Personal Protective Equipment Model Report

August 26th 2020, Jeanny Zhang

Introduction

As COVID-19 has quickly evolved into a global epidemic, a lot of countries and areas are experiencing an abrupt surge in coronavirus cases, which leads to a severe shortage in personal protective equipment (PPE). PPE is equipment worn to minimize exposure to hazards that cause serious illness and injuries. It includes items such as gloves, safety glasses and shoes, earplugs or muffs, hard hats, respirators, or coveralls, vests and full body suits. During the current pandemic, it is crucial for health care workers to have sufficient PPE to protect themselves from infections. Although many organizations have been endeavoring to prompt PPE donation or fundraising, few have been able to model PPE shortage using statistical models.

In order to distribute supplies more adequately and effectively, it is important for us to model PPE consumption and shortage in specific areas. This report is dedicated to discovering a statistical model that best predicts personal protective equipment consumption. The resulting model will be used to make predictions of the relative adversty of PPE shortage in specific areas. It will be integrated into PureFDA's data platform as a public resource. All the data used in this study are gathered from the website of The COVID Tracking Project and the website of Get Us PPE.

Method

In order to predict and model PPE shortage of a certain area, particular COVID-19 data of that area are needed. In this study, nine variables are used to predict the amount of PPE needed, and they are the current number of positive cases, the current number of hospitalized COVID-19 patients, the accumulated number of hospitalized COVID-19 patients, the current number of COVID-19 patients in ICU, the accumulated number of COVID-19 patients in ICU, the current number of COVID-19 patients on ventilator, the accumulated number of COVID-19 patients on ventilator, the number of recovered cases, and the number of deaths.

The PPE data were collected from the Get Us PPE website and the COVID-19 data were collected from the COVID Tracking Project website. These data were used as the training data for the regression model with the COVID-19 data being the independent variables and the PPE data being the dependent variable.

Because the website of Get Us PPE only provides the PPE data on May 12th 2020 and it is considered as insufficient for the training dataset to contain only one day of data, KNN imputation was used to replace missing values with substituted values for May 10th, 11th, 13th, and 14th 2020. These consecutive dates were selected because they have the most similar data to the one on May 12th 2020, decreasing the error rate in imputation. To introduce uncertainty and generate more accurate data, all the imputed data points were multiplied by a Gaussian distribution (1, 0.1), so that the imputed values vary between 90% and 110% from the original values. This process is known as a Gaussian modification.

A dataset that contains only five days of complete data was used as the training dataset for this study. A small-sized training dataset is acceptable in this case because the model will be used to forecast the relative urgency of PPE demand among areas instead of predicting the exact numbers of PPE needed.

Figure 1 is a summary chart illustrating all types of regression models tested in this study.

