Mining Reddit: Data Driven study to forecast post scores using Machine Learning

Dhruvil Gajaria, Prashansa Dhomane
{dgajaria,pdhomane}@ucsd.edu
University of California, San Diego

Abstract
Social Media platforms generate million terabytes of data every day. This data when analyzed can yield insights into user behavior and content engagement. This research paper explores the development and evaluation of predictive models for Reddit submission scores. Reddit is a popular social media platform where users can submit posts, and each post receives a score based on user interactions. Leveraging a dataset of 132,303 Reddit entries, we employ various machine learning models and techniques to predict submission scores based on features such as post metadata and sentiment analysis of post titles. Our analysis includes linear regression, random forest, XGBoost, and support vector regression (SVR) models. The findings from this analysis not only provide a predictive framework for post popularity but also offer a window into the collective behavior of Reddit users. This work has implications for content creators and marketers seeking to enhance their visibility and engagement on the platform.

1. Introduction

In the era of digital communication, social media has been integral to the way information is shared and consumed. Reddit is distinct from the rest of these platforms because of its user-generated content and distinctive layout. Reddit, sometimes referred to as "the front page of the internet," offers a wide variety of user-generated material, such as debates, reviews, news, and entertainment.

Users can upvote or downvote posts on Reddit, creating a dynamic ecosystem where content exposure is determined by community acceptance. This rating system affects a post's reach and engagement in addition to determining its popularity. Predicting a Reddit post's score thus becomes a subject of great interest for social media analytics professionals as well as for content providers looking to increase their visibility. Through the lens of machine learning, we explored a multitude of factors that contribute to the popularity of Reddit posts, including temporal aspects, content features, and community dynamics.

This project aims to investigate the complex dynamics of Reddit content popularity, specifically how the same content can receive varied levels of engagement when reposted to different communities, at different times, and with different titles. Our analysis focuses on contributions made to reddit.com, where each post usually consists of an image along with a title, which is then uploaded to a certain subreddit at a given time and voted on by the community with upvotes and downvotes.

One essential aspect of Reddit is the scoring system, which assigns a numerical score to each submission. This score reflects the post's popularity and is determined by factors like the number of upvotes, downvotes, and comments. The study looks at several variables, including posting time, subreddit, and user engagement data, in an effort to identify the key elements that influence a post's effectiveness.

Understanding the drivers behind submission scores can be beneficial for both content creators and platform administrators. Predictive models can help identify the features that influence a post's success, enabling users to optimize their content strategy. Moreover, platform administrators can use such models to detect anomalies or potentially harmful content.
In this paper, we aim to build and evaluate predictive models for Reddit submission scores. We consider a range of features, including

metadata such as the time of submission, the subreddit it belongs to, and sentiment analysis of post titles. We employ machine learning algorithms to develop predictive models and assess their performance using metrics like Mean Squared Error (MSE) and R-squared (R^2).

The following sections will describe the dataset in detail, outline the predictive task and the chosen features, discuss the modeling approach, review relevant literature, and finally, present and analyze the results of this study. Through this comprehensive approach, the paper aims to contribute to the growing field of social media analytics and provide valuable insights into the dynamics of user engagement on Reddit.

2. Dataset Description

The dataset under examination is sourced from Reddit. It is widely known for its community driven content and comprises a large array of user generated posts, each associated with specific subreddits and topic-focused communities.

2.1 Key Characteristics

Volume: The dataset includes a total of 132,303 entries each representing individual post on Reddit

Unique users: 63334

Timeframe: The dataset spans from 2008-07-28 04:26:54 to 2013-01-25 06:27:33

Attributes: Each entry in the dataset contains 13 distinct attributes, offering a multidimensional view of the posts

Attribute	Description
#image_id	A unique identifier for the image associated with the post.
unixtime and rawtime	Timestamps marking when the post was created, presented in both Unix time format and a more human-readable format.
title	The title of the post, as crafted by the user.
total_votes, number_of_upvotes, and number_of_downvotes	Metrics indicating the level of engagement a post received, in terms of user voting.
subreddit	The specific community (subreddit) to which the post was submitted.
score	A calculated metric representing the post's net popularity (upvotes minus downvotes)
number_of_comments	The count of user comments on the post, indicating the level of discussion it generated.
username	The Reddit username of the post's author.

