that these changes make sense, given the Exploratory analysis above, or any other exploratory analysis you choose to do.

Ans -a) The two changes which we made are as below:

1) We changed our model to random forest & this made our RMSE to 0.2833 from 0.2967. We did this to use a more complex model than a linear model

2) We changed our dataset by adding X^2 and X^3 features. This resulted in RMSE of 0.2954 which is slight improvement. We did this to add non linearity in the data set.

3.3.2 - By how much did your RMSE improve? Which change that you made improved it the most? How do you know?

Ans – Initial RMSE which we achieved is – 0.2967

By using RandomForest model we were able to achieve 0.2833, and by adding X^2 and X^3 we achieved a slight improvement of 0.2954. We can be certain because no other thing was changed and hence the change made has to be attributed to the improvements.