A machine learning approach to predicting and classifying the spread of contagious viruses in the US using COVID-19 data

Rohit Rungta, Brian Law

Abstract

This work applies machine learning algorithms to analyze the spread of COVID-19 in the United States while predicting and classifying counties as high or low risk. An artificial spread factor is created to quantify the rate at which the virus spreads in a county. Lasso regression analysis is conducted to search for features that are the most important or promising in predicting the spread of COVID-19 in US counties. The lasso regression model has a test RMSE of 0.79 for our most accurate model and the most important features in predicting the spread of COVID-19 are the population density, hospital infrastructure, and the day at which the governor implemented stay-at-home policies. For classification, three methods were implemented which utilized different feature matrices. The best overall classification method involved using the top three features of the lasso regression model where the training accuracy was 85% and testing accuracy was 60%.

Introduction

Although initially localized in China, the COVID-19 virus has transformed into a global pandemic and forced countries to re-evaluate their health systems and infrastructure while crippling the global economy. While some countries such as China and South Korea have acted swiftly in their actions to contain the virus, some such as Italy and the United States have been slow to react. Their ineffective response to this global pandemic has costed hundreds of thousands of lives. One main issue that the United States, specifically, faces is their lack of testing kits and hospital infrastructure. Academic institutions such as Johns Hopkins University have been compiling and tracking the number of confirmed cases all over the world; however, the United States is facing a severe shortage of testing kits. Thus, hospitals are not able to test everyone that believes they may have come in contact with the virus. This makes analyzing COVID-19 cases in the United States very difficult, as confirmed cases are not a measure of how many people have the virus, but rather how many people have been tested. For this reason, this study does not look at any data involving confirmed cases.

This study analyzes the deaths in each county over time to understand how quickly the virus has spread. We are then able to determine how susceptible counties are to highly contagious viruses such as COVID-19. While this study does not fight against COVID-19 directly, it does provide key insight in fighting the rest of this pandemic as well as future and inevitable global pandemics. Identifying counties that are of high risk will allow policy makers to provide a rough framework of where to spend the most resources and manpower to curb the spread of the virus in addition to successful policies or actions.

Methods

Description of Data

All data for this study was provided by the Johns Hopkins COVID-19 dataset made publicly available on GitHub. Specifically, we used the timeseries datasets regarding information on each county and the deaths for each county per day. In order to get an overall picture of the data in both time series data, the number of deaths and confirmed cases recorded each day in the US was calculated and plotted (Fig. 1). Both confirmed cases and deaths curves share the same steep, rising slope near April 18. This is a clear indication that COVID-19 has no signs of slowing down.

