Identification and Analysis of Health Vulnerabilities in Los Angeles

PICMath

Jeffrey Marin, Leonard Garcia, Anh Le, Derek Lin
Department of Mathematics
California State University Los Angeles

Advisor Dr. Jie Zhong, Professor at CSULA
Eva Pereira, Chief Data Officer from Los Angeles city

May 15, 2022

Identification and Analysis of Health Vulnerabilities in Los Angeles

Abstract

CalEnviroScreen (CES) is a mapping tool that allows people to identify communities in California that are most affected by different sources of pollution. CES uses socioeconomic, health, and environmental data to identify communities that are disproportionately burdened by pollution. The purpose of this work is to identify the most vulnerable communities in Los Angeles and make a comparison to COVID case rates, COVID death rates, low birth rate, Asthma, and cardiovascular disease. To identify the most vulnerable communities, we found the high pollution groups that have the highest negative health rates. These high pollution groups include toxic releases from facilities, Lead, and drinking water contamination. We define vulnerable communities as those which belong to all three high pollution groups. We used statistical inference to find the top features that are correlated to the COVID case rate and COVID death rate. It was found that low education and Asthma heavily impact both COVID cases and death rates. We also found that linguistic isolation and poverty disproportionately affect COVID death rates. We use different machine learning techniques to predict the case rate, death rate, low birth weight, asthma, and cardiovascular diseases such as random forest, XGBoost, lasso regression, and ridge regression. We were able to identify the best prediction models for each of our health problems.

1 Introduction

The Office of Environmental Health Hazard Assessment (OEHHA) has created the California Communities Environmental Health Screening Tool: CalEnviroScreen 4.0 (CES). CES is a mapping tool for identifying impacted communities by taking into account socioeconomic status, health, and environmental data. This tool uses 21 statewide indicators to characterize both pollution burden and population characteristics. It uses a scoring system in which the percentiles are averaged for this set of indicators.

Our objective is to conduct an exploratory analysis of the CES data to better understand the health of our local communities. We will use census tract-level data to identify the most vulnerable communities in Los Angeles. We will define the most vulnerable communities as those that bear a disproportionate burden of pollution, socioeconomic stressors, and health conditions that render them more vulnerable to the impacts of pollution.

2 Results

High vulnerable communities results: Performed an investigation to identify the most vulnerable communities in Los Angeles.

- Vulnerable communities are those that belong to these three high pollution groups:
  1. Toxic releases from facilities: concentrations of chemical emissions.
  2. Lead: potential risk for lead exposure in children.
  3. Drinking water contamination: amount of contaminants in drinking water.
- The most vulnerable communities are in the area around the intersection of the 105 and 110 freeways.

Feature Correlation with results: Performed correlation testing to identify features most correlated to COVID case rate and COVID death rate.

- Features highly correlated with COVID case rate and COVID death rate include:
  1. Low education

2. Poverty
3. Lead
4. Housing burden

- Asthma is disproportionately correlated with COVID case rate.
- Linguistic isolation is disproportionately correlated with COVID death rate.

Best Predictive models results: Developed and tuned various machine learning based predictive models on our health issues.

- COVID case rate: Lasso Regression on normalized data.
- COVID death rate: Random Forest Regression on raw data.
- Asthma: Lasso Regression or Ridge Regression on percentile data.
- Cardiovascular disease: XGBoost or Random Forest Regression on percentile data.
- Low birth weight: Random Forest Regression on percentile data.

3 Methods

3.1 Data

Our data consist of two different data sets: CalEnviroScreen and COVID data. CES has 21 features split into two categories Pollution Burden and Population Characteristics. Two components represent Pollution Burden – Exposures and Environmental Effects – and two components represent Population Characteristics – Sensitive Populations and Socioeconomic Factors.

