Interpretable Kernel Regression for COVID-19 Death Rate Prediction

Skyler Sprecker

ABSTRACT

The COVID-19 pandemic has strained healthcare systems across the United States. This study aims to identify factors that impact the COVID-19 death rate using kernel ridge regression. Because kernel ridge regression models are often difficult to interpret, we used Permutation-based Variable Importance (PVI) to see which factors most correlated to the death rate, and therefore are the most important. Our results show that the percentage of a population that is Black, the percentage that is vaccinated, and the elderly population percentage were the three most important factors of the 13 studied. Additionally, we show that PVI is robust at picking out the most important features across a range of alpha and gamma hyperparameter values, but can struggle ranking less important features consistently.

CCS CONCEPTS

• **Machine Learning**; • **Kernel Ridge Regression**; • **Permutation-based Variable Importance**;

KEYWORDS

Machine Learning, Kernel Ridge Regression, Permutation-based Variable Importance

ACM Reference Format:

Skyler Sprecker. 2018. Interpretable Kernel Regression for COVID-19 Death Rate Prediction. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

- xxx

1 INTRODUCTION

COVID-19 first appeared in China in late 2019 and spread worldwide throughout early 2020, being declared a pandemic in March of that year. It is a respiratory disease caused by the highly contagious SARS-CoV-2 virus. Given the public health threat COVID-19 poses, being able to estimate the death rate on a county level would be valuable to public health experts as they work to combat the spread of this disease. The initial idea was to collect data on different public health factors and then use machine learning to analyze them and determine which ones most strongly correlated to the COVID-19 death rate. At first linear regression was the technique chosen to perform the analysis, and while it did yield results, there was hope that a more powerful technique would result in a better, more accurate model.

It was decided to use kernel ridge regression for the analysis, however this raises at least one issue of its own. With linear regression, the coefficients generated by the regression were being used to see which features most strongly correlated to the COVID-19 death rate. Since kernel ridge regression extrapolates the data into a higher dimensional space, it is not possible to use these coefficients, thereby making the task of seeing which features most strongly correlate challenging.

One option for interpreting kernel regression is Permutation-based Variable Importance (PVI). Used regularly with random forest machine learning models, it takes each feature and shuffles it. The error of the model is then recalculated and the difference is used to determine which features are the most important for the model. Since PVI has not been used much with kernel-based models, it was decided to use this method to investigate its potential.

2 METHODOLOGY

Kernel ridge regression is a machine learning technique where the data is mapped to a higher dimensional space in an effort to make the data linearly separable, represented by the objective function:

$$J = \sum_{i=1}^{n} \Big(\sum_{j=1}^{n} \alpha_j \phi(x_i)^T \phi(x_j) - y_i \Big)^2 + \lambda \sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_j \alpha_k \phi(x_j)^T \phi(x_k)$$

. The kernel function is a function that returns the dot products in the higher dimensional space, of vectors taken as inputs. Given x, z $\in$ X and a mapping function $\phi : X \to \mathbb{R}^n$, then we have a kernel function k(x, z) = $\langle \phi(x), \phi(z) \rangle$.[6] One of the most common kernels, and the one used for this study, is the Radial Basis Function (RBF) kernel, which is $K(\mathbf{x}, \mathbf{z}) = \exp(-\gamma \|\mathbf{x} - \mathbf{z}\|^2)$, where γ is a hyperparameter that can be tuned to improve the accuracy of the kernel function and therefore model. Kernel functions take advantage of the "kernel trick" to avoid being very computationally expensive. The kernel trick allows us to avoid having to explicitly define the mapping function, but can instead calculate the dot product of x and z. For example, let $\mathrm{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\phi(x) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{bmatrix}$, $\mathrm{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$, $\phi(z) = \begin{bmatrix} z_1^2 \\ z_2^2 \\ \sqrt{2}z_1z_2 \end{bmatrix}$. Then $\phi(x)^T\phi(z) = (x_1z_1)^2 + (x_2z_2)^2 + 2(x_1z_1)(x_2z_2) = (x_1z_1 + x_2z_2)^2 = (x^Tz)^2 = k(x, z)$. This allows us avoid having to make calculations in a higher dimensional space, this preventing the computational cost from becoming excessive.

One problem with kernel ridge regression is that it is hard to interpret because of the mapping to a higher dimensional space. There are ways to interpret the results, but the model itself is not very interpretable. One way of interpreting the results of the kernel model is Permutation-based Variable Importance. PVI works by

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

calculating the original error of the model, then taking each feature of the model, shuffling (permuting) it, and recalculating the error. The idea is that by shuffling the feature, it breaks any association between that feature and the label. Thus, if the error after permuting is significantly different from the original error, that suggests the permuted feature is greatly important to the model. Likewise, if the error after permuting is not much different from the original error, the permuted feature is less important to the model.[5] PVI is largely used for random forest machine learning models, but not often for kernel regression models, which is why it was selected for this study.