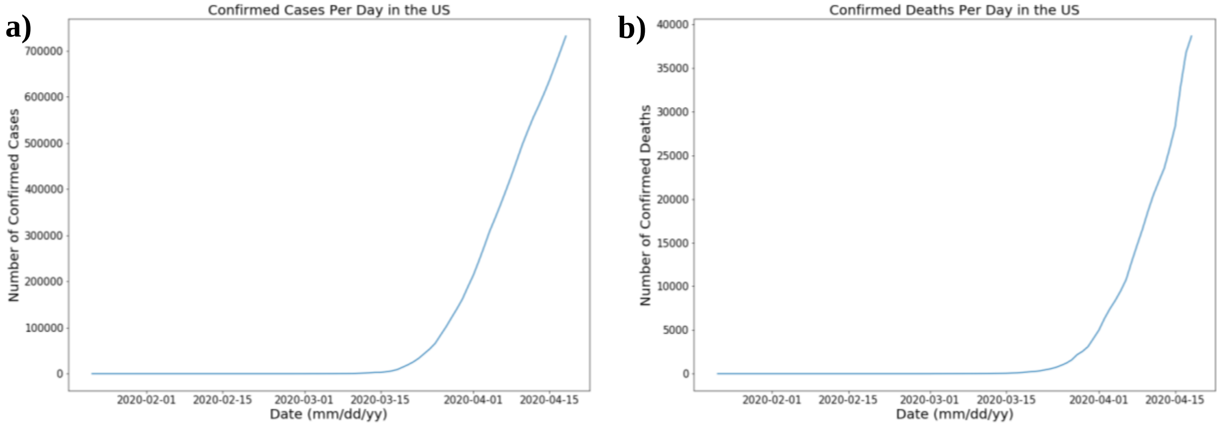

Figure 1: Total Number of Confirmed Cases and Deaths Over Time
a) The total number of confirmed cases started dramatically increasing in February, and shows no signs of stopping. **b)** The cumulative deaths are significantly lower than the number of confirmed cases, but also does not show any signs of stopping or slowing down.

Such trends demonstrate the importance of locating regions in the US that are at high risks while identifying successful mitigation strategies in slowing down the spread of COVID-19.

While comprehensive, the dataset contained many null values and slight inconsistencies. In order to link the information of each county and the number of deaths they were experiencing over time, we merged the two datasets together. This led to the elimination of the data involving US territories, such as Guam due to the severe lack of information regarding their counties. Thus, this study only looks at counties in the 50 states in the United States. Additionally, we found that counties with less than 2 deaths had severely lacking information. The deletion of such counties significantly limited the number of counties in our analysis – from 3140 to 801. We then further removed counties with less than 50 deaths. We found that features were overly sensitive in counties with very few deaths. This further limited our data set from 801 to 81 counties. Even upon, selecting counties with more than 50 deaths, several county features contained null values. Many of them were regarding mortality rate, and since we hypothesized that they were not related to the spread of COVID-19, we dropped such features. A comprehensive list of dropped features can be found in the supplementary Jupyter Notebook. All remaining null values we filled by the mean imputation method (only 40 null values across all features remained up before this point).

In order to capture the rate at which the virus has spread throughout the county, we developed a statistic which we call the 'spread factor' (Eq. 1). The spread factor is normalized by the population of each county to summarize how quickly the virus has spread – not the number of deaths in each county. Upon plotting the distribution of spread factors across all 81 counties, we found some counties to have abnormally large spread factors (Fig. 2). We decided to remove them during model development, as our models were extremely sensitive to these outliers. We do, however, add these outliers back into the testing dataset to ensure that we are capturing the rapid spread of the virus in these counties.

$$Spread\ Factor = \frac{Deaths\ between\ 4/12/20\ \ and\ 4/18/20}{7\ days} * 100{,}000. \qquad (Eq. 1)$$

We then conducted a 70-30 train/test split on our cleaned data for further analysis. After plotting a correlation heatmap for the different features, we found that many of the population features were heavily correlated. We thus remove several of the population features, as they present redundant information. We then found that there were two additional main sources of correlation.

