Project Report

Reddit Submissions Analysis

Zachary Fong
301410034

Johnny Mai
301421429

Context:

Reddit is a social media platform where users can share diverse content including text posts, links, images, and videos. Users can interact with these posts by providing an upvote/downvote, which affects the score of the submission. Discussion within a post is possible through comments. An individual post belongs to a certain subreddit, which is usually centered around a particular topic.

Description of the Problem:

The general problem we wanted to address is what makes a Reddit submission good or bad? To refine this question further, we wanted to take the question from the approach of a potential user. What are some specific actions that a user can take to make their post more successful? What are some of the attributes that affect the performance of a Reddit submission?

Our Strategy:

To tackle these problems, we first decided to analyze the various attributes of an individual Reddit submission. Do well performing submissions have a longer/shorter title than posts that don't perform as well? What about the time in which the post was created? We also wanted to look at user behaviors associated with the Reddit submissions, specifically, the post frequency/count of an individual user within a subreddit. Are more active users generally creating better submissions?

In addition to the various attributes of individual submissions, we trained a linear regression model to predict the performance (score) of a submission. This allowed us to see how accurately we could predict a score with our features and how each feature affected the score using the model's coefficients.

Data Collection and Cleaning Process:

After taking a look at the initial Reddit dataset provided for use on the cluster, we realized that we had to narrow down the data into a few subreddits. The two main reasons for this were: the sheer size of the full dataset, and the unique nature of each subreddit. Some subreddits had complex post content that we wouldn't be able to fully analyze and would likely have a high impact on its relative performance. For example links, images, or videos. As a result, we decided to stick to the 3 most popular subreddits that contained only text content: AskReddit, Jokes, and Showerthoughts. We also further subsetted the data to include only posts made during the year 2019, which was a decision made after realizing that the data collection process excludes a "retrieved_on" column after around April 2020. This column would have made it useful to determine the "age" of a submission (amount of time a submission was present on the platform). We also did not want any abnormalities in submissions which could result from the Covid19 Pandemic.

From the results of the 0-extract.py Spark script described above, we looked at the columns of the data to understand each of its purposes. In doing so, we noticed that a majority of the columns present in the provided schema of the Spark extraction script were missing. From this

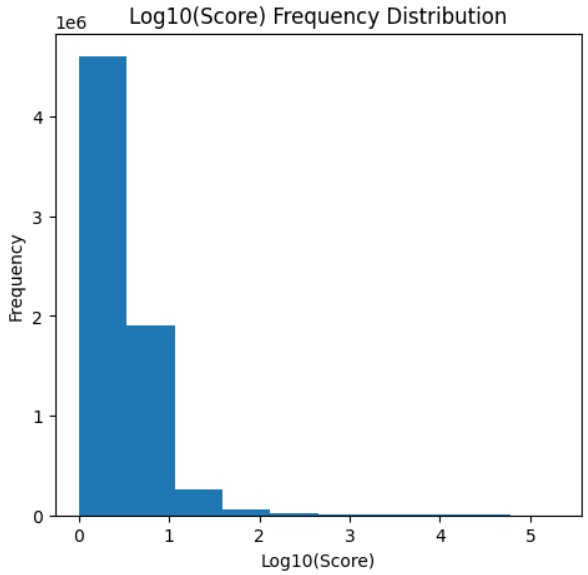
observation, we decided to further filter the data subset (with 1-filter.py) and only select the columns in which were relevant or were consistently present in the data.

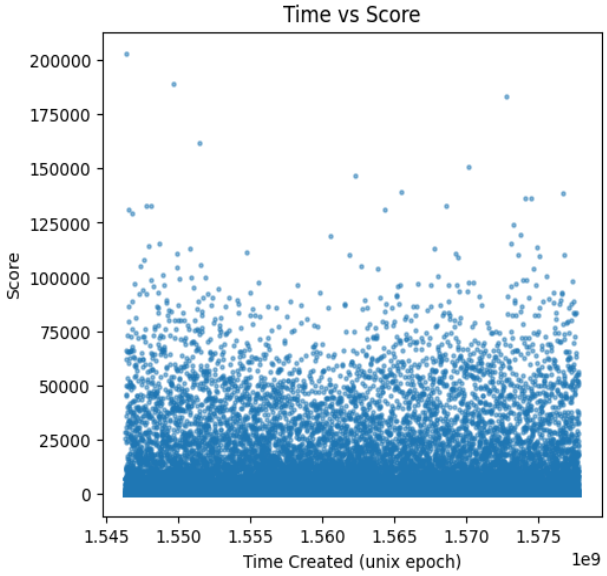
After filtering, we realized that the raw features we thought were important needed to be transformed to be usable (using 2-transform.py). We first wanted to apply sentiment analysis to the feature titles, but we quickly realized with the amount of data we had, that it would be computationally expensive. So, we decided to use the length of the title instead. For time analysis purposes, we extracted numerical values out of the creation timestamp for information such as the hour, day of the month, and day of the week. For user post frequency analysis, we extracted the number of posts a user created within a subreddit using a group by and an aggregate operation followed by a join. Lastly, we converted boolean values like over_18 and archived to binary values like 0 and 1 for the later purpose of training our regression model. The final result of this process was a dataset consisting of over 6.8 million submissions across three subreddits for the year 2019.

