COVID-19 Case Prediction in Virginia Based On Demographic Factors

William Zhang
Charlottesville, VA, USA
wyz3sp@virginia.edu

Rithik Yelisetty
Charlottesville, VA, USA
ry9bf@virginia.edu

Saiteja Bevara
Charlottesville, VA, USA
sb2xf@virginia.edu

ABSTRACT

COVID-19 has changed life as we know it - people are required to wear masks and social distance themselves to slow the spread of the virus. Since the COVID-19 pandemic began, people have been trying to accurately forecast how the novel coronavirus will spread and infect additional individuals. News organizations, along with individuals interested in analyzing trends, have all made various predictions since March 2020 as to how many people will become infected with this disease.

Various claims have been made regarding how different demographics have affected the trends regarding the spread of COVID-19. Prominent predictions that are noted in common media outlets have been based primarily on previous case counts. The goal of this project is to make predictions for the spread of COVID-19 with the inclusion of demographic data and income in order to better predict the next hotspots in the state of Virginia. The demographic data that will be considered includes age and race.

Author Keywords

covid-19, prediction, machine learning, demographics, data

BACKGROUND

COVID-19 has brought the world to a halt, infecting more than 68.5 million individuals across the globe and killing over 1.5 million. Just in the United States, more than 15 million people have been infected and over 286,000 people are now dead. Along with this, the virus does not seem to be slowing down, infecting close to two hundred thousand individuals on a daily basis, just in the United States. COVID-19 has also brought world economies to a halt and has caused borders to be closed for travel.

In order to complete the project of predicting Virginia case counts based on demographic data, several datasets will be used in order to increase the accuracy of the predictions. The first dataset that will be used is from the Virginia state government [4]. This dataset consists of the daily counts of cases, hospitalizations and deaths per locality in Virginia. Localities in Virginia will be determined using the FIPS code as this is a standardized method that assigns a code to each county and city.

The second dataset that will be used is from the United States Department of Agriculture [3]. This dataset consists of various socioeconomic data for each locality (by FIPS code). The data that the team plans to focus on include the unemployment rate and median household income.

The third and final dataset that will be used is from the Centers for Disease Control and Prevention [1]. This dataset contains the demographics of each locality within Virginia. The demographic data that the team intends to use includes the population density along with the distributions of age, race, and sex.

The combination of these three datasets will allow for analyze of various data to determine if there is correlation between these factors and the case counts of COVID-19 in the state of Virginia.

RELATED WORK

A study by Sarmadi, Marufi, and Moghaddam [6] examined various demographic factors over a course of three months in different countries. In particular, they examined the correlation of COVID-19 data with average GDP, temperature, and latitude/longitude. Two of the metrics established in the study were proportion of cases to population per 10^5 (PCP), as well as the proportion of deaths to population per 10^5 (PDP).

The results of the study showed that countries with higher gross domestic product (GDP) tended to have higher COVID-19 PCP and PDP. However, the authors note that the reason for this association could be a result of better availability of diagnostic kits and health care facilities to help identify cases, as well as the tendency of higher GDP countries to be more densely populated.

Additionally, the average temperature was found to be inversely associated with COVID-19 PCP and PDP. The authors draw on previous studies that show higher temperatures could be effective against different types of coronavirus such as SARS, as at lower temperatures, the performance of the immune system and liver are weakened. Since more northern latitudes tend to have colder temperatures, this association with higher COVID-19 PCP and PDP was also observed by the authors.

Other previous work done in the area also includes an Exploratory Data Analysis (EDA) available as a notebook on

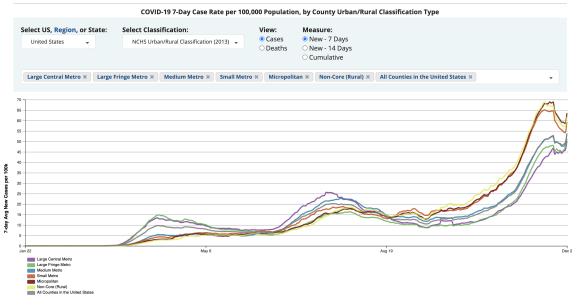

Figure 1. This figure shows the rising case counts in locations with different population densities across the United States. As evident by this graph, even though case counts are rising everywhere, cases per 100,000 individuals are rising in "non-core" or rural areas, along with large metro regions, the fastest across all population densities [2].