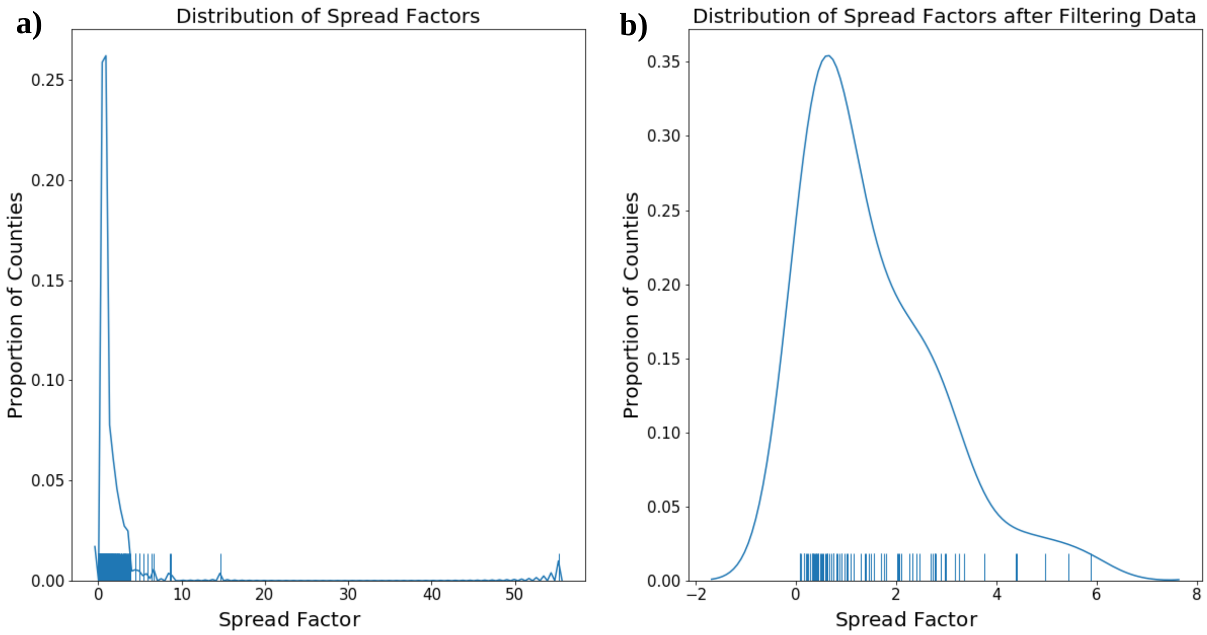

Figure 2: Distribution of Spread Factors
a) The distribution of spread factors immediately after cleaning the data indicate that there are some counties with abnormally high spread factors. These severely impact the model development process. **b)** Upon removing the outliers (counties with spread factors greater than 10), the distribution becomes much more uniform. The resulting distribution is skewed right.

In Fig. 3, the large white square towards the upper left is a representation of the health of the residents in the county, while the large white square towards the center of the heatmap is a representation of the hospital infrastructure in the county. We created another feature matrix removing redundancies in both of these groups and plotted a similar correlation matrix (Fig. 4). This feature matrix was shown to contain features that had little to no correlation to other features. We train and develop two different models on each of the feature matrices, and label them as the 'full' and 'simple' models.

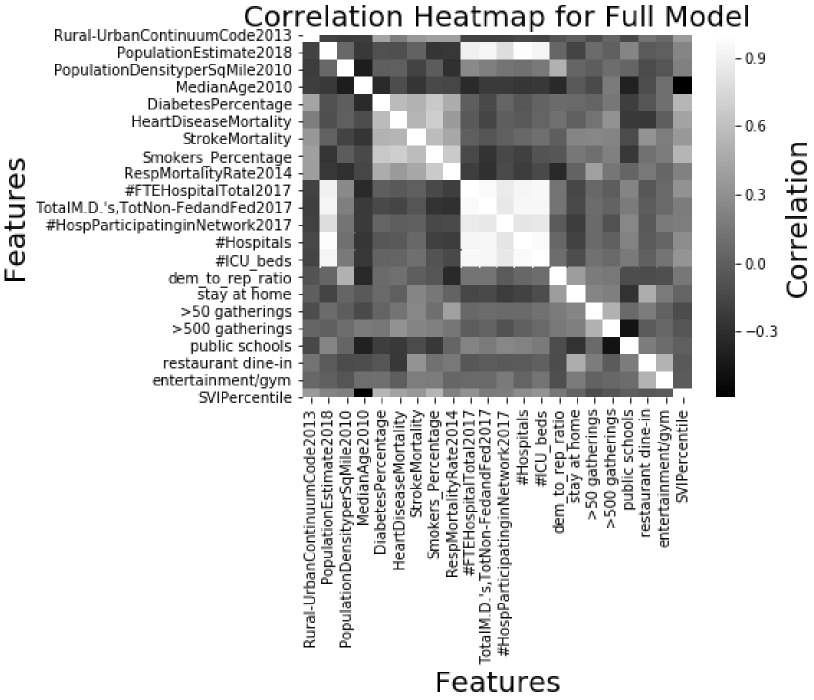

Figure 3: Correlation Heatmap for Full Model Feature Matrix
The correlation heatmap indicates that there are some redundancies involving features that describe the health of the county and the infrastructure of the hospital system.

