Reddit Submissions Analysis

Zachary Fong[a], Johnny Mai[1]

[a]*Simon Fraser University, Canada"*

Abstract

The paper "Reddit Submissions Analysis" is a project report that focuses on understanding factors contributing to the success of Reddit submissions. The authors, Zachary Fong and Johnny Mai, examine various aspects of Reddit submissions, including title length, submission time, and user behaviors. They employ data analysis techniques like linear regression and Mann Whitney U statistical tests to investigate these factors. The study is concentrated on popular text-only subreddits and uses data from 2019. Key findings include insights into the impact of title lengths, submission frequencies, and submissioning times on submission scores.

1. Introduction

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

2. Problem Description

Understanding what contributes to the success of submission on Reddit can often be a complex problem, especially to new users. This paper seeks to explore "What makes a Reddit submission successful?" and "Can we predict the score of a Reddit submission?". The goal of this exploration is to examine how submission components, from content attributes to user engagement, influence a submission's popularity and whether they can act as predictors for the success of a submission.

3. Approach

To address these questions, this paper starts by examining the various attributes of individual Reddit submissions from the Reddit corpus. These attributes include the title length of submissions and the temporal aspects of a submission. Additionally, this paper explores user behavior, with a focus on how submission frequency of individual users within a subreddit affects submission performance. The objective of this is to discern whether increased user activity correlates with the creation of higher-quality submissions.

In conjunction with the analysis of individual submissions, linear regression models were employed to predict the performance (score) of a submission. This enabled the evaluation of a basic model that used the attributes available on the Reddit corpus, as well as the degree of influence each attribute had on the score by viewing the model coefficients.

4. Data Processing Pipeline

After reviewing the initial dataset from the Reddit corpus, it was clear that a data processing pipeline was required. As a start, specific subreddits needed be chosen from the original dataset. The two main reasons for this were: the sheer size of the full dataset, and the unique nature of each subreddit.

4.1. Choosing Subreddit

Reddit houses a multitude of unique subreddits, each of which contain diverse content that has a high impact on a submissions performance. The inclusion of diverse content, such as links, images, or videos in submissions could pose challenges when analyzing it. Consequently, a decision was made to include only the most popular subreddits with text-based content from the Reddit corpus. These subreddits are: AskReddit, Jokes, and Showerthoughts.

4.2. Time Period Selection

After filtering for these chosen subreddits, an investigation into this data subset found null values in the "retrieved_on" column after April 2020. As this column would have been required to determine a posts age (how long it had been posted before being added to the Reddit corpus) and because the Covid19 pandemic started in early 2020 (which could cause abnormalities in submissions), only submissions made during 2019 were included. Submissions made before 2019 were not included as this paper aims to evaluate submissions in a similar time period.

4.3. Data Exploration

After processing the Reddit corpus through these initial stages of the pipeline using 0-extract.py, analysis began on each column of the new dataset to understand what information they held. During the analysis, it was found that a majority of the columns in this dataset were null or only contained a single value. From this observation, the processed dataset was further refined to remove these columns using 1-filter.py.