Kaggle [5]. The EDA looked at the demographic predictors broken down by country. As part of the analysis, the United States, China, and India were left out, as a scatter plot of GDP vs Population showed the three countries as outliers. With the rest of the data, mutliple different analyses were used, including an ordinary least-squares (OLS) regression, random forest, and single decision tree. In the final analysis, it was found that the GDP, mean age, and smoking were the best predictors of COVID-19 infections.

Another Kaggle notebook by William Shamma [7] was focused on finding the overall number of hospitalizations due to COVID-19 in the United States as a whole. The analysis showed that age was the factor that contributed the most to the rate of hospitalizations due to COVID-19.

CLAIM/TARGET TASK

The target task is to predict future case counts of COVID-19 based on information such as past case counts along with demographic information. This includes socioeconomic data such as unemployment rate and median household income, and other distributions of age and race. These will consist the features of the data, and the output will be predicted number of new cases expected for the next day. Whereas most models that consider some form of demographic data are at a national scale, this project focuses on localities within Virginia.

PROPOSED SOLUTION

Dataset Creation

As a part of the proposed solution, the first step is to organize and create the dataset for the model to be trained on. There are three separate datasets which serve as sources for the overall dataset. For each of the datasets, information will be matched based on FIPS, a unique identifier for localities. The focus will be on localities within Virginia. The first dataset provides total cases, hospitalizations, and deaths for each FIPS code for a given day [4]. Exporting this data as a CSV provides all the data starting from March 17th, 2020 up to the current date. Training samples for each locality will be created, identifying the case counts in sets of 10, starting from 3/17/20, which comprise the first 9 parameters and the output.

For each of these samples, there will be constant demographic and other information obtained from two other datasets for each locality. There will be constant parameters representing median household income and unemployment rate, for a total of 2 parameters [3]. The last dataset will provide distributions of age and race. Age distribution will be represented as percentages for ranges in (0-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65+) age groups. Population race distribution will be added as percentages for the categories of White, Asian, Black, and American Indian. Thus, each sample will have a total of 22 parameters. Each sample will create a row in a CSV file specific to a locality and set of 10 days [1].

Model

Given that the dataset is tabular and is essentially a regression problem with time series data but includes constant demographic data, an MLP would be an appropriate model to use. The number of hidden layers, hidden dimensions, optimizer, activation function, number of epochs, and learning rate will be the basic hyper-parameters for the model. Another hyper-parameter will be the number of previous days which are used in training samples for which to obtain case counts, with the baseline being 10 days.

Inputs and Outputs

In order for our model to accurately predict the number of cases, the team will be inputting a total of 22 variables and will be expecting 1 outputs. The team will be inputting 9 variables purely about COVID-19: one variable for each of the number of cases for each of the previous 9 days. The team will be inputting race and age inputs. This will compromise of a total of 11 variables - one for each race groups, age groups. The race groups are broken down into White, Asian, Black and American Indian. The age groups are broken down into the following classifications: 0-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-85. The final two input parameters will be the median household income and the unemployment rate from 2019. The output parameter will be the predicted number of cases for the next following day.

CONTRIBUTIONS

There are three main contributions that this project provides to analysis of COVID-19. By combining Virginia demographic data and COVID-19 case data, the project provided a better picture of the correlation between demographic factors and case counts. Furthermore, the project provided a better indication of exactly which demographic categories impacted case count predictions in Virginia. Finally, we provided predictions for case counts at a granular county level based on both demographic factors and past case counts.