Table 1: Description of each attribute in the dataset

	#image_id	unixtime	total_votes	number_of_upvotes	number_of_downvotes	localtime	score	number_of_comments
count	132303.000000	1.323020e+05	132302.000000	132302.000000	132302.000000	1.323020e+05	132302.000000	132302.000000
mean	10260.296441	1.340019e+09	1883.245748	1058.182900	825.062849	1.340036e+09	233.120051	39.063400
std	7317.407802	1.294615e+07	5970.693071	3181.146483	2796.540029	1.294027e+07	481.126196	142.742878
min	0.000000	1.217219e+09	0.000000	0.000000	0.000000	1.217244e+09	-264.000000	-1.000000
25%	3807.000000	1.333785e+09	15.000000	8.000000	6.000000	1.333811e+09	2.000000	0.000000
50%	9041.000000	1.344574e+09	45.000000	30.000000	14.000000	1.344599e+09	16.000000	3.000000
75%	16021.000000	1.348950e+09	376.000000	276.000000	96.000000	1.348952e+09	169.000000	15.000000
max	25887.000000	1.359095e+09	177103.000000	90396.000000	86707.000000	1.359095e+09	20570.000000	8357.000000

Figure 1: Summary Statistics

3. Exploratory Data Analysis Results

In Figure 2, the average number of comments does not vary significantly, however the votes line shows a pronounced upward trend starting from around the 10th hour, peaking during the later hours of the day. This suggests that posts tend to receive more votes as the day progresses, with the highest average number of votes occurring in the evening and night hours. The top 10 subreddits by number of posts can be observed in Figure 3. The ‘funny’ subreddit seems to have the most posts, followed by others like ‘pics’ and ‘gifs’. This can inform which subreddits are most popular.

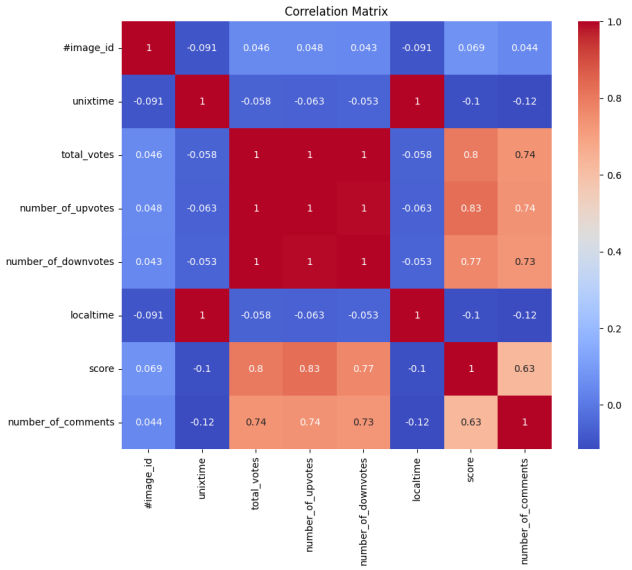

Figure 2: The correlation heat map shows strong correlation between total votes, upvotes and downvotes. Positive correlation observed between number of upvotes and score.

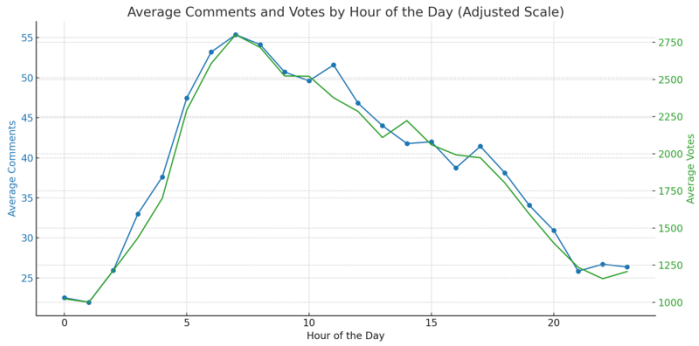

Figure 3: Times of day when submissions are most commented on or most rated.

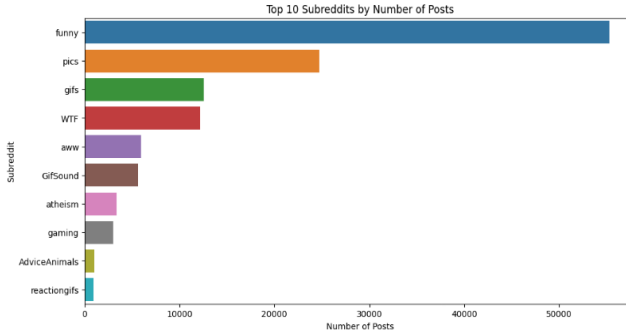

Figure 4: Activity differing across subreddits

4. Data Preprocessing

4.1 Data Loading and Feature Engineering

We begin by loading the Reddit submissions dataset, which includes information about posts, such as submission time, subreddit, and user engagement metrics. The 'unixtime' was converted to a human-readable datetime format from which additional temporal features were derived: hour, day of week, month, and year. These features can help capture trends and seasonality in submission scores.

4.2 Text Analysis- Sentiment Analysis

Sentiment analysis of post titles can provide valuable insights into how the emotional tone of a title influences user engagement. We utilize the TextBlob library to compute sentiment scores for post titles. The sentiment scores range from -1 (negative sentiment) to 1 (positive sentiment), with 0 indicating neutrality.