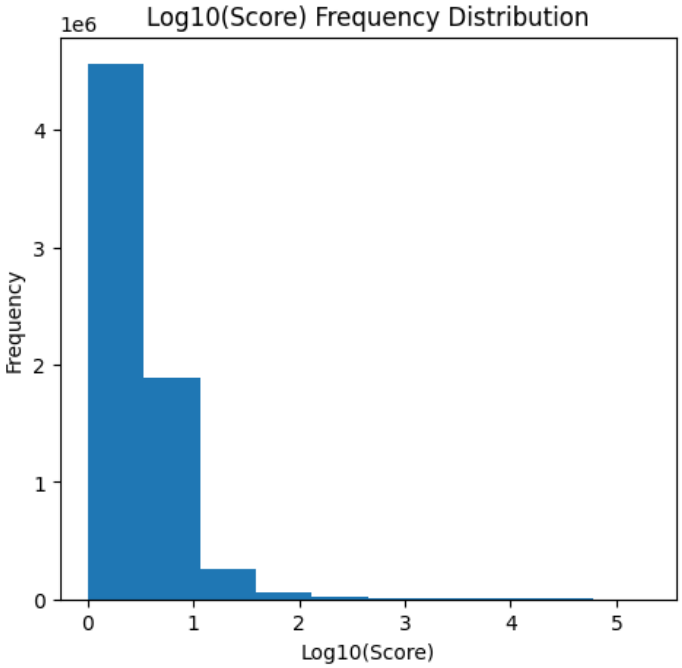

Figure 1: Frequency distribution of submission scores

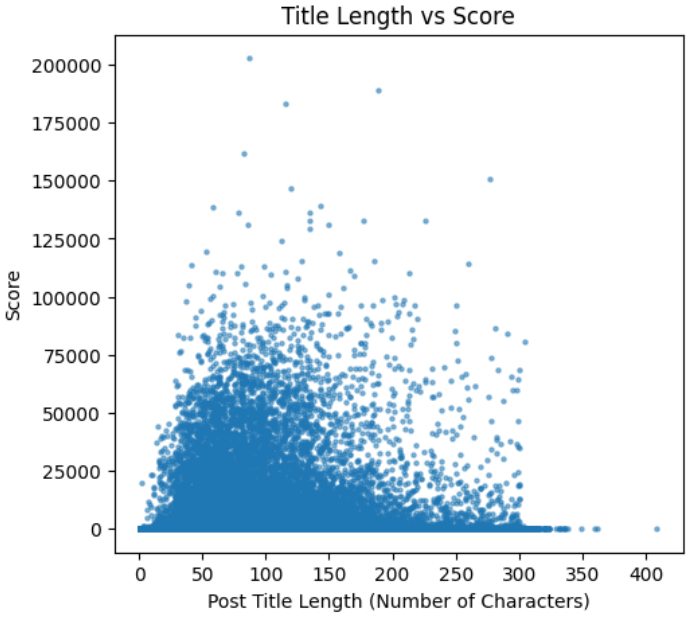

Figure 2: Scatterplot of submission title length vs score

4.4. Feature Transformation

After filtering, further review of the raw features intended for analysis and modelling indicated that transformations to them would be required. To generate temporal features for a submission, the hour, day of the month, and day of the week were extracted from the creation timestamp. For user submission frequency, the number of submissions a user created within a subreddit was extracted using a groupby and an aggregate operation, followed by a join. Boolean values like "over_18" and "archived" were converted to binary 0 and 1 values for false and true respectively. An initial goal during feature transformation was to apply sentiment analysis and text embeddings to each submissions title, generating two new features. However, after a small batch test run, the time to compute these values and store them were too large. So, the length of the title was used instead. The final result of this process was a dataset consisting of over 6.8 million submission across three subreddits for the year 2019. This was done by applying 2-transform.py to the dataset created from 1-filter.py

5. Data Analysis

After passing through the data processing pipeline, the resulting compressed gzip dataset was under 400 MB. Because of its small size, the dataset was moved off of the remote cluster initially used, and onto local machines. This allowed the data to be easily visualized using tools like Pandas and MatplotLib without the data partitioning or overhead that Spark and the HDFS applied.

5.1. Initial Observations

Initial analysis of the data consisted of summary statistics and visualization available in 3-initial_analysis.py. These visualizations included, frequency distributions and scatter plots for various features as seen in Figure 1 and Figure 2. These summary statistics and visualizations, specifically Figure 1, revealed a massively right skewed dataset. It showed an extremely large concentration of submissions with scores close to or at 0. In fact, 99% of the submissions in this dataset scored lower that 70. In comparison, the highest score for a submission was over 200,000. Unsurprisingly, the visualization in Figure 2 showed that there were more posts with shorter title lengths. It also showed that the number of high scoring submissions with a short title length exceeded the number of high scoring posts with long title length.

5.2. Stratified Sampling

To gain insights on what separated a high and low performing submission, quantile-based stratified sampling was applied to the dataset. This created two stratum: submissions with above average scores and submissions with below average scores. These stratum would represent successful and unsuccessful submissions respectfully which would then be used for statistical tests.