IMPLEMENTATION

Technical Solutions

In order to design a Machine Learning model to predict COVID-19 cases in different localities across Virginia, the team started out by shuffling all data from the dataset. 90% of this data was used as the training set and the remaining 10% was used as the testing set. The team implemented a MLP-based Sequential model using the Keras library, which itself is

built on top of Tensorflow. The structure of the model involved having an input layer of size hidden dimension with the "relu" activation function. This is then connected to several hidden layers, all with the same activation function and the same hidden dimension as the initial input layer. This is then fed into a dropout layer that is used to remove extraneous data from the set. This also helps with generalization of the model to ensure that the model is not overfit to the training data. The output layer returns one number (the case count for the next day) and uses the "linear" activation function since this is similar to a regression problem where the model is expected to predict one thing. Two optimizers were used during the hyperparameter training portion: SGD and Adam. The loss function that was chosen was Mean Squared Logarithmic Error.

In order to determine the best set of hyperparameters, the team used the Stratified K Fold algorithm to divide the training data into compartments. All compartments (except for 1) were used for training and the remaining compartment was used for cross validation. This process was repeated across all of the folds and the average of the accuracies was taken. The set of hyperparameters that resulted in the highest average accuracy was chosen. After significant hyperparameter tuning, the team determined that the following set of parameters resulted in the best model:

```
 { 'batch_size':  8, 'epoch':  10,
'learning_rate':  0.000001, 'hidden_dim':
200, 'hidden_layer':  20, 'opt':  'adam',
'x_train_shape':  x_train.shape }
```

Technical Challenges

Due to the nature of this problem, there were many technical challenges that the team encountered. Since demographic data by itself cannot be used directly to predict the number of cases, the team noticed that the accuracy of the model hit a limit of about 65%. The team also faced challenges implementing the Stratified K Fold algorithm due to the fact that the team had difficulty trying the split the data into an appropriate number of classes. The last technical challenge the team faced was determining a way to accurately measure what constituted as an accurate prediction from the model.

Technical Novelty

For this project, the team implemented a new accuracy function in order to help determine what is an accurate prediction. The true values are subtracted from the predicted values. If the difference between the true values and the predicted values is less than 2.5 cases, the model marks the prediction as accurate. If the difference is greater than 2.5 cases, the prediction is inaccurate. This differs from the usual accuracy function which compares the two values exactly. The second part of the project that involved significant work was the dataset creation. The dataset creation involved combining data from three different data sources and properly identifying relevant data needed for each row.

Method Variations

In addition to the main project of the trying to determine the exact number of cases for each locality, the team also tried working on improving the accuracy of the model by trying to do a binary classification problem: predicting whether the number of cases will be higher or lower than the previous day. This was accomplished using both an MLP model and a SVM model. Both models performed relatively similarly with accuracies around 70%. The team also tried a base model of having no hidden layers. This resulted in a significantly worse accuracy of 21%.

DATA SUMMARY

In order to successfully solve the problem determining COVID-19 cases in localities in Virginia, the team started by gathering data from three distinct sources: Virginia Department of Health (VDH), US Department of Agriculture (USDA) and the Centers for Disease Control and Prevention (CDC). All datasets contained real and up-to-date information about the data being portrayed.

Virginia Department of Health Data

The VDH data consisted about information regarding the number of cases per day for each locality. The team found each unique day and FIPS code combination, along with the associated number of cases for that locality on that day, and inserted data into a SQL table. Once this was complete, the team then used a separate SQL table (along with a Python script) to find the number of cases for the past 9 days and store this information alongside the current day and the FIPS code.

US Department of Agriculture Data

The USDA data consisted of the median household income and the unemployment percent for each locality. The unemployment percent and median household income, along with the associated FIPS code, was inserted into a separate SQL table.

Centers for Disease Control and Prevention Data

The CDC data showed various demographic data like race and age breakdowns per locality. For each district, the team first computed the percent of individuals that fell into each age and race bucket. This data, along with the total number of individuals in each district and the FIPS code, was added to a separate SQL table.

