Identifying Features Correlating to COVID-19 Decease Cases Using Machine Learning Regression Techniques

Jace Mixon
University of Central Florida

***Abstract*-The COVID-19 virus outbreak was declared as a global health emergency since the beginning of the year 2020 and has persisted in being a threat towards human health for over 12 months. Measures have been taken in analyzing what causes the spread of the virus and how to prevent further Infection rates, but limited study was done on the decease rates for the virus. In this research, a regression model Is constructed to analyze potential characteristics of a geographical location, primarily the states within the United States, to observe potential contributors towards decease cases. The results of the model creation and experimentation has lead to the conclusion that common hypothesis towards decease case rates, such as race and ethnicity backgrounds, do not have a significant contributing background towards decease rates, but further research on socio-economical backgrounds and other characteristics of a society can still be further explored to find and construct a more accurate model.**

I. INTRODUCTION

The first reported case of COVID-19 was reported back on December 31st, 2019 and would not be considered a global health emergency from the world health organization until January 30th, 2020. In response to the global health emergency declaration, multiple countries locked down their countries and mandated a stay-at-home order to contain the spread of the virus. This response caused global economic instability throughout the year and impacted social behaviors, such as relying on non-contact interactions and communications using live-streaming applications for work or educational reasons. As of writing this report, vaccines have been produced and distributed in several countries, but it is still deemed 'unsafe" to reopen countries fully.

Throughout the 2020 year, there has been a large effort in understanding the causes, symptoms, and treatments to the virus and were researched in an expedited amount of time to quickly re-establish normal workflow. Because of the expedited research conducted in 2020, there exist a vast amount of data repositories containing country and state statistics and their relations to population infection rates as well as details on the virus itself. However, due to the abundant information detailing infection rates and causations for those rates, creating more research on the same topic only creates redundant information and leads to no further exploration of the COVID-19 impacts on the global population. To create a novel discussion on the virus's impacts, this paper attempts to address the decease rates rather than the infection rates to explore any new insights on the virus's impacts.

II. PROBLEM STATEMENT

Since there is a larger emphasis on the analysis of decease cases, two considerations are to me made: the geographical scope of the analysis and the specific problem statement to analyze. The geographical scope and specificity on the problem statement will impact the tradeoff between the implications of the findings between a broad-base implication and a closer-to-individual level of implication.

A. Data Collection

As mentioned from the introduction, there exists a large amount of data collection pertaining to COVID-19 statistics juxtapose to characteristics of individuals and geographical locations with the virus. These data sources include the Johns Hopkins University: COVID-19 Data Repository, Milken Institute: COVID-19 Treatment and Vaccine Tracker, The World Bank: Global Health Statistics, and several more. Some of the data sources are static surveys taken at specific timestamps to gauge the virus's impacts and other data sources are dynamic repositories that are updated either daily or weekly to ensure consistent results.

The main issue with these data sources is that they partially contain overlapping results while also containing unique results not explicitly recorded in any other data source. This means that it is necessary to collect all the data sources available pertaining to the problem statement, even if most of the data sources contain duplicate results and perform a thorough data cleaning procedure to ensure each data point is unique. Additionally, since some of the sources are updated daily, it would imply that gathering data at a certain timestamp would then be considered 'outdated' after 24-hours. To overcome these challenges, the C3.ai organization compiled a data lake that gathered all COVID-19 related data sources and organized it into a database that is updated periodically, to the same schedule that the data sources are also updated. This completes the data cleaning portion of the data collection task and allows the study to progress further by constructing the data set by querying to the data lake at any given time.