Data Analysis Techniques:

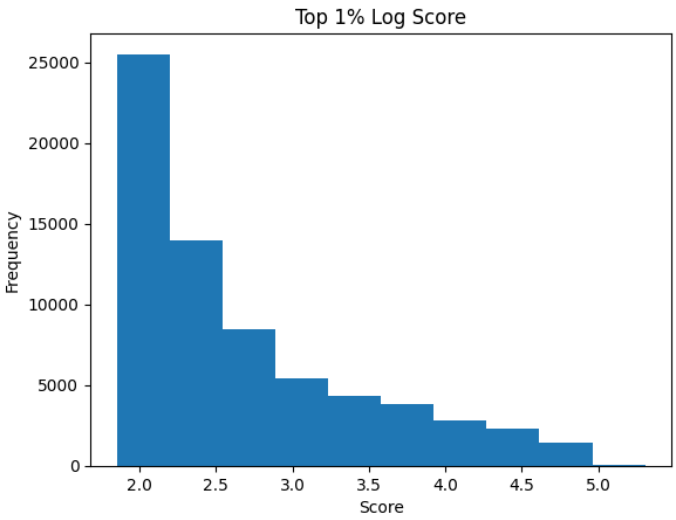
The resulting compressed json output of the transformation script (2-transform.py) was around 300 MB. We decided to move off of Spark at this stage and onto our local machines to easily visualize our data with Pandas and Matplotlib. We made this decision since 300 MB easily falls within the memory limitations of our devices.

Initial analysis of the data consisted of overall summary statistics and visualizations (with 3-initial_analysis.py). We wanted to plot the frequency distributions and scatter plots for various attributes of our data.





From these visualizations and summary statistics, we quickly realized that our data was massively right skewed. There was a large concentration of Reddit submissions with scores close to 0. In fact, a majority of the submissions we collected could be considered “not good”. According to the percentile statistics, 99% of our submissions have a score less than 70.



Even within the top 1% of all our submissions, the distribution of scores is still massively right skewed. To somewhat mitigate the skewness of our data, we decided to split the data into two main subgroups, submissions with above average and below average score of the dataset. The below average submission group was obviously the majority group, so we decided to randomly subsample from the majority until we had an equal amount of records in both groups.

To answer questions regarding various attributes of reddit submissions and their effect on score performance, we conducted Mann Whitney U statistical tests. This specific test was chosen due to the non parametric nature and the lack of requirements for normality or equal variance. Between the above average and below average submission groups, we looked at attributes regarding the length of a submission title, the creation time, and the post frequency of a user. We wanted to know whether the observations of attributes seen in one group are larger than in the other. Specifically, the null hypothesis for the tests being “the sort order of observations between the two submission groups are roughly equal”. And the alternative hypothesis as “the sort order of observations between the two submission groups are not equal”. Rejecting the null would suggest that the attributes in question have some effect on the submission score.

To gather even more results from this dataset, we also created an ML model. We used linear regression by converting the features into vectors with VectorAssembler(), then scaled them using StandardScaler(). We tested the models accuracy by splitting the data into testing/training data and calculating the root mean squared error and mean absolute error. To compare our error, we also calculated the mean absolute error from a model that always predicted using the mean score value. This functioned as our base case. After training this linear regression model, we were able to extract the coefficients to find the weighting assigned to each feature and whether the feature positively or negatively impacted the score.

Findings:

The results of our Mann Whitney U statistical tests are as follow:

Question	P Value	Conclusion (for a = 0.05)
Are title lengths different between above average submissions vs below average submissions?	4.715424283005718e-289	Reject the null, there is a difference in title length between the two groups.
Do users who create above average submissions have a different post frequency than users who create below average submissions within AskReddit?	0.7195124362627138	Cannot make a conclusion about the difference in post frequencies.
Do users who create above average submissions have a different post frequency than users who create below average submissions within Jokes?	1.226486215502343e-72	Reject the null, there is a difference in post frequency between the two groups.
Do users who create above average submissions have a different post frequency than users who create below average submissions within ShowerThoughts?	2.875269787922374e-44	Reject the null, there is a difference in post frequency between the two groups.
For above average posts, are the scores for posts created during working hours (between 9am and 5pm) on weekdays different from posts made outside these hours?	2.5853368743875955e-09	Reject the null, there is a difference in score between the two groups.