Description of Methods

A Lasso model was used to determine which features were the most important in predicting the spread factor. In order to find the optimal alpha parameter, 5-fold cross validation was performed. Two different models were created: one for the full feature matrix and the other for the simple feature matrix as described above.

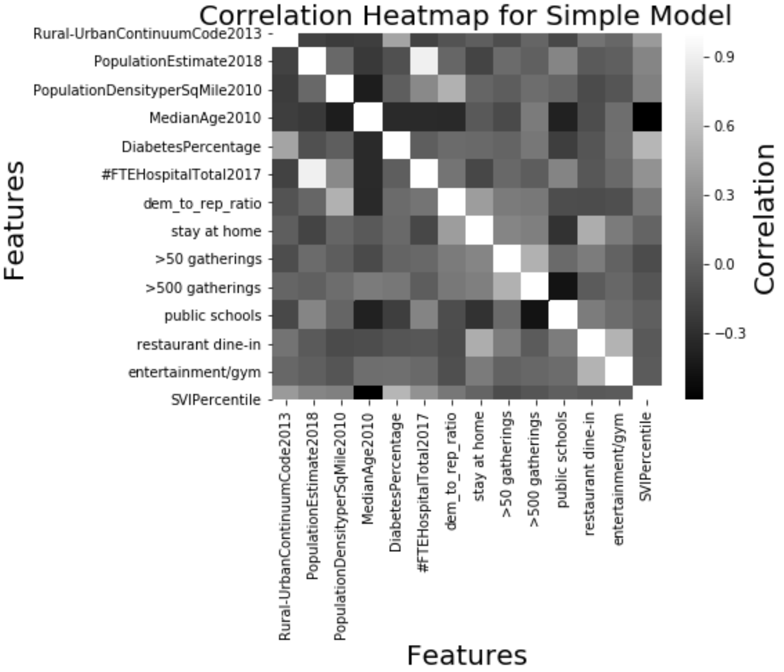

Figure 4: Correlation Matrix for Simple Model Feature Matrix The correlation heatmap indicates that all of the features are not highly correlated. Only essential and independent features have been extracted from the full feature matrix.

The method of classification is mainly used to address the following: will a region be at a relatively low or high risk of COVID-19 based on select features. There were three methods of classification performed where the form of the feature matrix varied. The first method involved using the top three features decided by the simple lasso regression model. The second method involved performing principal component analysis on the simple model feature matrix from the lasso regression analysis and utilizing the top three principal components. The third method involved utilizing all the features from the simple model feature matrix of the lasso regression analysis.

To answer the overall classification question, we first assigned classifications based on the spread factors. We then used K-means clustering to determine the number of clusters upon which to divide the data. To perform this unsupervised learning analysis, the top three most important features decided by lasso regression would be used for K-means clustering which are: 'stay at home', 'Total Hospitals', 'Population Density per Sq. Mile.' Once the number of classifications was decided, we assigned these classifications according to the distribution of the spread factor data. For example, if clusterings of 2 were utilized, then the data would be split using the median of the spread factors in the training set.

After the assignment of classification labels, a K-nearest neighbors classification was trained according to the three methods as aforementioned. The optimal k-parameter for each KNN classification model was found by performing cross validation and calculating the mean error percentage for various values of k. The k-value with the minimum cross validation error was selected for the final model which was then applied to the test set in calculating testing accuracy.