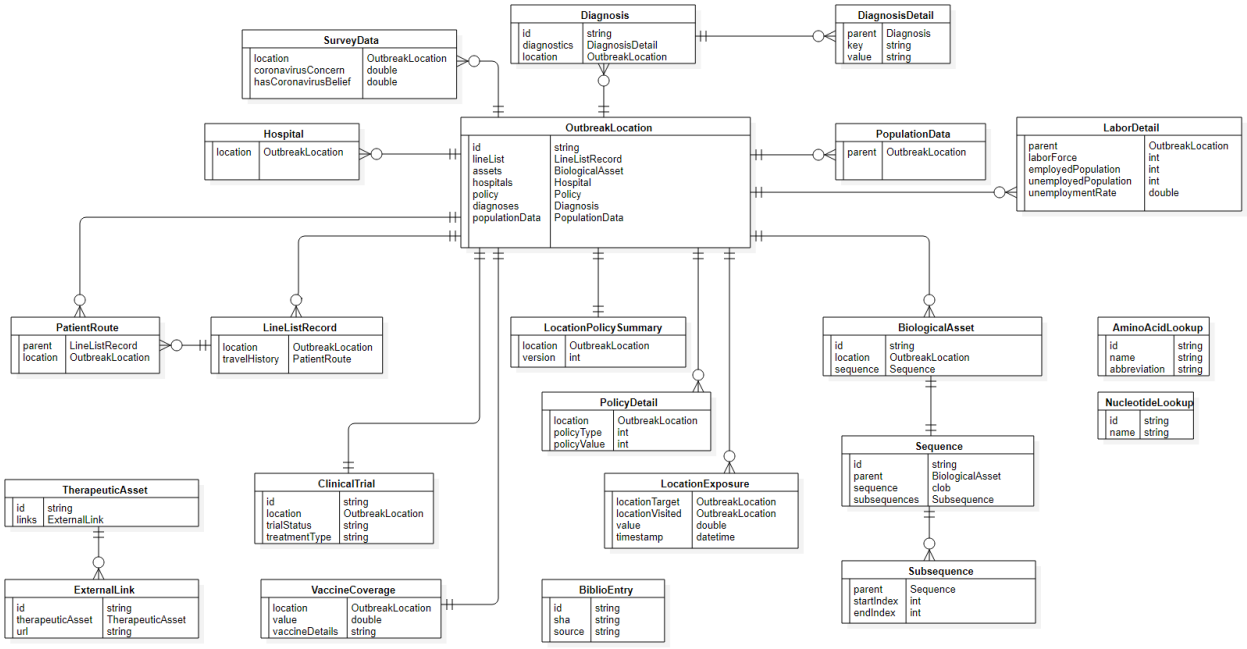

Figure 1. C3.ai COVID-19 Data Lake ERD, detailing the structure of the data warehouse.

B. Specific Problem Considerations

One of the problems that was considered is the correlations between a patient's characteristics and their survival rates. This problem statement, however, could not be resolved within the data lake as there were a limited sample size for decease cases concerning individual patients due to the patients being only laboratory-confirmed cases.

Another problem statement that was considered were the potential time-elapsed until a country could be completely vaccinated, as to create an updated discussion on the current progress towards global vaccination for the virus. Unfortunately, this problem statement was not possible to address within the data lake either since the results were restricted to locations within the United States and vaccination rates were based on previous vaccinations instead of the current vaccine for COVID-19.

In analyzing the structure of the data points, an observation was made that most of the data points in the data lake pertain to geographical location characteristics in relations to the COVID-19 infection rate. These characteristics include age range, race diversity, gender ratio, and population count. These variations of statistics pertaining to different geographical locations, along with values of total infected and decease individuals in each location, became the foundation for the leading problem statement to be addressed in this research: how do different location characteristics contribute to the number of decease cases?

C. Geographical Scope

Along with considering the problem statement for the research, the geographical scope of the research was also questioned. Initially, the research was conducted by analyzing the characteristics of the countries around the globe. However, more statistical characteristics about the United States, primarily describing the labor force from the US Bureau of Labor Statistics, were more plentiful than describing the characteristics of a given country. Additionally, the counties in each state offered more samples to collect for the data set, but it created a concern that the research would be too focused on a smaller geographical scale than normal. Therefore, the scale of the research is on a state-level in the United States.