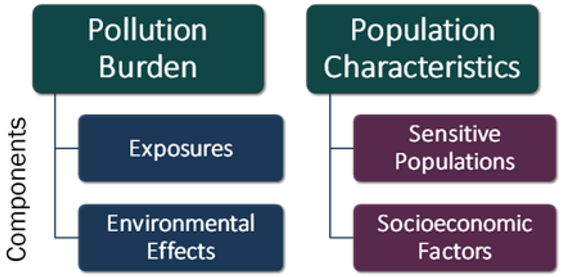

Figure 1: Components for CalEnviroScreen

CES dataset:

- Exposure indicators refer to those pollutants that people may be exposed to through direct contact, such as by breathing contaminated air. Includes 8 features: Air Quality: Ozone, Air Quality: PM2.5, Diesel Particulate Matter, Drinking-Water Contaminants, Children's lead risk from housing, Pesticide use, Toxic releases from facilities, and Traffic impacts.
- Environmental Effects are adverse environmental conditions caused by pollutants. Includes 5 features: Cleanup sites, Groundwater threats, Hazardous waste generators and facilities, Impaired waters, and Solid waste sites and facilities.
- Sensitive populations are populations with physiological conditions that result in increased vulnerability to pollutants. Includes 3 features: Asthma, Cardiovascular disease: Heart attack rate, and Low birth weight in infants.

- Socioeconomic factors are community characteristics that result in increased vulnerability to pollutants. Includes 5 features: Educational attainment, Linguistic isolation, Poverty, Unemployment, and Housing-burdened low-income households. Each of the features includes percentile scores for each of the indicators.

COVID data consist of 15 features:

- City Type, LCity, Community, Label, Source, ShapeSTAre, ShapeSTLen, City, F5 Fully Vaccinated, Population 5, F5 Pop Vaccinated, Cases, Case rate, Deaths, Death rate.
- Using geospatial GIS coordinates, we combined the CES and COVID into one consisting of 134 rows and 63 columns.

To use the data for training the machine model, we split our data into three types: Raw, Percentile, and Normalized Data. Raw data consist of all the columns with the raw value from the CES and COVID data. Percentile Data are all the percentile values in the data set. The normalized data is the raw data. We are normalizing it between 0 and 1.

3.2 Statistical Inference

3.2.1 Hypothesis Testing

Hypothesis testing is a type of statistical inference that uses data from a sample to make conclusions about a population. One begins by considering two hypotheses about the parameter or distribution, the null hypothesis H_0 and the alternative hypothesis H_a. The null hypothesis H_0 is a statement of no difference between the variables. The alternative hypothesis H_a is a claim about the population contradictory to H_0 and what we conclude if H_0 is rejected. This is normally what we are trying to prove. The hypothesis-testing procedure involves using the sample data to determine whether or not H_0 can be rejected.

During the hypothesis-testing procedure, the possibility of errors must be considered. A type I error is the result of rejecting H_0 when H_0 is true. A type II error is the result of accepting H_0 when H_0 is false. The probability of making a type I error is denoted α. The probability of making a type II error is denoted by β. When determining if H_0 should be rejected, one specifies the maximum allowable probability of making a type I error. This is known as the level of significance. A common choice for the level of significance is $\alpha = 0.05$. The p-value is a measure of how likely the sample results are, assuming the null hypothesis is true. If the p-value is less than α, the null hypothesis can be rejected.

Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data. It is denoted by r_s and constrained as follows

$$-1 \leqslant r_s \leqslant 1,$$

the close r_s is to ±1 the stronger the monotonic relationship. The calculation of Spearman's correlation coefficient requires that the data is interval, ratio level, or ordinal, and that it is monotonically related. There is no requirement of normality and is thus a non-parametric statistic.

There are two methods to calculate Spearman's correlation coefficient. The formula for when there are no tied ranks is[3]

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where d_i = difference in paired ranks and n = number of cases. The formula for when there are tied ranks is

$$\rho = \frac{\sum_1 (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}},$$

where i = paired score.

When calculating Spearman's rank coefficient, we must perform a significance test to decide whether, based on the sample, there is evidence that a linear correlation is present. To do so, we let our null hypothesis, H_0, be that there is no monotonic correlation and test it against the the alternative hypothesis, H_a, that there is a monotonic correlation, that is

$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0.$$

Like Spearman's correlation coefficient, Kendall's rank coefficient is a non-parametric statistical measure of the strength of a monotonic relationship between paired data[2]. It is more suitable for smaller sample sizes or data sets with many tied ranks. Moreover, like Spearman's Rank, there are two methods for calculating Kendall's Rank depending on whether or not there are tied ranks. The formula for when there are no tied ranks is

$$\tau = \frac{n_c - n_d}{n(n-1)/2},$$

where n_c is the number of concordant pairs and n_d is the number of discordant pairs. The formula for when there are tied ranks is

$$\tau_b = \frac{S}{\sqrt{[n(n-1)/2 - \sum t_i(t_i-1)/2][n(n-1)/2 - \sum u_i(u_i-1)/2]}},$$

where t_i is the number of observations tied at a particular rank of x and u_i is the number tied at a rank of y. Both correlation tests are used in this study to compare the results and estimate the better fit.