Feature	datasets	Subreddit	P Value	Conclusion (for a = 0.05)
Title Length	Stratified sampling datasets	all	4.715424283005718e-289	Reject the null
Post frequency of an author	Stratified sampling datasets	AskReddit	0.7195124362627138	Fail to reject the null
Post frequency of an author	Stratified sampling datasets	Jokes	1.226486215502343e-72	Reject the null
Post frequency of an author	Stratified sampling datasets	Showerthoughts	2.875269787922374e-44	Reject the null
Submission date/time	Above average stratified sampling dataset split into work hours/days and non-work hours/days	all	2.5853368743875955e-09	Reject the null

Figure 3: Analysis of data using Mann Whitney U tests

	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

Figure 4: Errors from all models

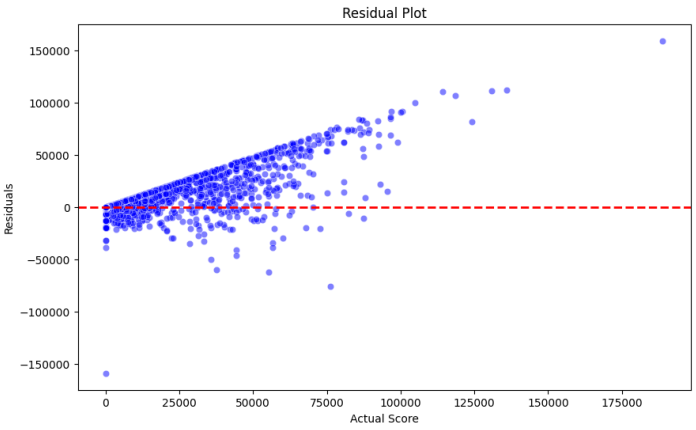

Figure 5: Residuals from a linear regression model using all features

5.3. Statistical Tests

As a main goal of this paper was to identify what factors affect the performance of a submission, statistical tests were applied to assess the statistical significance of these factors. The statistical test chosen to evaluate the stratum was the Mann Whiteny U test . This test was chosen as it was non-parametric and did not require normality or equal variances. Between these two stratum, title length, creation time, and submission frequency of a user were evaluated. This test would find whether the distribution of attributes in one group were different than the attributes in another. The results of this test on the stratum can be seen in Figure 3.

5.4. Initial Modelling

While the Mann Whitney U tests could assess whether a significant difference between factors in the stratum existed, it was limited in quantifying the influence by these factors. Using linear regression, it was possible to see how accurately a model could predict a submissions score and more importantly what coefficients the model was using. For this portion of the project, three models were created in predict_score_old.py. One linear regression model utilized all features from the Reddit corpus post data processing (model 1), one linear regression model utilized the three most influential features from the previous linear regression model (model 2), and the last model used a mean dummy regressor functioning as a baseline (model 3). These linear regression models were created with a pipeline that converted the features into vectors with VectorAssembler(), then scaled using StandardScaler(). The models were then trained with a 80/20 split and measured using RMSE and MAE. The resulting errors were then compared as seen in Figure4, the residuals for model 1 displayed in Figure5, and model 1's coefficients documented in Figure6.