III. Overview of Proposed Approach

To create the data set that the selected machine learning approach will adapt to, a query of all 50 states is called on the C3.ai data lake, which includes features such as number of hospital beds, population of state, population of children, and state name. From there, a second query is made to find characteristics of the states, such as number of males and females, age ranges, number of white citizens, number of African American citizens, number of Asian citizens, number of Hispanic citizens, and number of citizens with a disability. Lastly, the number of decease cases from April 4th, 2020 to today created the third query and the results per state were divided by the number of days to create a daily decease case. The state characteristics were selected due to their correlation analysis to the decease case scoring ~70% or greater. The state characteristics was supplied from the 2019 US Census and the decease cases were generated from the John Hopkins University repository.

After creating the data set that contains each state in the United States and their respective daily decease case count, a proposed list of machine learning regression algorithms is tested with initial parameters set to predict the total number of decease cases per state and the algorithm that holds the best results are pursued further for improved results. Once the model is fine-tuned to the data set, a series of model predictions are analyzed based on the average decease rate to discover insights towards the characteristics of the states in relations to the overall decease rate.

IV. Technical Details of Proposed Approach

For the preliminaries of the data set that will be used for testing and training, data integration has been incorporated at the start by using the C3.ai data lake, as mentioned in the problem statement section of this report. Additionally, the data set integrates data that is used from the 2019 US Census estimations, US Bureau of Labor Statistics results, and John Hopkins University COVID-19 Tracker repository. Since all 50 states are taken into consideration, there are no noise points or missing values to be considered. Lastly, the C3.ai data lake handles duplicate values, as mentioned from the problem statement section, by only having one value used for values that overlap with multiple sources. For feature selection, all attributes about a state have been tested using a correlation analysis with the decease case rate and have scored more than 70% on average. For feature engineering, the hospital beds were the only features that were combined to create one feature to reduce redundancy in the data set and still maintain high correlation to the target values.

The list of proposed machine learning regression model was as follows: linear regression, naïve bayes modeling, decision tree regressor, random forest regressor, support vector regressor, gradient boosting regressor, and AdaBoost regressor. For regression models that require categorical values, a separate data set has been preprocessed that modifies the feature vector's values to be within the ranges [0, 3] indicating different thresholds determined by the mean and standard deviation of each feature in the data set, but it does create issues to the overall reliance on the model to be an accurate regression model.

A. Linear Regression and Naïve Bayes

A linear regression model assumes a linear relationship between the feature vector and the target values in the following equation:

$$y = m_1x_1 + m_2x_2 + m_3x_3 + \ldots + m_nx_n$$

where $[x_1, x_2, x_3, \ldots, x_n]$ is the feature vector and $[m_1, m_2, m_3, \ldots, m_n]$ are the weights associated for each feature in the data set. While the calculation for each weight is simple by using a residual sum of square error loss method, the assumption of a linear relationship between the feature vector and target values can be detrimental to finding a good fit to the data set.

Naïve Bayes is another simple modeling algorithm that utilizes Bayes theorem to find the best fit:

$$P(y \mid x) = ((P(x \mid y) * P(y)) / P(x)$$

Alternatively, the value for x and y can be swapped and the equation becomes the probability of a feature vector given a target value, which then becomes a generative model and becomes ideal for the problem statement. However, such computation is expensive to conduct and is therefore assumed to be on a gaussian distribution, which then leads to further issues on the assumption of the data set in relations to the target values. model to be an accurate regression model.

B. Decision Trees and Random Forest

A decision tree can be used to create a non-linear discriminative model that also produces an easy-to-observe diagram on the different attributes of the data set and their contributions to the target values. While ideal to have an understandable depiction of which features contribute most to decease cases, Decision Trees are solely dependent on the samples in the data set and are sensitive to changes to the data set overall.

For a better-performing decision tree algorithm, a bagging method known as Random Forest can be used for a non-linear predictive model. What makes the algorithm more robust to the data set than regular Decision Trees are that it randomly samples from the data set and trains multiple, smaller trees rather than having one decision tree, which then prevents chances of overfitting to the data set and overall obtains better performance than one decision tree.