Combining Data

All of the data is combined on FIPS district using the SQL NATURAL JOIN command. This data is then read by the pandas library in Python and outputted to a csv format to be ingested by the ML model. Given that our dataset has over 30000 unique rows with data about each day of the pandemic (since March 17, 2020) for each locality, the team believes that we had sufficient data in regards to the parameters chosen to be explored. Since the data chosen has sufficient information about the demographics, the team hopes to be able to predict the number of cases in each locality for any given day assuming that the team has access to both the demographic data and the previous days of COVID-19 cases.

EXPERIMENTAL RESULTS

Metrics

The final model had a validation accuracy of about 64-65% and a test accuracy of about 64-65% as well. As the two accuracies are effectively the same, it implies that the model fit the training and test data well, and did not overfit or underfit.

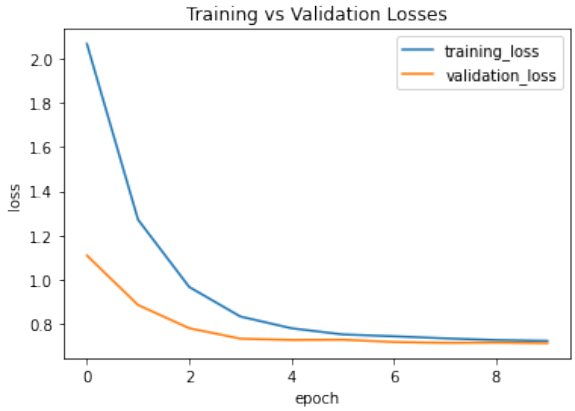

Figure 2. This figure shows the training and validation losses for the final model parameters. It can be seen that the training loss decreases at a good rate and is not too shallow, implying the learning rate is appropriate.

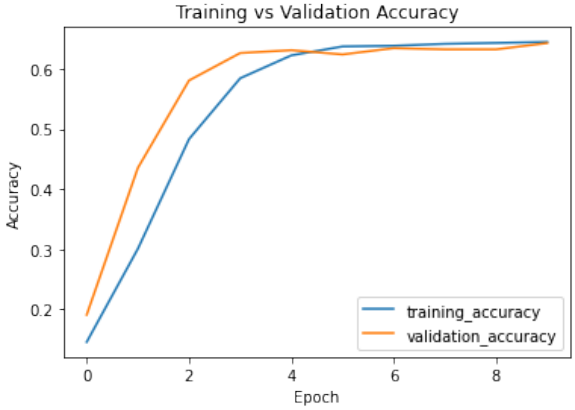

Figure 3. This figure shows the training and validation accuracy for the final model parameters. It can be seen that the training and validation losses converge to about the same, implying the model is a good fit for the data.

Baselines

As there were no publicly available studies done on the impact of demographics on daily COVID-19 case counts in Virginia at the time of writing, two appropriate baseline were determined to be the accuracy of random guessing, based on the range of daily COVID-19 cases in the dataset, as well as the accuracy of a simple linear regression model predicting daily COVID-19 case counts without any demographic data.

A rough worst-case estimate of the baseline accuracy from purely guessing based on the dataset was determined by finding the locality with the largest range of daily new COVID-19 cases. In the dataset, this was found to be Fairfax County, which had a maximum of 496 daily new COVID-19 cases. Since we can assume the minimum daily new COVID-19 case count can be 0, the probability of correctly random guessing the next day's case count is approximately 0.2%.

The linear regression model used for the baseline took into account the same previous nine days of COVID-19 case counts and excluded all demographic data, producing a testing accuracy of 30.45%. The random guessing baseline and the results of the simple linear regression model imply that the final MLP model that takes the demographic data into account provides about twice as accurate predictions than a baseline linear regression model without any demographic data.