3 EXPERIMENTS

3.1 Data Preparation

The first step in the experiment process was collecting data on different public health demographics to determine which most strongly correlate to the COVID-19 death rate. The features initially chosen were population density, obesity rate, percentage of the population that smokes, diabetes rate, elderly population percentage, and the vaccination rate. These factors were chosen as they generally represented how urban or rural a county is, the percentage of the population that have health issues that can make it harder to fight disease, and the percentage of the population that has been vaccinated to prevent severe disease and death from COVID-19.

After reading some of the available literature about other studies on COVID-19 death rate, the racial demographics of each county and the percentage of the population with limited access to healthy foods were added to represent the racial makeup of each county and whether or not they had the ability to easily purchase healthy foods. This brought us to 13 features total. The data was collected from the following places:

- Population density data came from WorldPopulationReview.com.
- Obesity, smoking, and diabetes rates, racial demographic data, and the percentage of the population with limited access to healthy food all came from countyhealthrankings.org, which is part of the University of Wisconsin Population Health Institute.
- Percentage of the population that is elderly data came from Census.gov.
- Vaccination rate data came from the Springfield Missouri New-Leader website.
- Death rate data came from usafacts.org. It has the total number of cases and deaths for each county, from which we can calculate the death rate.

The data was collected into an Excel worksheet, which was then saved as a CSV file to be read in by the code. Once the data was collected, the features were standardized so that population density is on the same scale as the rest of the data.

3.2 Kernel Model Creation and Variable Permutation

Once the data was prepared, the kernel ridge regression was set up and performed with 2/3 of the data used for training and the remaining 1/3 data used for testing, an alpha of 0.1, and a default gamma value, which for the scikit-learn kernel ridge regression function is 1/the number of features so 1/13 for this study. The regression was performed 20 times with data randomly split each time to avoid any potential biases from the inherent order of the data. The mean squared error (MSE) and coefficients of the regression were recorded after each regression and averaged after all regressions completed to mitigate any potential erroneous results. The MSE of the training data was 0.00003932 and the MSE of the testing data was 0.00007091. Both values were sufficiently small to provide confidence that the model was accurate.

When the model has been successfully created, it was then time to score the variables using PVI. The open-source scikit-learn Python library has a PVI method in its inspections module that was used to perform the PVI. Following the same process as the model creation, the PVI was performed 20 times for each feature, with the training and testing data randomly split each time and the scores returned averaged. To determine how robust the results returned by PVI were, we then decided to vary both the alpha and gamma hyperparameters of the kernel model to see how this affected the results of the PVI. For alpha values ranging from 0.1 to 0.75 were chosen and for gamma values between 0.01 and 0.1 were used.

3.3 Results

Using an alpha value of 0.1 and a gamma value of 1/13, the four features with the highest score returned by PVI were the percent of the population that was Black, the percentage of the population that was vaccinated, the percentage of the population that was elderly, and the smoking rate of the population of a county. The full rankings returned by PVI can be seen in Table 1.

Table 1: PVI scores for alpha=0.1, gamma=1/13

Feature	PVI Score	PVI Standard Deviation
Percent Black	0.36408	+/- 0.0291
Vaccination Percentage	0.21764	+/- 0.02705
Elderly Population Percentage	0.18251	+/- 0.02054
Smoking Rate	0.15167	+/- 0.023
Percent White	0.14195	+/- 0.01779
Percent Hispanic	0.11492	+/- 0.0168
Obesity Rate	0.06493	+/- 0.01451
Percent American Indian	0.0592	+/- 0.01581
Percent Asian	0.04955	+/- 0.01497
Diabetes Rate	0.0377	+/- 0.01412
% Limited Access to Healthy Foods	0.03482	+/- 0.01348
Population Density	0.02458	+/- 0.01133
Percent Pacific Islander	0.01773	+/- 0.00963

Across alpha and gamma values, the ranking of the top three features was unchanged, with the exception of that when alpha value = 0.75, vaccination percentage overtook percent black as the feature with the highest score. Figures 1 and 2 show the variable importance scores for each feature across alpha and gamma values. It is interesting to note that as the alpha value increased, the variable importance scores generally went down. The most dramatic change happened for the percent black feature, with its score when alpha = 0.75 less than half of what it was when alpha = 0.1. On the other hand, there was a positive correlation between gamma values and

feature scores, as increasing the gamma values generally made feature importance scores increase. This is most dramatic for the percent white feature, which saw its variable importance score dramatically increase as gamma increased. The rank of each features across alpha and gamma values and be seen in Figures 3 and 4.