C. Support Vector Regressor

Another discriminative model that retains a potential non-linear boundary construction is the Support Vector Regressor, which is sometimes abbreviated to SVR. Having a soft-margin SVR makes the discriminative model robust to noise points and have a good fit on data sets that have either linear or non-linear relationships between the feature vector and the target values. The issue that SVRs have is that for non-linear data sets, a kernel function is required to remap the feature vector, which becomes a trial-and-error approach to find the best kernel function for the data set and each attempt at the kernel function is computationally expensive.

D. Gradient Boosting Regressor and AdaBoost Regressor

Relating closely to the Random Forest algorithm is the Gradient Boosting Regressor model and AdaBoost Regressor model. The AdaBoost Regressor model follows similarly to the Random Forest algorithm in that it randomly samples from the data set and creates several decision tree estimators that, in combination, creates the boundaries for the data set and predicts the outcome. The difference with AdaBoost is that for every sample that it gets incorrect, the algorithm gives further weight to those samples so that they can be trained for again in the next iteration of training. Additionally, the Decision Trees that are created are decision "stumps" in that they do not go pass a max depth of 1.

Gradient Boosting Regressor follows a similar structure to AdaBoost Regressor with the multiple Decision Trees as estimators to the boundaries and re-weights the samples based on if the model gets it incorrect or not. What makes Gradient Boosting different is that the depth of each tree can be increased and is not limited to being a decision "stump", making it more comparable to Random Forest than AdaBoost.

V. EXPERIMENTS

The total number of features used in the initial data set was 14. This included: population of all children, labor force population, latest kindergarten population, hospital beds, male, female, under 18, 18 and over, 62 and over, white, African American, Asian, Hispanic, and Disability. The last column of the data set contained the decease rate for each state in the data set. A copy of the data set is created to transform the continuous data set into categorical data for the purpose of the naïve bayes model to function.

hospitalBeds	male	female	under_18	18_over	62_over	white	af_am	asian	hispanic	disability	DeathCases
35126	2359355.0	2516895.0	1096376.0	3779874.0	989899.0	3320247.0	1299048.0	66270.0	208626.0	781503.0	3950.058824
3517	384915.0	352153.0	184394.0	552674.0	107739.0	476015.0	24205.0	45920.0	51870.0	86874.0	114.350384
35150	3504509.0	3545790.0	1635344.0	5414955.0	1454601.0	5444453.0	317462.0	233213.0	2208663.0	903268.0	7029.214834
23440	1471760.0	1527610.0	704268.0	2295102.0	607439.0	2301044.0	459542.0	45504.0	224130.0	510910.0	2229.153453
170733	19526298.0	19757199.0	9022146.0	30261351.0	6769198.0	23453222.0	2274108.0	5692423.0	15327688.0	4101034.0	21881.030691

latestKindergartenPopulation	hospitalBeds	male	female	under_18	18_over	62_over	white	af_am	asian	hispanic	disability	DeathCases
2.0	2	2.0	2.0	2.0	2.0	2.0	1.0	2.0	1.0	1.0	2.0	0.0
0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
2.0	2	2.0	2.0	2.0	2.0	3.0	2.0	1.0	2.0	3.0	2.0	1.0
1.0	1	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	0.0
3.0	3	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	1.0

Figure 2. Snapshots of the continuous data set (top) and the categorical data set (bottom).

For each of the proposed methods, a 5-fold cross validation was used, in which the score metric that was used is the residual sum of square error. However, in the graphs provided, the y-axis indicates the model's accuracy rather than the residual SSE, so the graph with a higher score towards 1 indicates better performance.

A. Linear Regression and Naïve Bayes

While the performance for both linear regression and naïve bayes have average scores, the number of parameter-tuning that is available is limited and thus the performance is capped. Additionally, the scores for naïve bayes are misleading since all values calculated are categorical, which is not within the objective of this research.