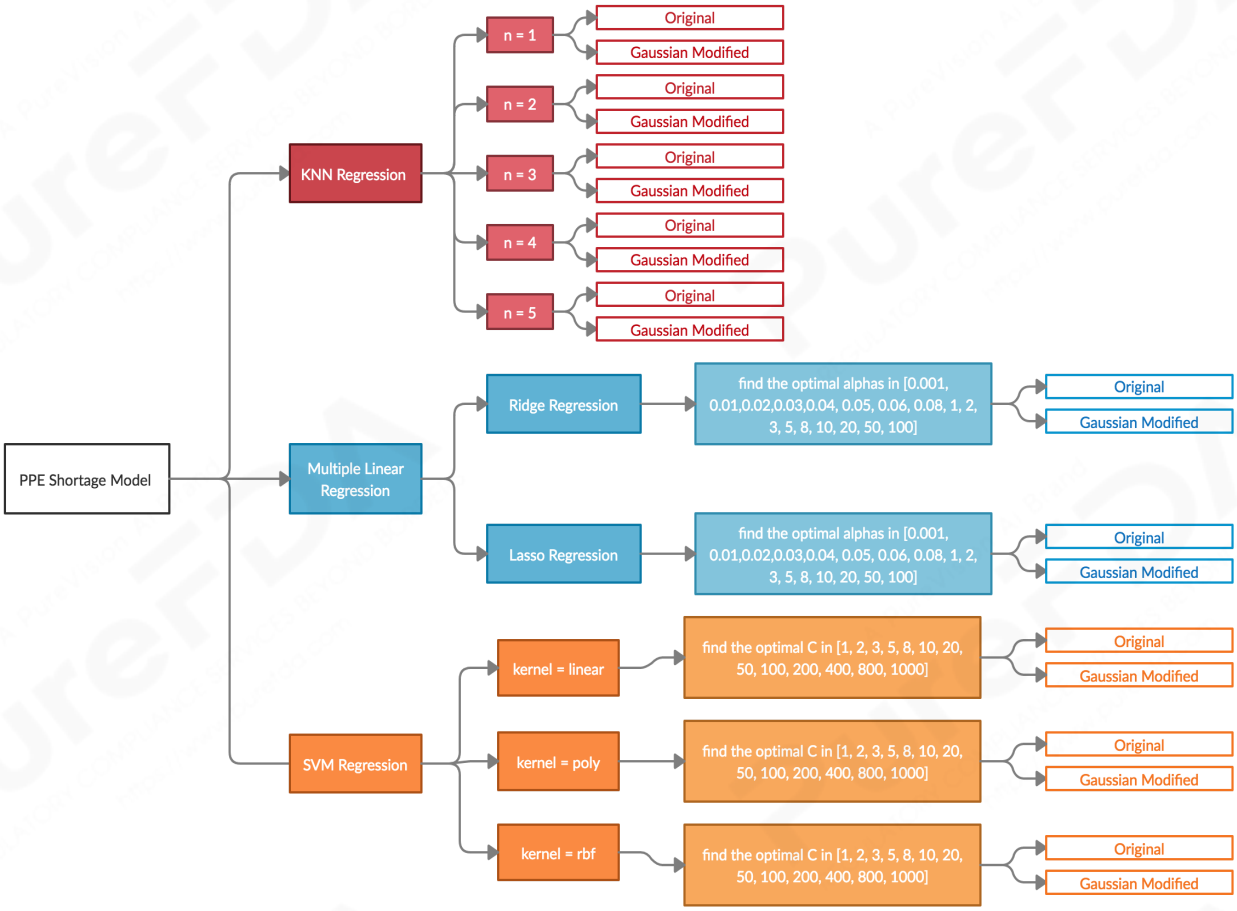

Figure 1. All Tested Regression Models

A total of twenty regression models were explored and compared in this study. The major three types of regression models are the k-nearest neighbors regression (KNN), the multiple linear regression (MLR), and the support vector machine regression (SVM). In order to find the most suitable model for PPE shortage, an average mean standard error (MSE) is calculated for each regression model. MSE is calculated by running the model a hundred times and taking the average of a hundred mean standard errors. Because regularization can have a huge impact on the model's performance, an optimal regularization parameter was calculated and the MSE at the optimal regularization parameter would be the average MSE for the model.

The KNN regression model was tested by varying the n-value, meaning the first n number of data points would be selected for calculation from the collection of data entries sorted by distances in ascending order. The n-values of 1, 2, 3, 4, and 5 were tested respectively. The multiple linear regression model was investigated by applying different types of regularization, Ridge regression penalty and Lasso regression penalty. The optimal parameter of each regularization was found and the average MSE was calculated. Various kernel functions, including linear, polynomial, and radial basis function (rbf), were examined for the SVM

regression model. The optimal regularization parameter of each kernel was also found and the average MSE was calculated.

Results

Table 1 contains the average MSE for all the tested models.

Table 1. Average MSE of Tested Models

Model	Parameter	Average MSE (original)	Optimal Regularization Parameter	Average MSE (Gaussian modified)	Optimal Regularization Parameter
KNN Regression	K = 1	0.0169	/	0.0211	/
	K = 2	0.0143	/	0.0136	/
	K = 3	0.0195	/	0.0260	/
	K = 4	0.0272	/	0.0260	/
	K = 5	0.0221	/	0.0262	/
Multiple Linear Regression	Ridge Regularization	0.252	alphas = 0.08	0.255	alphas = 0.08
	Lasso Regularization	0.121	alphas = 0.001	0.103	alphas = 0.001
SVM Regression	kernel = linear	0.458	C = 1000	0.571	C = 1000
	kernel = poly	0.0702	C = 10	0.169	C = 20
	kernel = rbf	0.0252	C = 100	0.0412	C = 100

Figure 2 is a graphic representation of the average MSE of all tested models.

