MiniProject 1: Predicting the Popularity of Reddit Comments using Linear Regression

Jean-Sébastien Grondin
McGill Id:260345519
jean-sebastien.grondin@mail.mcgill.ca

Zhourong Lee
McGill Id:260674414
zhourong.li@mail.mcgill.ca

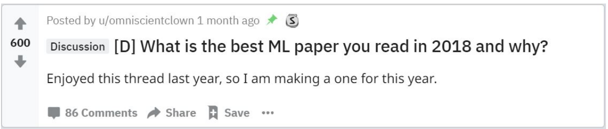
Zhenze Han
McGill Id:260675404
zhenze.han@mail.mcgill.ca

Abstract

This paper presents the methodology and results obtained when predicting the popularity of Reddit comments using linear regression. This study used a dataset of 12000 instances which was split into training, validation and test sets. Several additional features were extracted and analyzed for improving the model performance. **We found that...** The model achieved a mean squared error (MSE) of **X.XXX** on the test set.

1 Introduction

Reddit is a popular social media site which is currently the 5th most popular website in the United States according to Alexa. It is essentially a massive collection of forums and threads where users can upvote or downvote comments that they like or dislike (see an example below).



Posted by u/omniscientclown 1 month ago

600

Discussion [D] What is the best ML paper you read in 2018 and why?

Enjoyed this thread last year, so I am making a one for this year.

86 Comments Share Save ···

The objective of this study was to implement and evaluate a linear regression model for predicting the popularity score of Reddit comments. A total of 12000 instances of text comments were available, in addition to other features such as the controversiality score and the number of replies each comment received. Text comments were processed in order to extract a word count occurence feature for the most frequently occuring words. Several other features were generated, and the two features which exhibited the highest correlation with the popularity score were retained. Both the closed form and the gradient descent solutions were implemented and compared. The mean squared error was the metric used for all comparisons and analysis. The closed-form approach was used for subsequent experiments as it was found to be faster and more stable and also because it yields the exact solution. Models which included word occurence features were found to overfit compared to models that did not include word occurence features, so these features were droped from the final model. The two new features that were extracted were shown to improve the model performance without causing overfitting. The final model achieved a mean squared error of **X.XXX** on the test set.

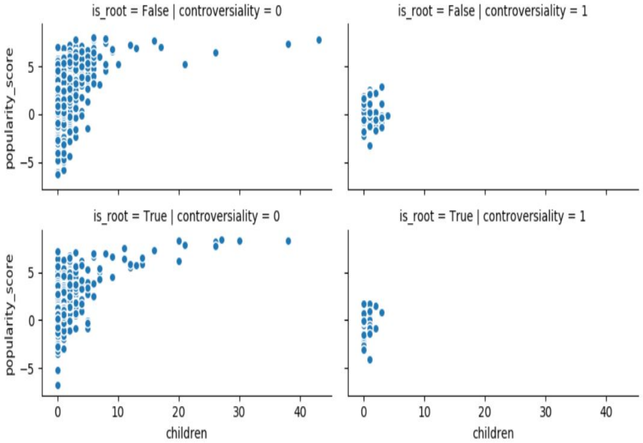
2 Dataset

The original dataset was obtained from the subreddit community **r/Askreddit**. The dataset contained 12000 instances with the following information:

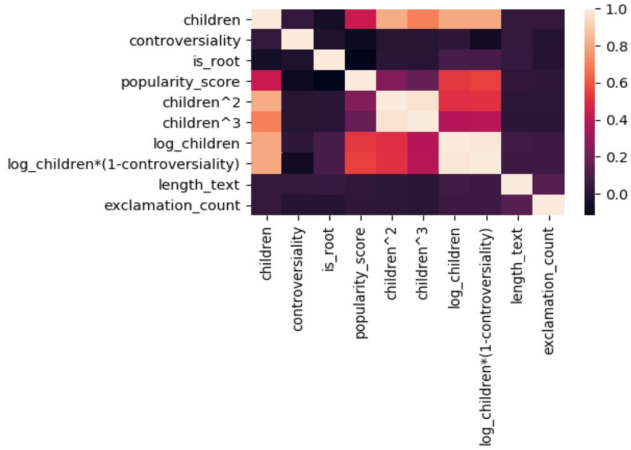
- **popularity_score:** This metric indicates how popular each comment is and is computed by Reddit. This is the target variable in this study.
- **children:** This feature indicates how many replies each comment received.
- **is_root:** This feature is True when a comment is the "root" of a conversation and is False when it is a reply to another comment.
- **controversiality:** This feature is a binary variable that indicates if a comment is controversial (when equals 1) or not controversial (when equals 0). A controversial comment is one that receives close to the same amount of upvotes and downvotes.
- **text:** The raw text of each comment.

The is_root boolean values were encoded using 0 and 1. The dataset was then split into training (80%), validation (10%) and test (10%) sets and the training set was subsequently used for text processing, exploratory data analysis and feature extraction. Simple pre-processing was applied to text comments by first a) lower-casing all text and then b) spliting text based on whitespace tokens to generate a list of all words. A list of the 160 most frequently occurring words in the training data was then computed. For every text comments, a 160-dimensional feature vector, x_{counts}, was extracted, with $x_{counts}[i]$ being equal to the number of times the ith most frequent word occured in the comment. Using the same list, word occurence feature vectors were extracted for the test and validation sets.

Some exploratory data analysis was also achieved. The following figure gives a visual representation of the original non-text features. Several observations were made. The popularity score was found to increase with an increasing number of children. Also, controversial comments generally did not lead to high popularity nor high number of children. While it is not readily noticeable in the plot, the mean popularity score is slightly higher (0.981) for non-root comments compared to root comments (0.703), and the [min, max] range is also shorter [-6.267, 7.981] for non-root comments than for root comments [-6.779, 8.373].



Since children seemed to be somewhat correlated with the popularity, the following additional features were generated for further analysis: $children^2$, $children^3$, $\ln children$, $(\ln children) * (1 - controversiality)$. Also, two additional features were also extracted from the text comments: a) the length of each comment and b) the number of exclamation marks occuring in each comment. A correlation matrix heatmap was then computed to assess the correlation of features with the target variable and is displayed below:



The two additional text related features mentioned above were not investigated further as they did not seem to be sufficiently correlated to the target variable. As for the others, a discussion of how these contribute to the model performance is included in the following section.

Before moving on to the results section, readers should be wary of the ethical implications of using machine learning with social media data for analysing or filtrering information. For example, if machine learning models are used to provide intelligence, they can introduce several types of bias related to the way data is collected, models are built, etc. Hence the authors wish that readers are cautious about a potential misuse of the models or a flawed interpretation of the results.

3 Results

4 Discussion and Conclusion

In this study, we have outlined a machine learning workflow that uses linear regression to predict the Reddit comments popularity score. Main takeaways are **that ......**

The study discussed in this report could definitely be improved. Possible directions for future improvements include:

- **A)** Try out different machine learning models and ensembling methods for better performance
- **B)** Obtain additional features directly from Reddit (e.g. parent score, whether comment is gilded, a time lapse feature between comments or since root comment was created, etc.)
- **C)** Remove punctuations from text words.
- **D)** Assess benefit of removing stop words (e.g. "the", "a", "an", "in") that may no be effective for distinguishing popular comments than non popular ones.
- **E)** Identify most frequent 2-gram and 3-gram sequences of words and assess whether the model performance can be improved using these.

5 Statement of Contributions

- **J-S Grondin** Primary role was to implement the linear regression algorithms and building the machine learning pipeline for this project. Also was responsible for undertaking the exploratory data analysis which led to the selection of additional features. Finally, contributed to the writing of the report.
- **Z. Lee** ...
- **Z. Han** ...