3.2.2 Bootstrapping

Bootstrapping is a statistical method of estimating sampling distributions using random sampling methods. This process creates a bootstrapped data set that is then used for hypothesis testing. This study uses non-parametric bootstrapping which generates data points through the Monte Carlos method. Each sample is created by repeatedly sampling from the original data set, with replacement, and taking the mean of the sampled set.

We chose this method as it is suited for making statistical inferences when parametric inferences are impossible due to the lack of sufficient original data to accurately assume a parametric model. A bootstrapped data set of size $10,000$ is generated for hypothesis testing

3.3 Predictive Models

3.3.1 Ridge Regression

Ridge regression is similar in working to linear regression. The only difference is the addition of the L1 penalty. L2 penalty terms were added to the equation to ensure the shrinking weights of the model to zero or close to zero to ensure the model isn't over-fitting the data[5]. The cost function is

$$\text{cost(w)} = \frac{1}{2*n}\sum_{i=1}^{i=n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{j=D} w_j^2,$$

y_i is the predicted value, $\hat{y}_i$ is the value we try to predict or estimate. w_j^2 is the parameter the machine will learn and adjust by itself. λ is hyper-parameter that we have to adjust it by our-self. We want to minimize

$$\frac{1}{2*n}\sum_{i=1}^{i=n}(y_i - \hat{y}_i)^2,$$

which is subject to

$$\sum_{j=1}^{j=D} w_j^2,$$

The w_j^2 can be represented as

$$w_0^2 + w_1^2 = c^2,$$

The equation above is the equation of the circle with the origin at (0,0) and the radius c. The ridge regression create a solution that minimizes the cost function so that the w_0 and w_1 can only be from points within or on the circumference of the circle. The equation we tries to minimize becomes

$$f(w_0, w_1, \lambda) \sum_{i=1}^{i=n} (y_i - w_0 - w_1 x_i) + \lambda(w_0^2 + w_1^2 - c^2),$$

we are try to minimize the cost function so we have to set the value for λ if λ is 0 we have a linear regression equation. The machine will learn the value for w and auto-adjust itself to give us the best-optimized value for the function with the lowest cost.

We use grid search CV to select the best λ value between 0.01 to 25 to find out the best value to optimize the R^2 value. We run ridge regression on raw data, percentile data, and normalized data to predict the new value for Covid case rate, death rate, asthma, cardiovascular disease, and low birth rate.

3.3.2 Lasso Regression

Ridge regression has one small flaw when two features are highly correlated with each other; the weights are equally distributed between those two features implying there will be two features with the lesser value of coefficients rather than one feature with strong coefficients. For this reason, we use Lasso. Lasso's cost function is

$$\text{cost(w)} = \frac{1}{2 * n} \sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{j=D} |w_j|,$$

where y_i is the predicted value and $\hat{y}_i$ is the value we try to predict or estimate. $|w|$ is the parameter the machine will learn and adjust by itself. λ is hyper-parameter that we have to adjust it by our-self.that we have to adjust it by our-self. We want to minimize

$$\frac{1}{2 * n} \sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2,$$

subject to

$$\sum_{j=1}^{j=D} |w_j|,$$

the lasso regression shrinks the coefficient to zero. This is important when there are large number of features to model the the machine learning algorithm. The way it does is by trying to minimize the cost function i.e. the Residual sum of squares subject to a constrain

$$\sum_{j=1}^{j=D} |w_j| \leqslant c,$$

we can expand this equation for w_0 and w_1 we get

$$w_0 + w_1 \leqslant c,$$

Lasso regression minimizes the cost subject to constraints, but for the lasso, when we plot the points for the constrain, there will be a diamond that is created with (0,0) as the center. We will have the contour plot with the Residual sum of squares value which is bound to be within or on the circumference of the diamond.Since the chances of the contour plot touching the end points of the diamond are quite high, thereby driving the weights for certain features zero[5].

We use grid search CV to select the best λ value between .0001 to 10 to find out the best value to optimize the R^2 value. We run lasso regression on raw data, percentile data, and normalized data to predict the new value for Covid case rate, death rate, asthma, cardiovascular disease, and low birth rate.