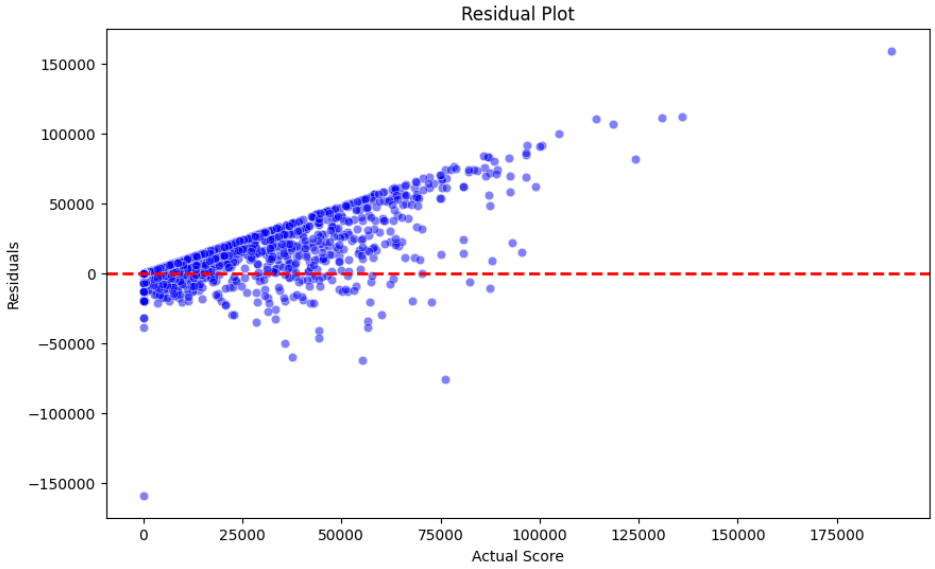
While interpreting the results, it is important to understand that the data samples used for each statistical test is a biased subsample of the original dataset. Therefore, the conclusions made cannot be applied to the original data.

Results from running linear regression on all features, top 2 features, and using the mean score as predicted value:

	All features	Num_comments and gilded	Compare against mean_score
Root mean squared error	850.8500810499847	793.1119301859801	-
Mean absolute error	50.83035424247715	43.47686890947218	67.95221037300425

From the table above that describes the errors of each prediction method, we found that the model using the top 2 features predicted the score of a post more accurately than the model using all the features. We also found that using the model with the top 2 features performed roughly 1/3 better than predicting using the mean score alone.

Scatterplot of the residuals vs the actual score of the post from the model using all features:



To visualize the performance of the model, we created a scatter plot where the blue dots represented how much error the prediction had in relation to the score and the dotted red line represented an accurate prediction. What we can gather from this plot is that as the post grows in score, so does the error. It's interesting to note that the model regularly overpredicts the success of a post rather than underpredict.

Model coefficients from linear regression on the dataset with all features.

Feature	**Coefficient**
created_on	16.29070932261182
age	3.1010997485409026
month	-19.880368539351174
day	-1.4677736175693687
hour	-5.586812713386523
day_of_week	0.31342902122267635
post_count	-8.84975473377815
over_18	-0.022429246904766803

gilded (number of awards)	327.2248991128074
archived	-0.1529727661698511
stickied	0.7659805780092003
num_comments	530.5186587717683
title_length	16.301779165735464

From the table above of the features and their coefficient from linear regression, we found that the most impactful features of a post were the number of comments they had and the number of awards they had. Most of the other features had a very small weight in comparison.

Conclusion:

The initial questions that we introduced were “What are some specific actions that a user can take to make their post more successful?“ and “What are some of the attributes that affect the performance of a Reddit submission?”

Despite the conclusions within the findings of our statistical tests, we cannot confidently recommend a set of actions a user can take to improve the score of their submissions. This is due to the fact that we had to use a biased sample of the original dataset to conduct the tests, which we believe would not be representative of the population.

As for the second question, we can say that the attributes that impact a posts score the most are the number of comments and awards it receives. We were also able to create a model that made predictions 1/3 better than with the mean_score prediction alone.

Limitations:

Dealing with the massive skew in distribution of submission scores within our dataset was a massive problem. A large portion of our data had score values concentrated at or near 0. Additionally, we had very infrequent submissions that had extremely high score values. We could have considered these extremely high performing submissions as outliers, but we wanted to include them within our analysis so as to not reduce the amount of data we had for good submissions. In retrospect, we could have better defined and split our initial dataset into portions that could be more usable.

In regards to the model, we were able to predict the score of a post more accurately than using the mean alone, however they had a mean absolute error of 43 at best. This error is larger than the mean score of roughly 38, which means our model’s performance is **very poor and its predictions should not be viewed as accurate whatsoever**. A better approach may have been to engineer the date features differently, like having a boolean weekday/weekend feature, or to sample the dataset differently to draw more meaningful conclusions and focus on popular posts.

Accomplishment statement (per person)

Johnny Mai

Extracted Reddit submissions data off of cluster using PySpark and Hadoop to begin data analysis process. Explored the collected data and generated visualizations and summary statistics using Pandas and Matplotlib in Jupyter Notebook to understand various features of 6.8 million Reddit submissions. Conducted statistical tests using Scipy stats to evaluate relationships between submission attributes and performance.

Zachary Fong

Discovered and evaluated the most impactful attributes of a Reddit post affecting its score. Performed ETL processes on Reddit data located on a remote cluster using Hadoop, Spark, and Python. Visualized refined datasets and applied statistical methods like Mann-Whitney U to analyze refined datasets using Jupyter and Pandas. Created linear regression models to predict a Reddit post's score 35% more effectively than using the mean score alone.