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

Figure 6: Coefficients from linear regression model using all features

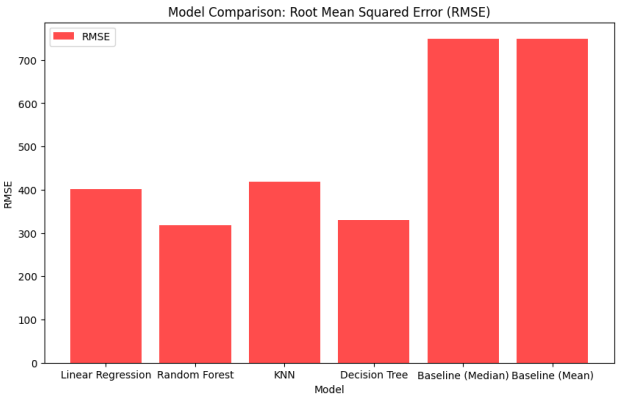

Figure 7: Coefficients from linear regression model using all features

5.5. *Final Modelling*

The results of these initial linear regression models, while better than the mean dummy regressor, were quite poor. So after review, 4 additional models were created in predict_score_new.ipynb. These new models were: Random Forest regression, KNN regression, Decision Tree regressor, and an additional baseline median dummy regressor. These models were chosen as research after the initial modelling stage suggested they outperformed other models in similar projects. While the same models were also included in this final modelling section, the dataset it used only contained the AskReddit subreddit because of reduced computing power available. These additional models were trained and hyperparameter tuning was applied where possible with the computing power available.

6. Findings

6.1. *Statistical Tests*

After conducting the Mann Whitney U tests on the subsets derived from quantile-based stratified sampling, a number of findings can be made. Title length of a submission was found to be statistically different between the two stratum, meaning that title length had an impact on the success of a submission. The post frequency of an author within a subreddit also was found to be statistically different between the two stratum, but only in Jokes and Showerthoughts. Surprisingly, the Mann Whitney U test for post frequency in AskReddit failed to reject the null hypothesis, resulting in no conclusion. However, the most interesting result is the Mann Whitney U test on the date/time feature of altered datasets. These datasets were created by taking the above average strata and splitting it into two subsets. One containing submissions made within working days/hours and one containing submissions made outside of working days/hours. The assumption behind this test was that users may interact differently on Reddit whether they are working or not. The test resulted in a rejection of the null hypothesis, meaning that above average posts were influenced on whether they were posted during work hours or not.

6.2. *Modelling*

During the inital modelling stage, despite the linear regression models having high error, the model coefficients were calculated and could be used to check the influence of each feature. Unsurprisingly, the most influential features were the number of comments and number of awards a submission had. These are both attributes that come as a submission gains popularity, so this discovery is not that revealing. The next most influential features, however less influential in comparison, is the month when posted and title length. The coefficient for the month posted was negative, meaning predictions for more recent posts were lower, and the coefficient for title length was negative, meaning predictions were higher for submissions with longer titles.

After the final modelling stage, a model with significant improvements over the inital modelling stage was found. This model was a random forest regressor and achieved an R^2 value of 82%, with 76% lower MAE and 58% lower RMSE scores when compared to the mean dummy regressor. It also achieved 15% higher R^2, 65% lower MAE, and 21% lower RMSE values when compared to the original linear regression model approach. While this model is much more accurate than both the linear regression model and the mean dummy regressor model, it took much longer to train. The RMSE of all models can be seen in Figure 7. It may be more valuable to look at RMSE over MAE as it's more sensitive to large errors when compared to MAE.

7. Conclusion

The main goal of this paper was to find what made a Reddit submission successful and if the score of a Reddit submission could be predicted. Based on the statistical tests, title length, post frequency, and whether a submission was posted during working hours, all affect a submissions success. Based on the inital linear regression models coefficients, a recommendation could be made to create submissions that encourage comments and awards, that are posted earlier in the year, and that have long title lengths. In terms of predictability, this paper trained a random forest model that had significantly lower error and significantly higher R^2 values when compared to a mean dummy regressor. So, attributes were found that affected a Reddit submissions success and models were created that could accurately predict the score of a Reddit submission, meeting the goals of this paper.

8. Limitations

Although the Mann Whitney U tests found statistically significant evidence for differences in the distributions of certain attributes, it did not find the degree of this difference. Instead, this paper relied on a linear regression model with moderately low accuracy to predict this degree through model coefficients. Because of this, it would be acceptable to trust the statistical conclusions from the Mann Whitney U tests, but not acceptable to trust the conclusions made from the linear regression

model coefficients. Additionally, when considering the predictions from the regression models, although the most accurate model significantly outperforms the mean dummy regressor, its residual errors are still occasionally large. These models also heavily rely on the number of comments and awards, which often both come with the success of a post, and is not something a submission author can readily control.