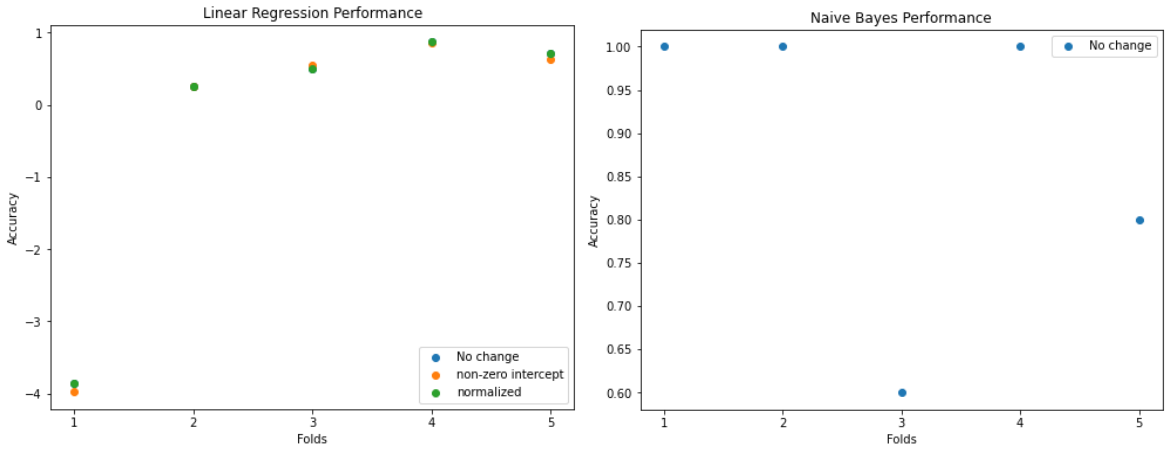

Figure 3. Comparison between Linear Regression performance and Naïve Bayes performance.

B. Decision Trees and Random Forest

With more parameters to adjust, the Decision Tree and Random Forest models were modified by adjusting the number of estimators, the depth of the tree, and the different evaluation metrics to observe the best performing combination of initial parameters. From the results, the Decision Tree model scores below a 50% accuracy while Random Forest models tended towards a consistent, higher accuracy with more estimators and depth to the trees.

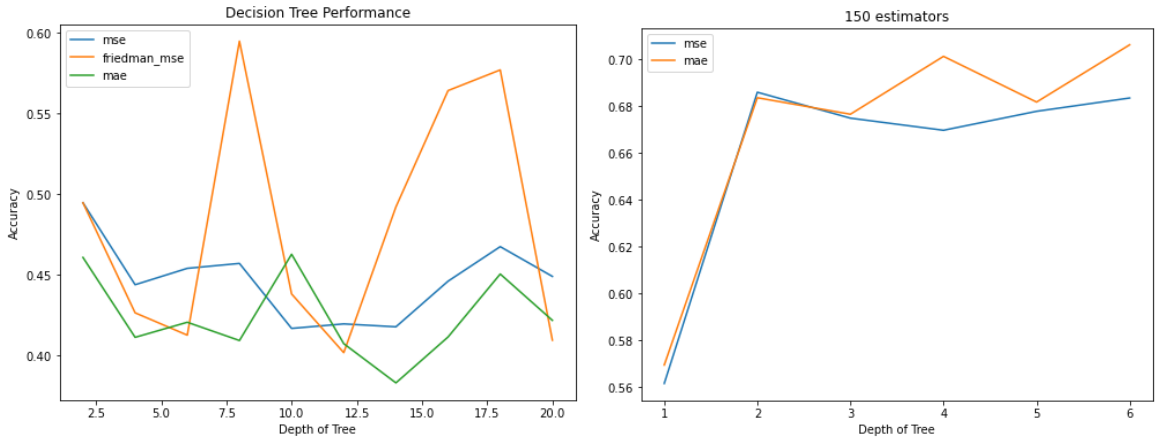

Figure 4. Comparison between Decision Tree performance and Random Forest performance.

C. Support Vector Regressor

The main parameters that were adjusted when testing the performance of the Support Vector Regressor were the different kernel functions to remap the feature vectors. Additionally, for the polynomial kernel function, the different degree sizes were modified to test for any performance improvements. However, all performance results on the 5-fold cross validation tests show either 0% accuracy or less, making this model to be less than ideal for the current research.