3.3.3 Random Forest Regression

The random forest regression model combines the output of multiple decision trees to reach a single result.Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature[4]. Here is the formula representation of when we calculate the feature importance of two nodes

$$ni_j = w_j C_j - w_{\text{left}(j)} C_{\text{left}(j)} - w_{\text{right}(j)} C_{\text{right}(j)},$$

where ni_j is the importance of node j, w_j is weighted number of samples reaching node j, C_j is the impurity value of node j, left(j) is child node from left split on node j, right(j) is child node from right split on node j. Each column in our data will represent a feature. In this case, the left node will be one feature, and the right node will be another feature. We calculate the value of importance between two features.

Then we calculate the importance of each feature on a decision tree

$$fi_i = \frac{\sum_{j:\text{node j splits on feature i}} ni_j}{\sum_{k \in \text{all nodes}} ni_k},$$

where fi_i is the importance of feature i, ni_j is the importance of node j. We normalized the value between 0 and 1 by dividing by the sum of all feature importance values

$$normfi_i = \frac{fi_i}{\sum_{j \in \text{all features}} fi_j}.$$

By combining all the trees, we will get the formula for the random forest regression. Here is the final formula presentation

$$RFfi_i = \frac{\sum_{j \in \text{all trees}} \text{norm}fi_{ij}}{T},$$

$RFfi_i$ is the importance of feature i calculated from all trees in the Random Forest model, $\text{norm}fi_{ij}$ is the normalized feature importance for i in tree j, T is total number of trees.

We use 100 decision trees to create our random forest regression. We run lasso regression on raw data, percentile data, and normalized data to predict the new value for COVID case rate, death rate, asthma, cardiovascular disease, and low birth rate. We select the best result with the lowest root mean square error value.

3.3.4 XGBoost Regression

XGboost is a supervised learning method based on function approximation by optimizing loss functions and applying several regularization techniques. Here is the equation for the loss function for XGboost

$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t),$$

where $y_i, \hat{y}_i^{(t-1)}$ are the result of the actual value and predicted value. $\Omega(f_t)$ is the hyper-parameter we have to adjust. $f_t(x_i)$ is the regularization that we try to minimize by applying the Taylor formula[1]. After using second-order Taylor approximation, we have

$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i)\right) + \frac{1}{2} h_i f_t^2(x_i) + \Omega(f_t).$$

We can minimizing a simple quadratic function as below

$$\text{argmin}_x Gx + \frac{1}{2}Hx^2 = -\frac{G}{H}, H > 0 \qquad \text{min}_x Gx + \frac{1}{2}Hx^2 = -\frac{1G^2}{2H}.$$

We apply XGBoost on raw data, percentile data, and normalized data to predict the new value for COVID case rate, death rate, asthma, cardiovascular disease, and low birth rate to get the best root mean square error value.

4 Detailed Presentation of Results

4.1 Indentifying Vulnerable Communities

We began our analysis by looking at the areas of Los Angeles with the highest pollution. Our data set included measurements for 13 different pollutants each of which was measured differently. For this reason, we focused on the percentile values. For every pollutant in our data set, we defined high pollution areas as those that fell in the 90th and above percentile. For these high pollution areas, we calculated the average rates of the adverse health conditions we had in our data set. We then calculated the averages for all of Los Angeles and compared the two.

We identified three high pollution groups as having amongst the highest average adverse health rates:

1. Areas with high toxicity-weighted concentrations of modeled chemical releases to air from facility emissions and off-site incineration.

2. Areas with high potential risk for lead exposure in children living in low-income communities with older housing.

3. Areas with large amounts of contaminants in the drinking water.

We defined vulnerable communities as those that belonged to all three of these high pollution groups. For these vulnerable communities, we calculated the average rates of adverse health conditions and compared them to the average rates of Los Angeles as a whole.

Vulnerable Communities vs. All of Los Angeles		
Feature	**Mean of Vulnerable Communities**	**Mean for all of Los Angeles**
Emergency department visits for asthma per 10,000	100.14	57.29
Percent low birth weight	7.16	5.47
Emergency department visits for heart attacks per 10,000	21.47	13.91
COVID Case Rate: Number of cases per 100,000 residents	34,616	25,085
COVID Death Rate: Number of cases per 100,000 residents	324	260

Table 1: Rates of health issues in Los Angeles

For every health condition in our data set, these vulnerable communities had average rates significantly higher than the average for Los Angeles.