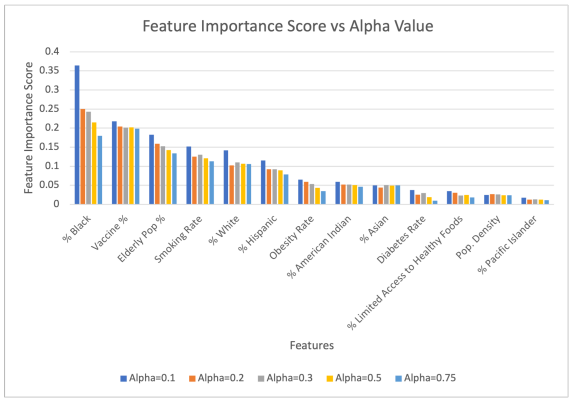

Figure 1: Feature Importance vs Alpha Value

It is interesting to note that as the alpha value increased, the variable importance scores generally went down. The most dramatic change happened for the percent black feature, with its score when alpha = 0.75 less than half of what it was when alpha = 0.1. On the other hand, there was a positive correlation between gamma values and feature scores, as increasing the gamma values generally made feature importance scores increase. This is most dramatic for the percent white feature, which saw its variable importance score dramatically increase as gamma increased. The rank of each features across alpha and gamma values and be seen in Figures 3 and 4.

As Figures 3 and 4 show, while PVI was good at consistently ranking the top few features across alpha and gamma values, it was not nearly as consistent when it came to the ranking of the lower importance features across alpha and gamma values. Interestingly, though, Figure 3 shows that when features did change rankings across alpha values, they so among groups of features. For example,

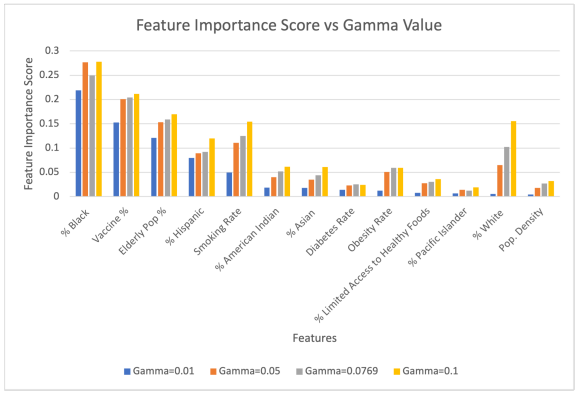

Figure 2: Feature Importance vs Gamma Value

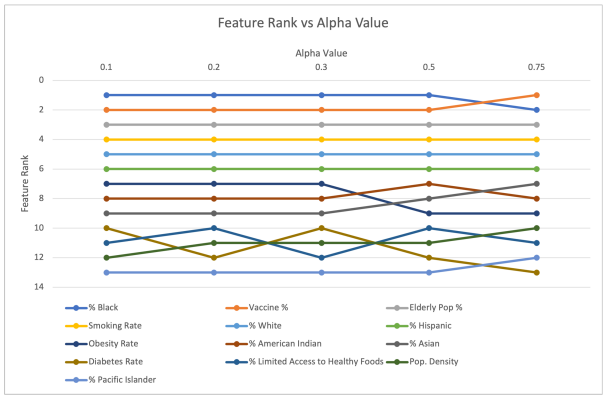

Figure 3: Feature Rank vs Alpha Value

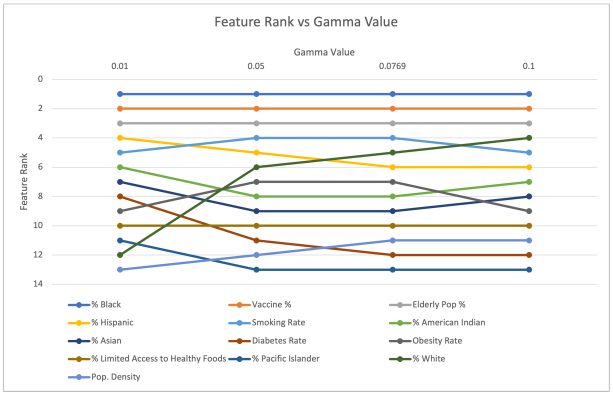

Figure 4: Feature Rank vs Gamma Value

while diabetes rate, percent with limited access to healthy foods, and population density change rankings every time the alpha value changed, they only changed places with each other. This indicates that they all have very similar variable importance scores.

