**Overview of impact of SVD Dimension Reduction**

The output of the CountVectorizer operation is a large matrix; for our relatively small dataset, with default settings, there were over 5000 columns. Given this, the curse of dimensionality is bound to come into effect. It’s naturally an attractive idea to consider reducing the dimension.

A brief review of the literature[[1]](#footnote-1) suggests SVD is widely used in NLP and is a powerful tool to reduce the dimensionality of the data. Please see below for a plot comparing variance explained by number of components in the SVD

Chart

Description automatically generated

At about 1000 components, it appears that ~80% of variance is explained. A legitimate question arises: What of this variance is “signal” and what is “noise”? To test this, we compare performance of models with 10, 100, 500, 1000, and 2000 components kept in PCA. Please see below for a chart comparing the performance in each model. Graphical user interface, application

Description automatically generated

As we can see, GaussianNB and RandomForest actually decrease. SVC and LogisticRegression both perform nearly as well with 500 components as they do with 2000, both of which are quite close to the performance with the full data. It may be noteworthy that, although we are not considering model fitting time as part of our evaluation criteria, most models, especially the SVC, fit noticeably faster on dimension-reduced data, with the fitting time decreasing as dimension decrease.

1. https://learning.oreilly.com/library/view/practical-text-mining/9780123869791/xhtml/CHP039.html [↑](#footnote-ref-1)