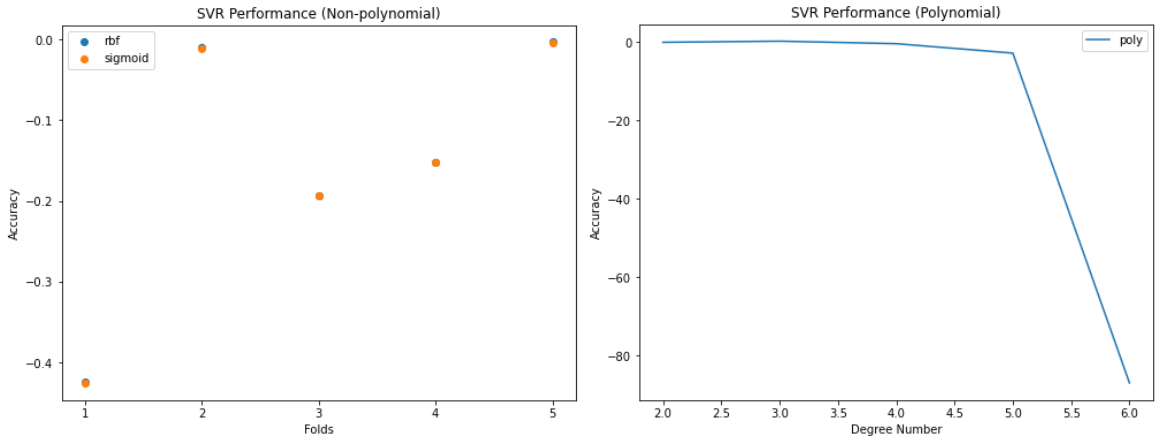

Figure 5. Comparison between the different kernel functions for the Support Vector Regressor model.

D. Gradient Boosting Regressor and AdaBoost Regressor

The parameters that were modified for the Gradient Boosting Regressor model and AdaBoost Regressor model are similar to the parameters that were changed for the Decision Tree and Random Forest model algorithms. The results for AdaBoost Regressor when modifying the number of estimators result in obscure performance depictions, whereas the Gradient Boosting Regressor performance has a clear peak in the number of estimators and the tree depth for best performance.

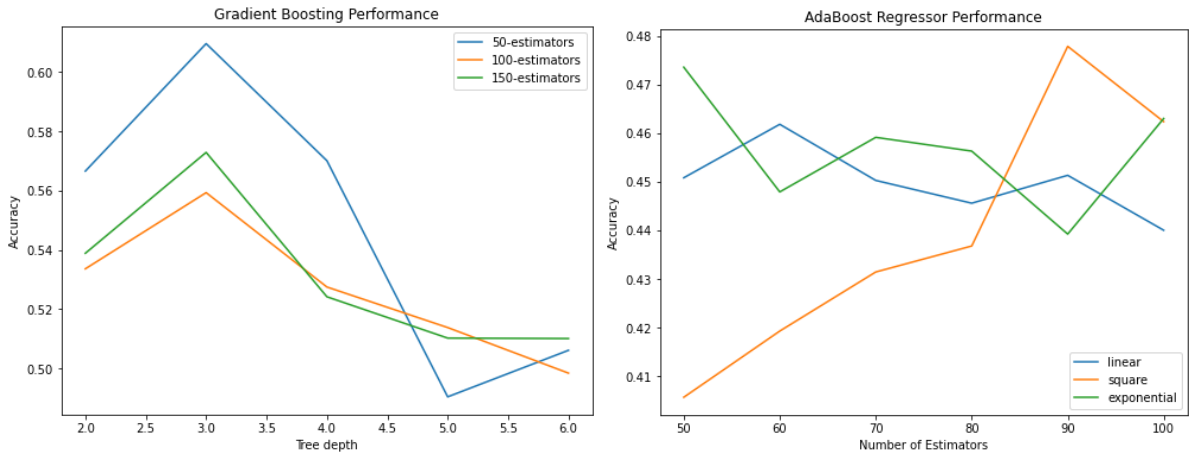

Figure 6. Comparison between Gradient Boosting performance and AdaBoost performance.

