What features did you consider?

To reduce the scope, I only considered predicting the virality of English posts. I preprocessed the ***article text*** using ***TF-IDF*** because I wanted to ***extract keywords*** from the articles. The intuition behind TF-IDF is that it looks for words that appear often in a particular document, but not in very many documents. This helps to find the most descriptive words in a text corpus. I also standardized the URLs and one hot encoded them based on the ***50 most frequent URLs***. I also one hot encoded the ***50 most frequent authors***. The reason I used Top 50 for both URLs and authors was to minimize the number of columns I added to the feature space.

What model did you use and why?

I went with traditional machine learning techniques instead of neural networks to keep the models simple and explainable. Since I converted the problem into a binary classification, I trained four of the most popular models for classification which were Logistic Regression, Naive Bayes with Bernoulli distribution, Decision Tree Classifier, and Random Forest Classifier. I decided to go with the Logistic Regression because it was able to meet my optimizing and satisficing metrics mentioned below.

What was your evaluation metric?

For my evaluation, I used Recall and Accuracy. I ***optimized for Recall*** (optimizing metric) and set an ***acceptable limit of 85% for accuracy*** (satisficing metric). I chose recall because I wanted to ensure most viral articles were being recommended and I didn't really care if some non-viral articles were recommended because that only cost me real estate on the users' feed.

What features would you like to add to the model in the future if you had more time?

I would like to incorporate Portegues articles into my system. I would do this by translating the Portegues into English and retraining the model. I could also build two models, one for Portegues, and another for English. I would also use the timestamp as a feature because there are prime times when people are browsing the web so articles posted at certain times may have a better chance of becoming viral. A feature that I would like to add that is not part of the dataset would be the influence of the author. This could be the number of followers they have on social networks or how long they have been reporting.

What other things would want to try before deploying this model?

I would analyze the coefficients of the Logistic Regression model. This will help give me a better understanding of its decision-making process so I can answer questions from users or stakeholders when they are wondering why it recommended certain articles. I would also train it on the full dataset instead of only on my training set. Afterward, the only thing I would do is serialize the model and deploy it.