**Machine Learning Assignment 1**

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

Answer: As per API Document, If the number of requests is fewer than 100, the PRAW API will complete the submission in a single API call.  As a result, we set a limit of 1000 requests each subreddit category, for a total of 3000 requests. As a result, 3\*10 API requests are required, totaling 30 API calls.

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

Answer: The Redditt API has a limit that permits us to only call 1000 requests at a time while collecting data.

**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? Hint: You'll have to consider how many API requests you are allowed to make in a given time period.**

Answer: 0.02 seconds

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

Answer: 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?**

Answer: The most popular Reddit posts are those that make the most impact and receive the most votes. As a result, during data extraction, only the most well-known and popular posts will be collected, leaving the bulk of postings ignored. As a result, I feel that these posts do not represent the entirety of that subreddit's information.

**2.1.1 What are the names (subreddit\_name\_prefixed) of the 25 different subreddits that are in part2\_data.csv?**

Answer:

1)'r/Jokes'

2)'r/news'

3)'r/science'

4)'r/WritingPrompts'

5)'r/Showerthoughts'

6)'r/worldnews'

7)'r/todayilearned'

8)'r/learnprogramming'

9)'r/announcements'

10)'r/funny'

11)'r/food'

12)'r/sports'

13)'r/gadgets'

14)'r/aww'

15)'r/mildlyinteresting'

16)'r/memes'

17)'r/technology'

18)'r/travel'

19)'r/books'

20)'r/gaming'

21)'r/cats'

22)'r/conspiracy'

23)'r/PoliticalHumor'

24)'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)?**

Answer: No of reddit authors that have a post in more than one unique subreddit in part2\_data.csv is “569”.

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

Answer: Mean number of upvotes (ups) for posts in r/Jokes is “41057. 78134”.

**2.1.4 What is the variance of the number of upvotes in r/news?**

Answer: Variance of the number of upvotes in r/news is “600707867.6203”

**2.1.5 What is the standard deviation of the number of upvotes received across the entire dataset?**

Answer: Standard deviation of the number of upvotes received across the entire dataset is “43102.48447”

**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?**

Answer:Variance is the square of Standard Deviation

**2.1.7 Which subreddit had the third highest median number of upvotes?**

Answer:The subreddit that had the third highest median number of upvotes ‘'r/aww'’ is 109811.0

**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?**

Answer:0.1009

**2.2.1 (3) - Submit your histogram image in your assignment**

A picture containing window, shoji, shrimp

Description automatically generated

**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)**

Answer: ‘r/ announcements' subreddit is the least popular among all subreddits.

**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)**

Answer: percentage of Top posts for each of the three sub reddits are:

* r/news - 84%
* r/science - 98%
* r/worldnews - 79%

**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)**

Answer: probability of Top posts for each of the three sub reddits are:

* r/news - 0.73
* r/science - 0.12
* r/worldnews - 0.96

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

Answer: 35 posts which were sent in 2010.

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!

Answer:

|  |
| --- |
| **year** | **ups** |
| **2015** | 0.000000 |
| **2016** | 0.000000 |
| **2017** | 0.000000 |
| **2018** | 65603.000000 |
| **2019** | 133736.328125 |
| **2020** | 140891.619469 |

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!.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

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.

Answer: r/news subreddit is one of the most up and coming .

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!)

Answer:

* downs because all the rows are zeros and we cannot give meaning information from this feature.
* num\_reports because all the values are null values we cannot perform data preprocessing with 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!)

Answer: is\_crosspostable and media\_only because they provide false values which are not useful for our Analysis.

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.

Answer: Subreddit\_id and subreddit\_name\_prefixed are both unique and refer to the same user. So, in order to eliminate repetition, we will select one of them.

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

Answer: Permalinks contain a unique url as well as the title of the post, which we don't need to use in our predictive analysis because we're only looking at the titles of the posts.

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

Answer:

Scatter chart

Description automatically generated

2.3.6 - Describe, briefly (a sentence) the relationship between num\_comments and upvotes.

Answer: The num\_comments and upvotes are directly proportionate, indicating the post's popularity.

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

Answer: num\_crossposts has the strongest positive correlation with ups i.e 0.538

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

Answer: created\_utc has the weakest positive correlation with ups i.e 0.1655

3.1.1 (5) - 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

Answer : 0.414169

3.1.2 - What did the whole one-hot encoding thing on subreddit\_name\_prefixed actually do?

Answer : The purpose of One hot encoding is to convert Categorical variable to Numerical data.

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

Answer: Dropping the dataset's first column is a feature of one hot encoding in which we drop the alpha numeric column. It helps in overcoming feature dropping's perfect collinearity.

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

Answer: StandardScaler scales a feature to unit variance after removing the mean. The standard deviation is divided by all of the values to get the unit variance. therefore, it makes mean = 0 and scales the data to unit variance.

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.

Answer:

Chart, scatter chart

Description automatically generated

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

Answer: From the above plot it shows the relationship between independent variable and dependent variable between Difference and Actuals.

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

Answer: 0.32925150236738066

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.

Answer: The old RMSE 0.4141690850780637 and after the converting independent variables into log scale we got the RMSE 0.32925150236738066

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?

Answer: the strongest positive predictor of upvotes is year 0.98

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?

Answer: the strongest negative predictor of upvotes is itself -0.1394

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 analyses above, or any other exploratory analysis you choose to do.

Answer: We used Lasso Regression Techniques, which is a feature selection approach that is used to improve the rmse score.

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

Answer: We considered Lasso Regression the rmse for lasso regression 0.5027935677984402 and the old Rmse is 0.414169. It is increased by 0.1.