E. Performance Improvements

After analyzing the performance of the proposed regression model methods at the baseline level, the Random Forest and Gradient Boosting Regressor models were deemed to have the best potential for high-performing results for the problem statement. Afterwards, two methods for improving performance for the models were implemented. The first approach was to reduce the data set feature vectors to reduce the calculations that the models must consider. This can be accomplished by the model's built-in feature importance per feature in the data set, which outputs the percentage of importance the calculations had in determining the output of the model by a percentage amount that sums to 1. An observation that was made was that most of the features listed were considered less than 1% important to the calculations of the trees in the model. The only features that were deemed important were the hospital beds and the Hispanic population count. The second approach was to use a GridSearch function that can test a range of value combinations and record which combination performs the best for the regression model. For Random Forest, it was observed that numerous, deep trees were favorable, but to retain shorter estimation trees, the size of the trees were maxed at heights of 7 with a total of 250 estimators. Additionally, Gradient Boosting Regressor favored decision tree "stumps" over trees with a larger height, and it could accomplish its task with as few as 50 estimators instead of the baseline 100 estimators.

Combining both approaches increased the performance of the Random Forest model and Gradient Boosting regression model from 64% to 72% and 68% to 75%, respectively. The difference between the two models is that in the 5-fold cross validation analysis, the Random Forest model had high-performing models in one of the folds while having two other folds performing less than the average line. In contrast, the Gradient Boosting model performed near the average performance line, showing more consistent performance in the 5-fold cross validation.

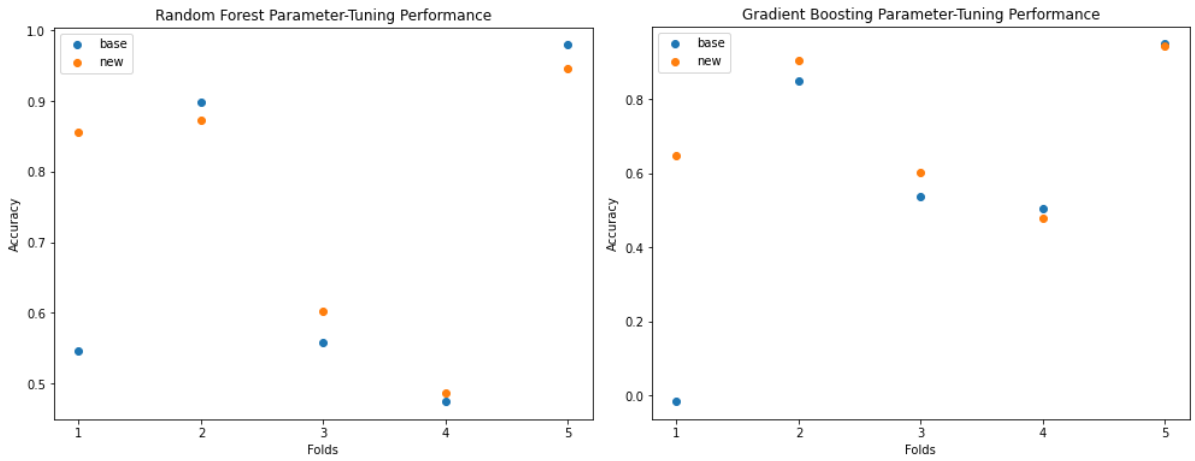

Figure 6. Comparison between Random Forest parameter-tuning performance and Gradient Boosting parameter-tuning performance.

We can thus conclude that the Gradient Boosting regression model performed the most consistently with a high performance-rating of 72% on the data set.

After fine-tuning the models and selecting the best-performing model, 7 different and unique points were generated and administered to the models to analyze the results from the model. Each point represents a combination of the varying sizes of hospital beds and Hispanic population. The output is then compared to the average decease case: if it falls below the average decease case, then the values associated to the generated data point produces low decease cases (indicated as a 0 in the output). Otherwise, the values associated to the data point produces high decease cases (indicated as a 1). The results can be viewed from Table 1. In identifying the decease case count with respects to the changes in the feature values, it was noted that changes in the hospital bed count was the leading cause for the decease number to increase, while the total count for the Hispanic population barely contributed enough decease cases to surpass the average decease case threshold, which was 5,619 per day.