Results and Discussion

Lasso Linear Regression

Due to the complex nature of the spreading mechanisms of COVID-19 and other infectious viruses, we performed a Lasso regression on both the full and simple feature matrix to predict the spread factor for a county simply given its characteristics. This would be very powerful in identifying counties that might be especially prone to biological threats. The model was evaluated using the root mean squared error or RMSE (Eq. 2). The RMSE is a powerful evaluation metric, as the error is in the same units as the dependent variable in question (the spread factor).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (Eq. 2)$$

Table 1 summarizes the results for both models developed. We see that the models' testing and training RMSE are reasonable. The spread factor ranges from zero to six, so a RMSE error of around one is decent. The model is not very accurate; however, it is able to predict the risk level of various counties. Overall, we find that both models perform very similarly, but the testing RMSE for the simple model is slightly more accurate than the full model. This might be due to chance, or it might be due to the fact that the simple model removes some noise that is arising from the extra features in the full model. Thus, we believe this model should not be used as comparison metric between counties of similar spread factors, but rather a rough guide to predict the general safety of various counties during a pandemic.

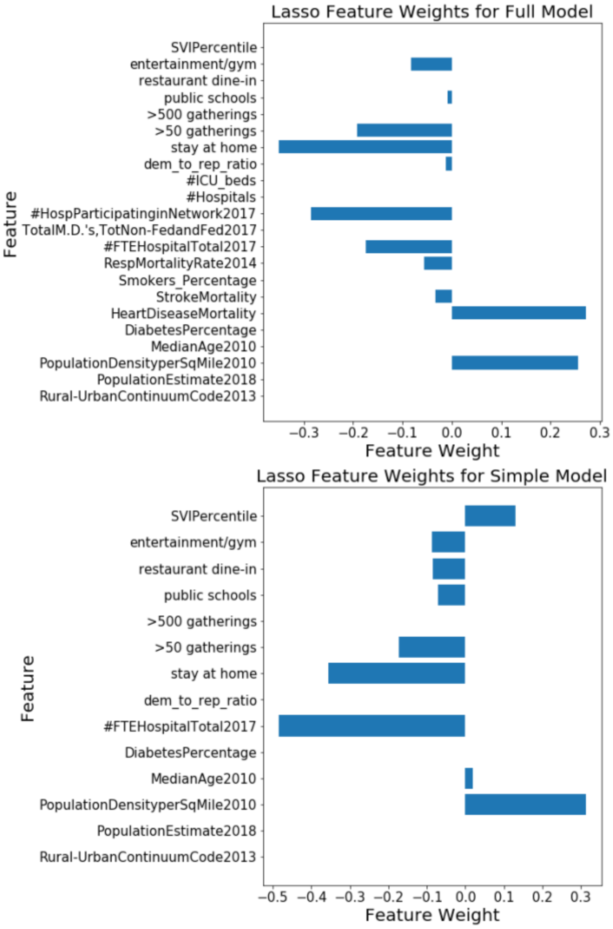

Figure 5: Weights for Each Feature Lasso Models
In the simple and full model, the most important weights are the population density, stay at home date, and total number of hospitals.

Table 1: Summary of Lasso Regression Results
The RMSE for both models are quite similar; however, the simple model has a slightly better testing RMSE. The optimal α parameter for both models were found using 5-fold cross validation.

Model	Optimal α Parameter	Mean Cross Validation RMSE	Training RMSE	Testing RMSE
Simple Model	.10	1.20	1.1	0.79
Full Model	.12	1.21	1.1	0.87

One reason why we employed a Lasso regression is to determine which features are most impactful in the spread of the virus. Figure 5 plots the weights of each feature in both models. As we might have expected, some of the most important features in both models are 'Stay at Home', 'Population Density', 'Hospital Total' (a measure of hospital infrastructure), and 'Gatherings With > 50 People'. Some of the redundant features in the full model have starkly different weights (i.e. 'Heart Disease Mortality' and 'Diabetes Percentage'). We were surprised to see that the median age had no impact on the model despite the virus adversely affecting the elderly.