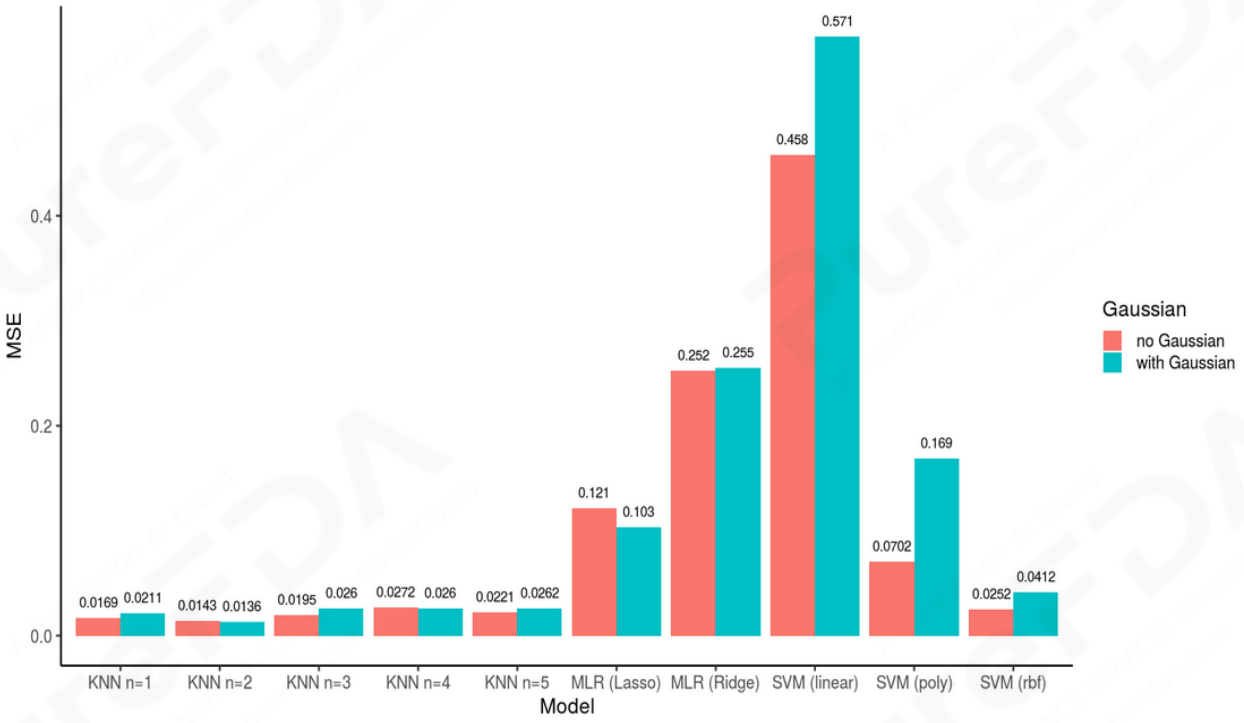

Figure 2. Average MSE of Tested Models

Figures below are graphical representations of the relationship between MSE and various regularization parameters for MLR models and SVM regression models. The optimal regularization parameter lies in where the MSE is the smallest.

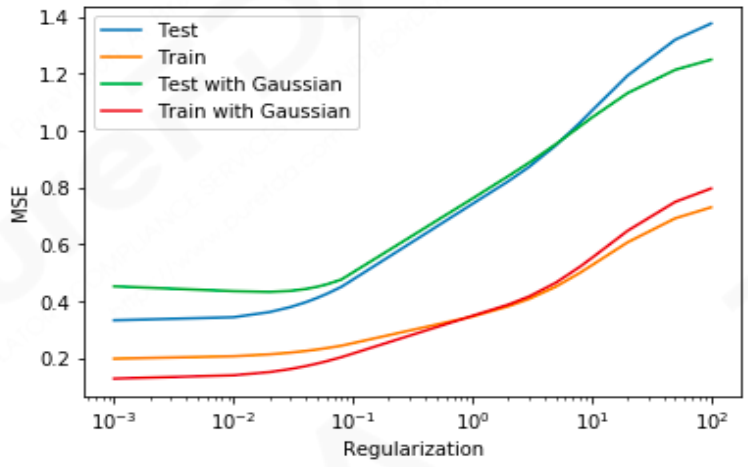

Figure 3. MSE Vs Regularization Parameters (MLR with Ridge Regularization)

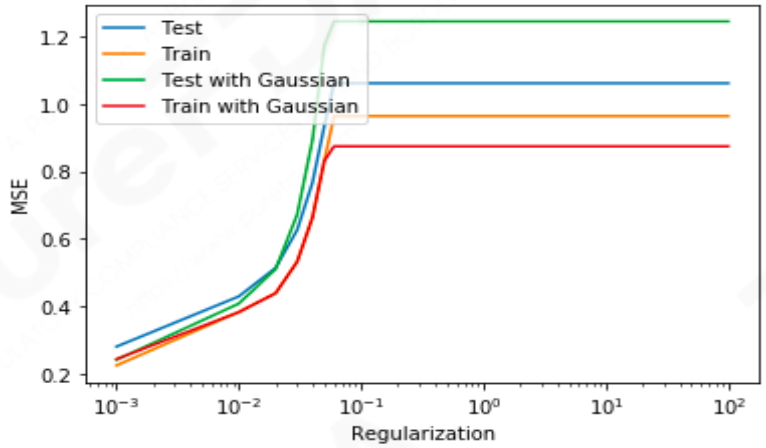

Figure 4. MSE Vs Regularization Parameters (MLR with Lasso Regularization)