After we developed a definition of vulnerability, we identified the location of the most vulnerable communities. Since we designated a community as being vulnerable using three percentile values, we added up the three percentile values to rank the most vulnerable. Focusing on the highest ranking, that is the most vulnerable, we see that these communities are located in the same general area.

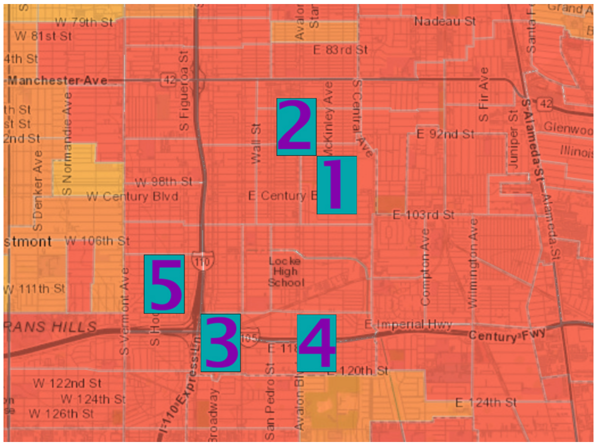

Figure 2: Location of most vulnerable communities

The most vulnerable communities are located in the area surrounding the intersection of the 105 and 110 freeways. Figure 2 has been marked to include the top 5 highest scoring, though all 10 highest scoring communities are located in the same general area.

4.2 Feature Correlation

Once we identified the most vulnerable communities of Los Angeles, we performed hypothesis testing on the bootstrapped data to identify monotonic correlation relations between the features, and COVID case rates and COVID death rates. We performed both Spearman's Rank correlation test and Kendall's Rank correlation test. All results listed are statistically significant.

Features Most Positively Correlated with COVID Case Rate		
Top Features	**Spearman's Rank**	**Kendall's Rank**
Percent of population over 25 with less than a high school education	0.799	0.605
Emergency department visits for asthma per 10,000	0.732	0.538
Percent of population living in poverty	0.726	0.533
Potential risk for lead exposure in children	0.650	0.466
Percent housing-burdened low-income households	0.632	0.451

Table 2: Correlation between features and COVID case rate

Features Most Positively Correlated with COVID Death Rate		
Top Features	**Spearman's Rank**	**Kendall's Rank**
Percentage of limited English-speaking households	0.515	0.357
Percent of population living in poverty	0.512	0.356
Percent of population over 25 with less than a high school education	0.486	0.336
Population Characteristic Score	0.468	0.323
Percentage of population that is Hispanic	0.446	0.306

Table 3: Correlation between features and COVID death rate

The percentage of the adult population with less than a high school education was highly correlated with both COVID case rate and COVID death rate, being the feature with the strongest correlation to case rate, and among the most strongly correlated with death rate. This is indicative of the effectiveness proper education has on the individual's COVID protection measures. The percentage of the population living in poverty also showed a strong correlation with both case rate and death rate, pointing to financial burden as a probable cause of escalated high-risk behavior. Besides poverty, the feature with the highest correlation to the case rate was the number of Emergency department visits for asthma. The feature that showed the strongest correlation to death rate was the percentage of limited English-speaking households, hinting at the lack of proper communication channels as an important detriment to receiving the necessary COVID treatment. Overall we found that there were both pollution and socio-economic features that showed a strong correlation to case rate. For the death rate, it was predominantly socio-economic features that showed the strongest correlation. After finding the features that were most positively correlated with the case and death rate, we looked at which were most negatively correlated.

Features Most Negatively Correlated with Case Rate		
Top Features	**Spearman's Rank**	**Kendall's Rank**
Percentage of population that is White	-0.724	-0.531
Percentage of population that is Other/Multiple	-0.663	-0.477
Percentage of population over 64	-0.541	-0.379
Pollutants in impaired water bodies	-0.203	-0.136
Leaking underground storage tank sites	-0.150	-0.100

Table 4: Inverse correlation between features and COVID case rate

Features Most Negatively Correlated with Death Rate		
Top Features	**Spearman's Rank**	**Kendall's Rank**
Percentage of population that is Other/Multiple	-0.450	-0.309
Percentage of population that is White	-0.406	-0.277
Percentage of population over 64	-0.208	-0.140
Leaking underground storage tank sites	-0.176	-0.118
Pollutants in impaired water bodies	-0.161	-0.108