Clustering

The results of K-means clustering for varying k is depicted in Fig. 6 which plots the sum of squared distances from the nearest cluster mean versus the number of clusters formed. We see that clusters of four are the most appropriate. The small decrease in the sum of squared distances for clusters larger than four is an indication of likely overfitting. A visual depiction of the training clusterings is plotted in Fig. 7. Furthermore, we created a box plot of the spread factors versus clustering prediction to understand which classes have higher risk factors (Fig. 8). The high degree of overlap between clusterings plot was highly concerning. In an ideal case, the four clusterings would represent distinct groups of spread factors. Nevertheless, the four classification labels were assigned according to the quartile percentages. As such, Class 1 represented the lowest 25% of the spread factors, Class 2 represented the second lowest 25% of the spread factors, and so on to create four classifications.

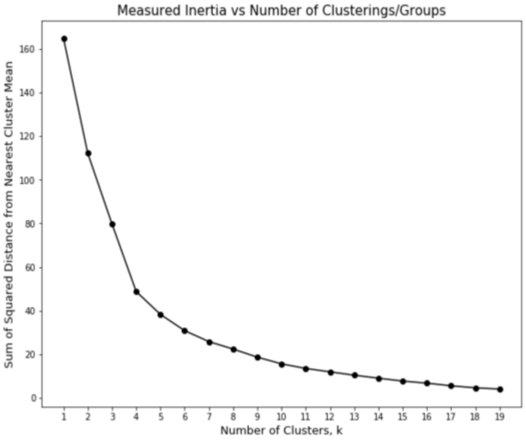

Figure 6: Measured Inertia Versus k for K-Mean Clustering
The summed squared distances from nearest cluster mean does not decrease as dramatically after four clusters. This hints that larger clusters will lead to overfitting.

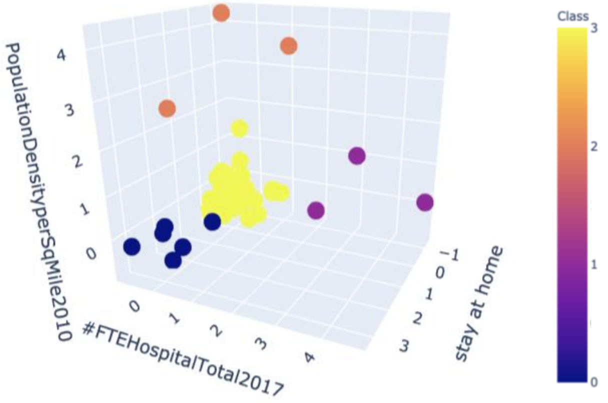

Figure 7: Scatter Plot of K-means Clustering for Training Dataset
Visually, the presence of four distinct clusters is evident. The largest cluster is centered around zero, as these counties have no distinguishing features.

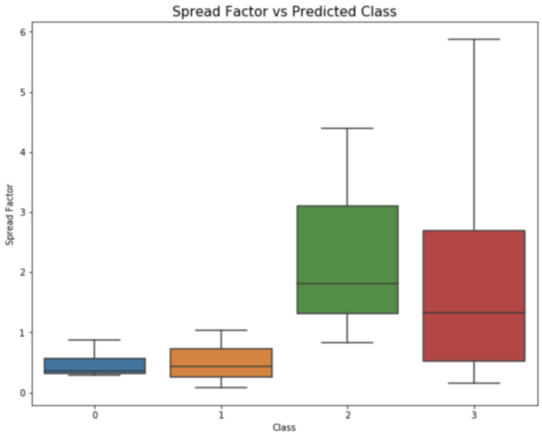

Figure 8: Spread Factor vs Predicted Class
Classes zero, one, and three have a significant amount of overlap in spread factors. This is highly concerning, as these classes would ideally represent counties with different spread factors.