TABLE I
SAMPLE RESULTS FROM GRADIENT BOOSTING REGRESSOR

Samples	Description	Decease Count Prediction	Output
1	25% of Hispanic population average	2694.77	0
2	75% of Hispanic population average	3172.73	0
3	25% of Hospital Bed average	1706.63	0
4	75% of Hospital Bed average	7436.08	1
5	25% of both in combination	1582.20	0
6	50% of both in combination	2819.19	0
7	75% of both in combination	7789.62	1

VI. RELATED WORKS

The potential to use machine learning algorithms to analyze the impacts of the COVID-19 virus is an idea that has been thoroughly explored and numerous publications in the attempts to understand and predict future outcomes of the virus based on current behaviors have been written. Earlier articles have been written since March 2020, when

the virus outbreak has been deemed a global health emergency and countries began implementing stay-at-home orders, and immediately research has been conducted using regular regression analysis to observe a quadratic trend of infection rates while having the issue of a limited amount of data due to the virus being recently discovered [1]. With this report, over 12 months have passed and by observing the report once again, we can clarify that the infection rates followed similarly to an exponential rate, which is like what was analyzed initially from the report.

As the virus persisted throughout the year 2020, more literature on using machine learning to better understand potential steps towards overcoming this obstacle were presented. As analyzed by Punn, Sonbhadra, and Agarwal, estimations based on infections and decease rates were shown to agree with previous literature that a vast number of lives were at potential to be lost should measures to counteract the virus are not chosen and upheld [2].

It would not be until after several months have passed that there existed several COVID-19 sources that data scientists can use to find further insights towards the spread of COVID-19. Surveys [3] have been conducted that reviews potential options that can further mitigate the spread of the COVID-19 virus and explore potential in using machine learning algorithms to help in diagnosing patients with the virus. Additionally, studies are conducted to find other leads towards the spread of the virus, in which it was found that cultural proximity and cleanliness influences the speed of infection rates while cultural habits or social distancing and social spaces did not have noticeable influence [4]. All the samples that were read from the existing literature would soon become the foundation of the data samples and features that would be used for the exploration of this research's problem statement. Additionally, the regression analysis would become key in further understanding the virus's impacts after a year has passed since the global health emergency declaration has been announced.

VII. Conclusion

The research that has been conducted was an attempt to address the possible characteristics that can lead to further decease cases within a geographical location. The characteristics that were analyzed were primarily on the different characteristics of the population within the location, which included age range, gender ratio, race and ethnic diversity, and disabilities. These feature characteristics were selected since they were initially hypothesized to have a contributing factor towards decease cases based on current understandings of the virus and its relations to infection and decease rates in highly populated areas along with correlation analysis during data preprocessing. However, upon creating and tuning a regression model that aimed towards predicting the decease rates within the states, it was found that these features did not contribute significantly towards decease cases, which contradicts the hypothesis of the research. We can then say that the hypothesis was incorrect, that neither race, gender, ethnic background, or current disability has a significant contributing factor towards decease cases in the states. Further research can be done with socio-economical backgrounds and broadening the scope of the data set to include countries instead of analyzing states within the United States.

References

[1] S. Deb, M. Majumdar, "A time series method to analyze incidence pattern and estimate reproduction number of COVID-19." *Applications*, 24 Mar. 2020.

[2] N. S. Punn, S. K. Sonbhadra, S. Agarwal, "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms", medRxiv 08 Apr. 2020.

[3] D. S. W. Ting, L. Carin, V. Dzau, T. Y. Wong, "Digital technology and COVID-19", *Nature Medicine*, vol. 26, pp 458-464, Apr. 2020.

[4] H. M. Singer, "Short-term predictions of country-specific Covid-19 infection rates based on power law scaling exponents.", *Physics and Society*, 25 Mar. 2020.

Source code: https://github.com/Jace-Mix/ML-Assignments/blob/main/Project/CAP%205610%20Project.ipynb