Table 5: Inverse correlation between features and COVID death rate

We found that the same set of features were highly inversely correlated to both COVID case and COVID death rate. The three most inversely correlated features for both case and death rates were the percentage of the population that is white, the percentage that is Other/multiple, and the percentage over the age of 64. The fourth and fifth most inversely correlated features for both case and death rate were the amount of leaking underground storage tank sites and the amount of pollutants in impaired water bodies. We see that amongst the most negatively correlated features, both pollution and socio-economic features were included. Overall, for both positive and negative correlations, there was a stronger relationship between our features and case rate than with death rate.

As seen above, the ordering of the correlation coefficient between different features is the same for Spearman's Rank and Kendall's Rank. This suggests that the Spearman's Rank correlation coefficient is likely to be more accurate as the additional assumptions made by Spearman's Rank test did not alter the results from Kendall's Rank test.

These results show that education and financial stability are among the most important factors in identifying vulnerable communities, while heavily suggesting the necessity to commit more resources to facilitate communication with foreign language speaking households to improve their COVID treatment.

4.3 Predictive Model

In addition to correlation testing, we performed predictive modeling. We developed several machine learning models to make predictions on the various health issues included in our data set. The models used included Ridge Regression, Lasso Regression, XGBoost, and Random Forest Regression. We trained the models on three variations of our data set:

1. Raw data
2. Normalized data
3. Percentile data

We trained each model to predict each of our health issues. After tuning, we determined which model was the most accurate for each health issue. In order to evaluate the accuracy of our models, we used K-fold cross-validation. The results are listed in the following chart.

Most Accurate Predictive Model		
Feature	**Most accurate model**	**Type of data used**
Emergency department visits for asthma per 10,000	Lasso Regression or Ridge Regression	Percentile Data
Percent low birth weight	Random Forest Regression	Percentile Data
Emergency department visits for heart attacks per 10,000	XGBoost or Random Forest	Percentile Data
COVID Case Rate: Number of cases per 100,000 residents	Lasso Regression	Normalized Data
COVID Death Rate: Number of cases per 100,000 residents	Random Forest Regression	Raw Data

Table 6: Results of predictive models on health issues

For asthma rate, heart attack rate, and percent low birth weight, the most accurate predictions were made with percentile data. For COVID case rate and death rate, the most accurate predictions were made with normalized and raw data respectively. Some models performed similarly, such as with Lasso Regression and Ridge Regression on predicting asthma rate. The models were the least accurate when predicting COVID death rate.

5 Conclusion with Implications and Suggestions for Future Work

Based on our findings, we were able to identify vulnerable communities as those with high chemical emissions, lead exposure, and drinking water contamination. These vulnerable communities suffered from high rates of COVID, COVID deaths, low birth weight, emergency department visits for heart attacks, and asthma. Our correlation tests found that education and financial stability are very important factors in identifying vulnerable communities. We recommend committing more resources to facilitate communication with foreign language speaking households to improve their COVID treatment. We also recommend that further research should be done to identify the causes of the Hispanic population's vulnerabilities to COVID.

In the future, we would focus on expanding our prediction models. We would like to predict how changes in pollution affect these health issues. We would also like to predict how will health in vulnerable communities change with reduced pollution levels. We would finally like to incorporate additional health condition datasets to see if we can expand our prediction models.

6 Acknowledgements

A special thanks to PIC Math which is a program of the Mathematical Association of America (MAA) and the Society for Industrial and Applied Mathematics (SIAM). Support is provided by the National Science Foundation (NSF grant DMS-1722275). In addition, the authors thank Dr. Jie Zhong, Professor at CSULA for his guidance that made this project possible. We also want to thank, Eva Pereira, Chief Data Officer of the city of Los Angeles, who served as a liaison and provided valuable feedback throughout the collaboration.

References

[1] Dimitris Leventis. XGBoost Mathematics Explained. 2018.

[2] Kendall Tau Metric. Encyclopedia of Mathematics. 2010.

[3] Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*, pages 69–71, 2012.

[4] Stacey Ronaghan. The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. 2018.

[5] Sidharth Sekhar. Math behind Linear, Ridge and Lasso Regression. 2019.