The results of applying Method 1 KNN classification based on these four classifications is depicted in Fig. 9. It is clear that classification predictions based on clusterings of four is unreliable as cross validation mean error percentages were as high as 70%. Therefore, the number of clusterings was reduced to two in the hopes of decreasing the complexity of the classification, resulting in a higher classification accuracy. In the end, the data was labeled as either "low" or "high" risk depending on whether the spread factor was below or above the median value of the spread factors in the training set. These particular set of classification was applied to all three methods involving various feature matrices.

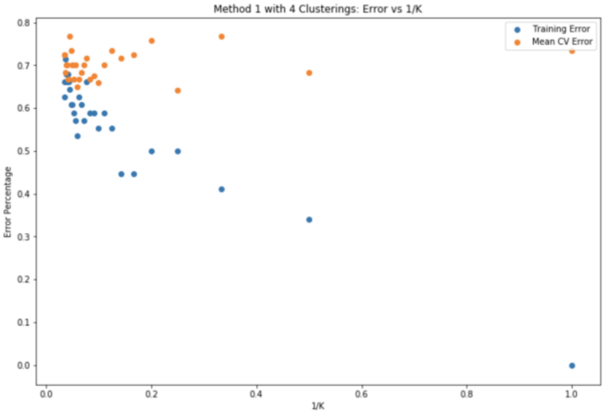

Figure 9: Training and CV Errors for Method 1 with Four Clusterings

The classification error rate is plotted with respect to the k parameter chosen by the KNN classification. The large CV errors indicates that our model does not capture many of the complexities in the data.

The optimal k parameter for each of the methods were found using 5-fold cross validation. Table 2 summarizes the optimal k-parameter, training accuracy, and testing accuracy for each respective method. We observe from Table 2 that method 1 and method 3 have similar accuracies while method 2 had a substantially lower accuracy compared with the other two methods. We attribute the poor performance of method 2 to the principal components. While the first three principle components do capture the largest variance in the simple feature matrix, they do not necessarily capture the variance of the three most important features deemed by the Lasso regression model. It is likely that the first three principle components are capturing the variance of features that are not important in predicting the spread of COVID-19. Ultimately, we conclude that the best classification model/method is method 1 where a KNN classification is based on the top 3 features of the lasso model. This model is quite capable of predicting whether a region in the US is at low or high risk of COVID-19 spread with only 3 statistics of a county. Since method 3 has similar performance to method 1, yet requires more data, we deem method 1 to be more powerful.

Table 2: Summary of KNN Classification Method Results

All three methods had comparable cross validation errors, but performed differently on the training and testing datasets. Model one had the highest training accuracy while model three had the best testing accuracy. Model two was the worst of the three models and likely suffered from a high degree of overfitting.

Method	Optimal k Parameter	Mean Cross Validation Error	Training Accuracy	Testing Accuracy
Method 1: Top 3 Features from Simple Lasso Model	3	0.33	0.85	0.60
Method 2: Top 3 PCA Components	12	0.28	0.68	0.48
Method 3: All Features from Simple Lasso Model	8	0.27	0.73	0.68

As seen through the 3D visualizations of the classifications in Fig. 10, there is an issue with Method 1 where points near (0, 0, 0) are ambiguous. Intuitively, this makes sense, as counties with few discerning or unique characteristics may lead to ambiguous classifications. Although not accurate enough to warrant extra resources to one county over the other, this model is able to predict counties with abnormally large spread factors. At a minimum it performs better than a random assignment of a high or low risk to various counties in the US.

