Predicting the Popularity of Reddit Posts

Ahmed Shuaibi
ashuaibi@stanford.edu

Problem

- Given a Reddit post, I would like to predict its popularity as measured by the number of upvotes it receives.
- The amount of upvotes a post receives impacts its visibility on the site. It is therefore important to understand the relative significance a post's features have on its popularity.
- I utilize NLP techniques such as GloVe word embeddings and sentiment classification in addition to post metadata using Linear, KNN, and Random Forest Regression tackle this problem.

Dataset & Features

- Raw data consists of json entries of all Reddit submissions over the first 6 months of 2018 with 96 fields that encompass the post's information and metadata.
- After processing and filtering out posts without text content, we obtain all submissions falling under the subreddit AskReddit.
- Features are generated using the post's text, such as sentiment classification and embedding representation. These features are used in combination with provided post features to predict the upvote score of a post.

Raw Features:

- num_comments - Number of post comments: Natural number
- gilded - Number of times gilded: Natural number
- over_18 - Is for mature audiences: 2D one hot encoded vector.

Derived Features:

- hour - Hour of day of post creation: 24D one-hot-encoded-vector.
- day - Day of week of post creation: 7D one-hot-encoded-vector.
- title_length- Length of title: Natural number
- sentiment – Positive/Negative: 2D one-hot-encoded-vector.
- embeddings – MOWE of post title: 300D vector.

References

- Adam Reevesman. Reddit Comment Karma, Dec 2018
- Andrei Terntiev and Alanna Tempest. Predicting Reddit Post Popularity Via Initial Commentary, 2014
- Daniel Poon, Yu Wu, and David Zhang. Reddit Recommendation System, 2011
- Jason Baumgartner. PushShift.io. Directory Contents
- Jordan Segall and Alex Zamoshchin. PREDICTING REDDIT POST POPULARITY, 2011
- Lyndon White, Roberto Togneri, Wei Liu, Mohammed Bennamoun. How Well Sentence Embeddings Capture Meaning, 2015
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532-1543, 2014
- Prasanna Chandamouli, Radhakrishnan Moni, and Varun Elango. Predicting reddit post popularity, 2017

Models + Metrics

1) Linear Regression

Ordinary least squared loss cost and default learning rate of 0.0001 were used

$$Cost = \sum_{i}^{n} ||y_i - \hat{y}_i||_2^2$$

This model is inflexible and subject to high bias.

2) K-Nearest-Neighbors Regression

KNN Regression considers examples that are closest together in the feature space. After finding the 5 nearest neighbors to a particular example, the average of these neighbors' scores is used as the predicted value. For this model, the standard Euclidian distance function is used:

$$distance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

This model is very flexible and subject to high variance.

3) Random Forest Regression

Utilizes a combination of multiple decision trees and the average score of these trees to arrive at a score prediction. Uses bootstrap aggregation to continuously randomly sample examples from the training data to fit decision trees. After doing this random sample with replacement and thereby obtaining B decision trees, the prediction for a test example is given by:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Bootstrap aggregation reduces variance and allows for good bias-variance tradeoff balance.

Metrics

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2}$$

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Experimental Results

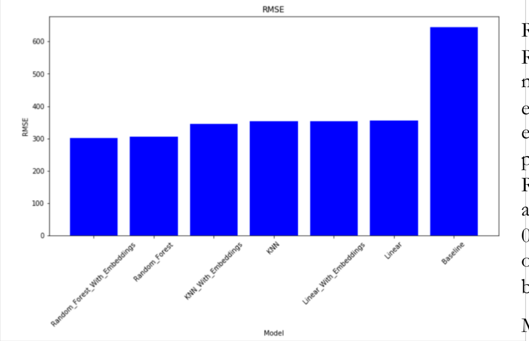

Figure 1: RMSE vs. Model

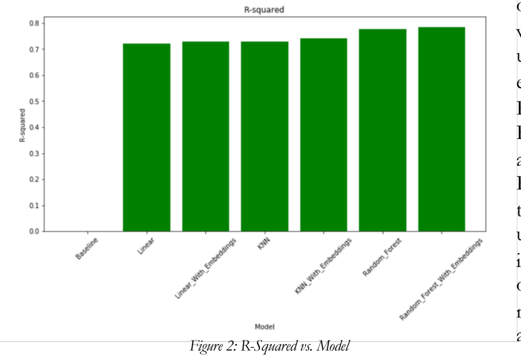

Figure 2: R-Squared vs. Model

Random Forest Regression utilizing the mean of word embeddings as features exhibited the best performance, with an RMSE value of 301.32 and and R² value of 0.7855. This outperformed the baseline by 53.49%.

Models utilizing word embeddings as marginally outperform those without. Specifically, utilizing word embeddings with Random Forest Regression brought about a 1.1% decrease in RMSE. Furthermore, a thorough ablative analysis uncovers that the most important features in order of importance are num_comments, gilded, and embeddings.

Future Work

- Embedding representations for a subset of the comments would also be a beneficial feature in popularity prediction.
- Reddit specific word embeddings can be generated with GloVe given a corpus of data that may be more suitable for this task.
- Finally, this popularity prediction can be applied to other subreddits and can utilize features that do not solely depend on text, such as attached images or links.