4.3 Handling Missing Data

To ensure the quality of our dataset, we handle missing data by imputing missing values with the mean of the respective numeric columns. This step is crucial for maintaining the integrity of our analysis. Username column has 20260 missing values, rest just have 1 missing value.

4.4 Feature Scaling

We apply feature scaling to numerical features using the Robust Scaler. This scaler is robust to outliers and ensures that features have similar scales, preventing one feature from dominating the others during modeling. Categorical features were one-hot encoded to transform them into a format suitable for machine learning models. Truncated Singular Value Decomposition (TruncatedSVD) was used as a dimensionality reduction technique to address potential problems of high dimensionality, particularly after encoding categorical variables.

4.5 Model Training and Hyperparameter Tuning

An 80/20 train-test data split was used to train the models. To maximize their performance, the RandomForest and XGBoost models' hyperparameters were tuned. By combining domain expertise with a grid search approach, the

tuning was done while taking overfitting concerns and model complexity into account.

5. Model Development and Evaluation

The models were assessed using the R-squared (R^2) and Mean Squared Error (MSE) values. While R2 shows how much of the variance in the dependent variable is predictable from the independent variables, the MSE gives an indication of the average squared difference between the observed actual outturns and the predictions made by the model.

5.1 Linear Regression with TruncatedSVD

We start our modeling process with linear regression augmented with dimensionality reduction using Truncated Singular Value Decomposition (TruncatedSVD). This approach allows us to capture latent patterns in the data while mitigating the curse of dimensionality.

$$y\hat{} = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta nxn$$

Here, $y\hat{}$ is the predicted score, $\beta 0$ is the intercept, βi is the coefficient for feature i, and xi is the value of feature i.

Mean Squared Error (MSE): 75,539.20
R-squared (R^2): 0.661

Our linear regression model provides valuable insights, but there is room for improvement.

5.2 Random Forest Regression

Next, we explore the application of a more complex model, Random Forest Regression. This ensemble method can capture nonlinear relationships and interactions among features.

$$y\hat{} = 1/N \sum y\hat{}T$$

Each tree T in the forest makes a prediction $y\hat{}T$, and the final prediction is the average across all trees, N is the number of trees in the forest.

Mean Squared Error (MSE): 18,756.79
R-squared (R^2): 0.916

The Random Forest model significantly outperforms the linear regression model, suggesting that complex interactions among features play a crucial role in predicting submission scores.

5.3 XGBoost Regressor

XGBoost is a powerful gradient boosting algorithm known for its exceptional performance in regression tasks. We incorporate XGBoost into our pipeline and observe the following:

$$y\hat{}i(t) = y\hat{}i(t-1) + \eta \cdot ft(xi)$$

Here, $ft(xi)$ is the prediction of the t-th tree, and η is the learning rate.

Mean Squared Error (MSE): 17,855.38
R-squared (R^2): 0.920

The XGBoost Regressor outperforms both linear regression and Random Forest models, indicating that gradient boosting techniques effectively capture the underlying patterns in our data.

5.4 Support Vector Regression (SVR)

Lastly, we experiment with Support Vector Regression (SVR), a kernel-based method. Our SVR model uses a linear kernel, but other kernels like polynomial or radial basis function (RBF) can be explored. Our findings include:

$$y\hat{} = w \cdot x + b$$

Where w is the weight vector, x is the feature vector, and b is the bias.

Mean Squared Error (MSE): 93,244.48
R-squared (R^2): 0.582

While SVR provides reasonable results, it lags behind the other models, suggesting that linear relationships may not fully capture the complexity of the data.

6. Discussion

6.1 Feature Importance

Understanding which features contribute most to predicting submission scores is essential for content creators. Random Forest and XGBoost models allow us to assess feature importance.

Some key insights include:
Time of Submission: The hour of submission has a significant impact, with certain hours leading to higher scores.
Subreddit: The subreddit to which a post belongs strongly influences its score.
Sentiment Analysis: Positive sentiment in post titles correlates with higher scores.

Model	MSE	R^2
Linear Regression	75,539.20	0.661
Random Forest Regression	18,756.79	0.916
XGBoost Regressor	17,855.38	0.920
Support Vector Regression (SVR)	93,244.48	0.582

Table 2: Comparison of models employed on MSE AND R^2

6.2 Model Selection
Linear Regression served as our baseline model. Despite achieving an MSE of 75,539.20, the model's R^2 was 0.661, meaning that it could account for roughly 66.1% of the variation in post scores.

The Linear Regression model did fairly well considering the complexity of the dataset, but it was evident that a model that could capture non-linear patterns might be able to produce better results.

Random Forest Regression significantly improved upon the baseline, with an MSE of 18,756.79 and an R^2 of 0.916. The significant improvement in performance indicated that the ensemble method of Random Forest—which constructs several decision trees and combines them to produce a more reliable and accurate prediction—was more appropriate for the complex Reddit data.