Figure 10: Visual Depiction of Classification Method 1

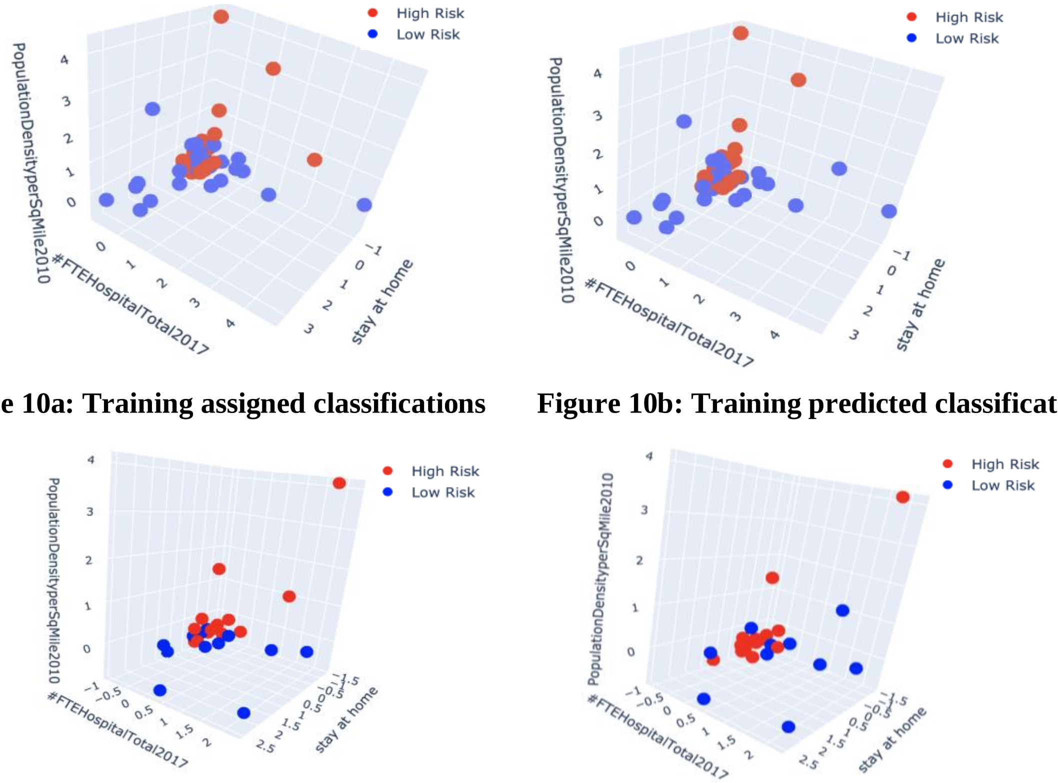

Figure 10a: Training assigned classifications **Figure 10b: Training predicted classifications**

Figure 10c: Testing assigned classifications **Figure 10d: Testing predicted classifications**

While both the Lasso regression and classification methods are able to predict the spread and risk of the virus in various counties in the US, there are severe limitations in both models. In cleaning the dataset, we omitted any counties with fewer than 50 deaths. Thus, we are not analyzing smaller counties in all of our models. This can have serious ethical dilemmas when using this model to allocating funding and resources to fight the virus. Since both of our models marginalize smaller counties, we do not recommend using them as a quantitative method in allocation resources. It is also important to note that our models are only accurate for the United States, and counties outside of the US may not be applicable for our models. In order to create similar models for counties outside of the US, we would need to collect and process similar data for the region in question.

Conclusion and Main Findings

Upon cleaning and removing redundant features from the original dataset, we conducted a Lasso regression analysis. We found that the most accurate regression model had a training RMSE error of 0.79, and the most important features in predicting the spread of the virus were the population density, hospital infrastructure, and the day which the governor implemented stay-at-home policies. Our classification attempts indicated that Method 1 of classification has the overall best performance as it has one of the highest overall accuracies (across the training and testing set). This model's prediction was only based on a minimal 3 features as deemed most important from lasso regression. Such a classification may be used in predicting the relative risk of COVID-19 in regions within the US and such information may be useful when trying to gauge which areas should receive the most resources in terms of mitigating the spread of COVID-19. To perhaps improve this classification model, more useful features could be found from other datasets and implementation of these features should increase the reliability and accuracy of the developed model. One example of such a feature might be the pervasiveness of particular health conditions that are strongly linked with the virus in counties across the US. Another future direction for this project might be to make our model more generalizable by incorporating data from cities or counties from around the world.