3.4 Discussion

The results found show that Permutation-based Variable Importance is capable of successfully picking out the most important features from a kernel ridge regression model, across a range of hyperparameter values. But in order to make sure that the important features according to PVI are actually important, we must see if there is other research out there that backs up our findings. The study "Ensemble machine learning of factors influencing COVID-19 across US counties" found that "Specifically, the CDC measures for minority populations ... and proportion of Black- and/or African-American individuals in a county were the most important features for per capita COVID-19 cases". [3] It used ensemble machine learning to look at over 100 features about counties in the United States and found that, among other things "10% increase in the proportion of Black- and/or African-American individuals in a county is associated with increases total deaths". This aligns with our findings that the percent of a county's population that is black has a positive correlation to the COVID-19 death rate of the county.

Vaccination percentage had the second highest variable importance score, but this almost certainly indicates a strong negative correlation to the death rate, as COVID-19 vaccines are upwards of 95% effective at preventing hospitalization and death from the disease.[2] This does highlight one issue with PVI in that because the variable importance scores can only be positive, it does not indicate whether a feature has a positive or negative correlation to the label. However, because of the available information about COVID-19 vaccines and their effectiveness, we can assume that there is a negative correlation between the percentage of the population that is vaccinated and the death rate.

The third highest variable importance score was elderly population percentage, which makes sense as a person's immune system declines as they age, meaning the elderly are more susceptible to disease. A study from early on in the pandemic found that the "percentage of people aged >70 years", among other features, "plays an important role in predicting COVID-19 occurrences."[4] And while COVID-19 occurrences are not deaths, more occurrences of the disease will lead to more people dying because of it.

Finally, the smoking rate of the population was the feature with the fourth highest score across alpha values. This makes sense, as smoking damages the lungs, and COVID-19 is primarily a respiratory disease that attacks the lung. A study in the United Kingdom found that "Compared with those who had never smoked, current smokers were 80% more likely to be admitted to hospital and significantly more likely to die from COVID-19". A statement from the lead researcher said "Our results strongly suggest that smoking is related to your risk of getting severe COVID, and just as smoking affects your risk of heart disease, different cancers, and all those other conditions we know smoking is linked to, it appears that it's the same for COVID. So now might be as good a time as any to quit cigarettes and quit smoking." [1]

Permutation-based variable importance for kernel models is not perfect, however. The same study that found that more elderly people increases COVID-19 occurrences also found that population density and "prevalence of comorbidities" also were important features in predicting COVID-19, whereas the results we gathered ranked the diabetes rate and population density among the four least important features. This indicates that there is still work to do to improve the accuracy of permutation variable importance.

REFERENCES

[1] Ashley K Clift, Adam von Ende, Pui San Tan, Hannah M Sallis, Nicola Lindson, Carol A C Coupland, Marcus R Munafò, Paul Aveyard, Julia Hippisley-Cox, and Jemma C Hopewell. 2022. Smoking and COVID-19 outcomes: an observational and Mendelian randomisation study using the UK Biobank cohort. *Thorax* 77, 1 (2022), 65–73. https://doi.org/10.1136/thoraxjnl-2021-217080 arXiv:https://thorax.bmj.com/content/77/1/65.full.pdf

[2] Dan-Yu Lin, Yu Gu, Bradford Wheeler, Hayley Young, Shannon Holloway, Shadia-Khan Sunny, Zack Moore, and Donglin Zeng. 2022. Effectiveness of Covid-19 Vaccines over a 9-Month Period in North Carolina. *New England Journal of Medicine* 386, 10 (2022), 933–941. https://doi.org/10.1056/NEJMoa2117128 arXiv:https://doi.org/10.1056/NEJMoa2117128

[3] David McCoy, Whitney Mgbara, Nir Horvitz, Wayne M Getz, and Alan Hubbard. 2021. Ensemble machine learning of factors influencing COVID-19 across US counties. *Scientific reports* 11, 1 (2021), 1–14.

[4] Mihir Mehta, Juxihong Julaiti, Paul Griffin, and Soundar Kumara. 2020. Early stage machine learning–based prediction of US county vulnerability to the COVID-19 pandemic: machine learning approach. *JMIR public health and surveillance* 6, 3 (2020), e19446.

[5] Christoph Molnar. 2022. Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/feature-importance.html

[6] Drew Wilimitis. 2019. The kernel trick. https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f#:~:text=The%20%E2%80%9Ctrick%E2%80%9D%20is%20that%20kernel,the%20data%20by%20these%20transformed