Some Parts

Data

Our main data source is the COVID-19 Open Research Dataset (CORD-19) assembled by the White House and leading research groups. It is a collection of scientific articles about COVID-19, SARS-CoV-2, and related coronaviruses. Our 12372 sample subset of this dataset is chosen by first selecting observations available on the NCBI website and second ensuring that the full body of the text is available in the dataset. Unfortunately citation data was not supplied in the CORD-19, to gather these values we scraped available citation data from the NCBI website. The information available in our compiled dataset included, journal of publication, full article text, authors, number of articles cited in the text, and number of citations received by the article. Since we are interested in understanding paper importance in terms of contribution to the field, these citations are key to our inquiry.

To transform the articles' text into usable features, each abstract and text body are run separately through gensim's python implementation of Doc2Vec. This resulted in a 100 dimensional vectorization of each text. We chose to analyze the abstracts independently because they are designed as a succinct summarization of the paper's content and could help focus the vectorization. Each of these vectorizations were then embedded into 5 dimensional space using a variational autoencoder, Isomap, and Spectral embedding. Using these embedded documents, we create three new features for the abstracts and the full text: mean Euclidean distance of 5 nearest neighbors, mean Euclidean distance of 5 nearest neighbors from the preceding six months, and a count of texts within an open ball of radius $0.1 *$ standard deviation of all distances. These were designed to measure where the work sits within the topic's literature, where it sits in relation to prior work, and density of work in the specific topic respectively.

The final features used in the model are as shown below:

Features
Journal name
Publication year
Average abstract distance
Six month average abstract distance
Count of close abstracts
Average body distance
Six month average body distance
Count of close abstracts

Due to heavy right tails and non-linearities, some of the variables are transformed. A log transform is applied to the average abstract distance, six month average abstract distance, and average body distance. A square root then log transform is applied to the second and third dimension of the embedded body texts.

Because our interest is in predicting paper influence for recent papers, we hold out all papers published in 2020. In addition to being the papers of interest, we felt the recent nature of their publication may have an influence on their total citation numbers. Our predicted ranking of these papers is reproduced in the appendices.

Results

The most successful model is an xgboost model with max depth of 5, learning rate of 0.5, and objective functions squared error and hinge loss for regression and binary classification respectively. Also shown below are the results from a random forests model. In addition to these models we tested linear models, a neural network, and Gaussian process regression and classification. However, the tree based methods performed best on the training data.

The results show that while the choice of embedding had an effect on accuracy, with the variational autoencoder embedding performing best, these improvements were only marginal. Similarly, the effect of the text based features, while helpful, were less important than date of publication and journal of publication. Inspection of feature importance from the xgboost model showed that far and away the most predictive feature for citations is the journal in which the work is published.

_	AUC	Accuracy	F1
Spectral Embedding			
xgboost	0.71	0.71	0.77
random_forests	0.70	0.70	0.72
Isomap Embedding			
xgboost	0.71	0.72	0.78
random_forests	0.69	0.70	0.71
Variational Autoencoder			
xgboost	0.72	0.73	0.77
random_forests	0.70	0.70	0.72
No Embedding			
xgboost	0.72	0.73	0.78
random_forests	0.71	0.71	0.73

_	Variance Explained	MSE	R2
Spectral Embedding			
xgboost	0.40	18.36	0.63
random_forests	0.35	19.94	0.60
Isomap Embedding			
xgboost	0.39	18.73	0.62
random_forests	0.33	20.46	0.59
Variational Autoencoder			
xgboost	0.38	19.10	0.61
random_forests	0.36	19.70	0.60
No Embedding			
xgboost	0.40	18.52	0.63
random_forests	0.35	19.98	0.60