Classifying Lego Subreddits
Executive Summary

Blake Wallace

April 8, 2019

1. Data science objective:

   Construct and compare two classification models that attempt to predict whether a single random reddit post is from a subreddit associated with lego.

2. Data:

   13,212 reddit posts are used to train our models. This included 6,566 from three different lego subreddits, along with 6646 from three different non-lego subreddits.

3. Results:

   Even though the first constructed models using a Random Forest classification varied significantly when tested on new data, the final results using unseen data on a Naive Bayes model tested quite closely to the initial training data.

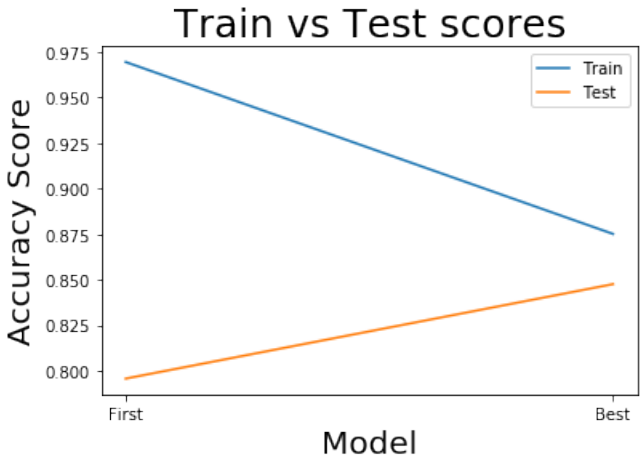

Figure: Comparing the variance between the first model and the best model.

4. Recommendations:

   Our recommendation is that the Multinomial Naive Bayes model should be used to make predictions when attempting to classify lego subreddits. We believe that the results are worth exploring more to see if greater accuracy can be obtained, and also to look into the generalizability of the model to the classification of other subreddits.