XGBoost Regressor further refined the predictions, demonstrating the efficacy of gradient boosting frameworks in handling tabular data. It yielded the highest MSE of 17,855.38 and R2 of 0.920 out of all the models that were tested. This demonstrates how well the model can manage overfitting while learning intricate relationships within the data.

Support Vector Regression (SVR), despite its sophistication in feature space transformation, yielded an MSE of 93,244.48 and an R^2 of 0.582. These numbers suggested that compared to tree-based models, SVR was less successful at capturing the predictive signals in the Reddit dataset, at least when using the linear kernel.

Our results indicate that the XGBoost Regressor outperforms other models in predicting submission scores. Its high R-squared value (0.920) suggests that it captures complex relationships effectively. However, the choice of the best model may vary depending on specific use cases and computational resources.

7. Literature Review
Numerous facets of content performance on social media platforms have been examined in earlier research. According to Tandoc and Vos (2016), engagement—which is gauged by likes, shares, and comments—is crucial when it comes to online content success metrics. The comparable metrics on Reddit are upvotes, downvotes, and comments. These factors add up to a post's score, which is the main focus of our predictive analysis.

Complementing the investigation into Reddit post-performance, Zamoshchin and Segall (2012) employed machine learning to predict post scores from initial comments and votes, substantiating the significance of early user interaction. This work serves as a foundation for our current study, highlighting the predictive potential of post timing and community context — aspects we aim to delve into with greater granularity.

Researchers have used machine learning techniques in the field of predictive modeling to predict social media content's popularity. McAuley, J., et al. (2013) investigated the relationship between features taken from Reddit post titles and the likelihood of a given score and volume of comments.

8. Future Scope

This research offers opportunities for further investigation:

Feature Engineering: Exploring additional features, such as post content analysis, user engagement patterns, or external factors like trending topics.

Hyperparameter Tuning: Fine-tuning model hyperparameters to optimize performance further.

Ensemble Methods: Combining predictions from multiple models to enhance predictive accuracy.

User-Specific Models: Developing models tailored to individual user behavior.

Advanced Text Analysis: Further research on advanced text analysis could use deep learning sentiment analysis or topic modeling, two more advanced natural language processing methods, to get deeper insights from post titles and content.

Cross platform Studies: Analyzing the dynamics of content popularity on various social media sites (such as Facebook, Twitter, and Reddit) may help identify engagement tactics unique to each platform.

Ethical and Societal Implications: It would be beneficial to investigate the moral and societal ramifications of predictive analytics in social media, especially as they relate to privacy and the possibility of manipulation.

9. Conclusion

In this research, we applied various machine learning models to predict Reddit submission scores based on post metadata and sentiment analysis of post titles. Our findings highlight the importance of features like submission time, subreddit, and sentiment in determining submission scores. The models created and assessed in this study, which ranged from simpler methods like Linear Regression to more complex ones like Random Forest and XGBoost, offered insightful information about the complex processes underlying Reddit user engagement.

The XGBoost Regressor emerged as the top-performing model, achieving an R-squared value of 0.920. This suggests that complex relationships and interactions among features significantly impact submission scores on Reddit.

Content creators, platform administrators, and data scientists can leverage the insights from this research to enhance content strategy, detect anomalies, and gain a deeper understanding of user engagement dynamics on Reddit.

As the Reddit platform continues to evolve, ongoing research in predictive modeling and data analysis will be essential to staying at the forefront of content performance optimization.

10. References

Tandoc, E. C., Jr., & Vos, T. P. (2016). The journalist is marketing the news: Social media in the gatekeeping process. Journalism Practice, 10(8), 950-966. https://doi.org/10.1080/17512786.2015.1087811

Zamoshchin, A., & Segall, R. (2012). Predicting Reddit post popularity via initial comments and votes. Stanford University, CS229 Machine Learning Projects. Retrieved from https://cs229.stanford.edu/proj2012/ZamoshchinSegall-PredictingRedditPostPopularity.pdf

Lakkaraju, H., McAuley, J., & Leskovec, J. (2013). What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (pp. 311-320). Retrieved from https://cseweb.ucsd.edu/~jmcauley/pdfs/icwsm13.pdf

Loria, S. (n.d.). TextBlob: Simplified Text Processing. Retrieved from https://textblob.readthedocs.io/en/dev/

Chen, T., & Guestrin, C. (n.d.). XGBoost Documentation. Retrieved from

https://xgboost.readthedocs.io/en/latest/index.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Dubourg, V. (n.d.). Scikit-learn: Machine Learning in Python. Retrieved from https://scikit-learn.org/stable/

Reddit. (n.d.). Reddit: The Front Page of the Internet. Retrieved from https://www.reddit.com/