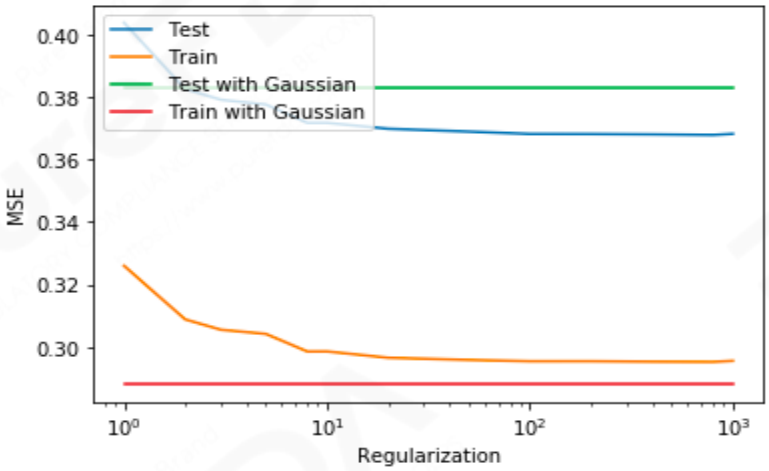

Figure 5. MSE Vs Regularization Parameters (SVM with Linear Kernel)

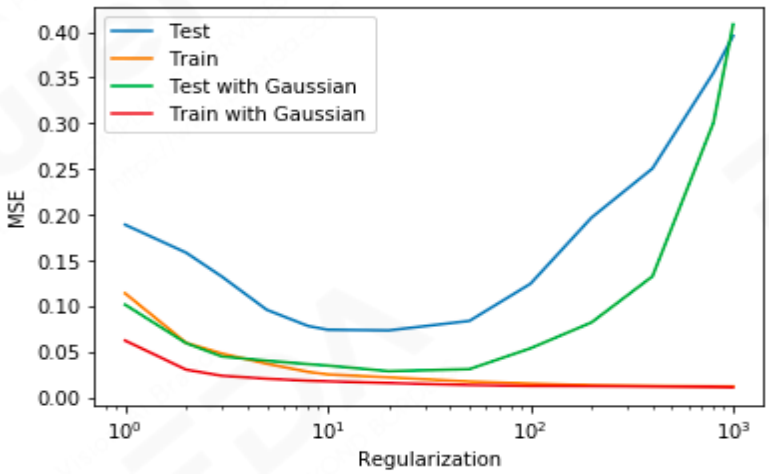

Figure 6. MSE Vs Regularization Parameters (SVM with Poly Kernel)

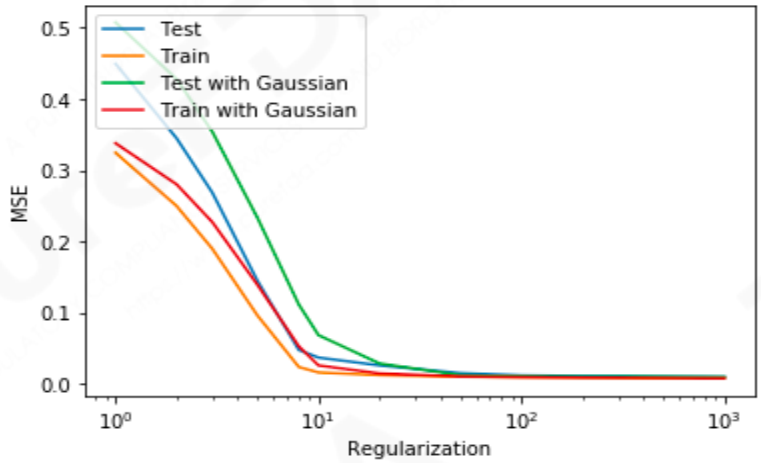

Figure 7. MSE Vs Regularization Parameters (SVM with RBF Kernel)

Discussion

Statistically, lower the value of MSE, the data fits the regression model better. Based on the results, the K=2 nearest neighbor regression model with a Gaussian modification has the lowest average MSE (0.0136). Although the KNN regression models performed exceptionally well in this study compared to others, it is not valid to deploy this type of model. Due to the fact that the training data set contains daily data for each country/area, the nearest neighbor of a country/area would always be the data from the same country/area on a different day. Therefore, the model is considered to be “cheating” in this case. It is in doubt if a KNN regression model will still have good performance after deployment.

The model with the second lowest average MSE is the SVM regression model with a rbf kernel, which has a MSE of 0.0252. The optimal regularization parameter for this model is C = 100. As shown, it is the most suitable model in predicting PPE consumption and shortage. It will be integrated into PureFDA’s data platform as a public resource.