Hyperparameters

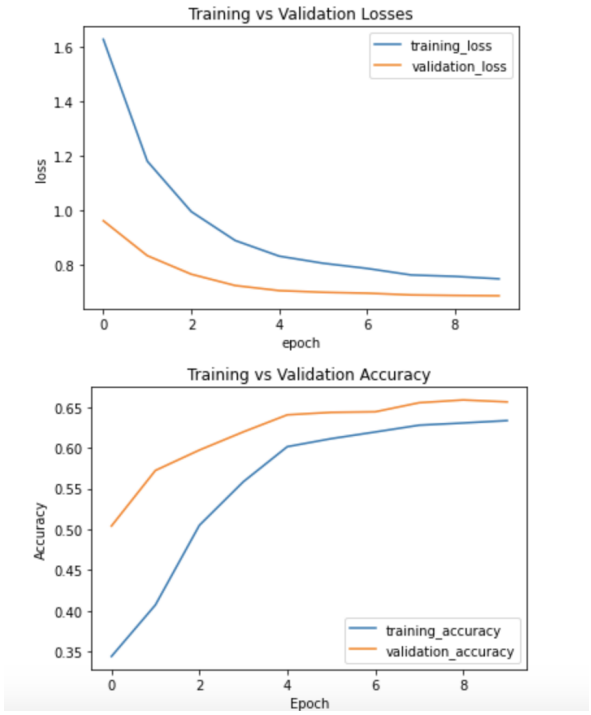

Figure 4. This figure shows the training results for a learning rate of 0.0000005 and 10 epochs, which causes the model to slightly underfit.

The hyperparameters that were tested and optimized for the model were the learning rate, the optimizer function, the number of epochs to train, the number of Keras Dense hidden layers, and the dimension of the hidden layers. A large range of learning rates was tried and Figure 4 shows a training result from a learning rate that was too low. The losses and accuracy for training and validation failed to converge fully and is slightly underfitting.

Conversely, a learning rate too high, as shown in Figure 5, causes the validation accuracy to be very jumpy. The particular training result caused the accuracies to fail to converge, and overfit the data. The final learning rate of 0.000001 was chosen due to it allowing for the model accuracy and losses to converge while also having the losses decrease at a rate that was not too shallow. Similarly, the number of epochs (10) was

Figure 5. This figure shows the training results for a learning rate of 0.000005 and 4 epochs, which causes the model accuracy to jump around and fail to converge.

chosen to be long enough that both the model accuracy and losses converge, but not too long such that the model begins to overfit or the accuracy no longer changes significantly.

The two optimizer functions that were tried were the ADAM and SGD functions. It was found that the ADAM optimizer function led to higher test accuracy, and was chosen as a result. Additionally, the hidden dimension was varied between the values in the set $\{10, 20, 50, 100, 200, 400\}$ and the number of hidden layers was varied between the values in the set $\{1, 2, 3, 5, 10, 15, 20, 30\}$. The resulting values of 200 and 20, respectively, were chosen as they resulted in the highest test accuracy.

EXPERIMENTAL ANALYSIS

With the given dataset, the best accuracy after hyperparameter tuning was about 65%. Even modifying the task to be a binary classification task, attempting to predict whether the number of cases would simply be higher or lower than the previous day, resulted in a similar accuracy. Thus, the dataset could be expanded to include more datapoints in order to better improve accuracy. Another option would be to incorporate other demographic factors besides the ones included in this study.

In order to gain insight into which category or value for each demographic metric influenced predictions, test data was manually created which highlighted these differences. For example, taking into consideration the age distribution, there are seven different values, ranging from ages 0-14 to ages 65-85.

Rank	Age Group
1	65-85
2	25-34
2	55-64
4	35-44
5	45-54
6	15-24
7	0-14

Rank	Race
1	Black
2	American-Indian
3	White
4	Asian

Figure 6. Ranking of age and race distributions based on predicted cases

Keeping all other data such as past case counts, distributions for race, median household income, and unemployment percentage constant, the values for these domains was modified: all age groups excluding one were given a value of 0, while the leftover was given a value of 1.0 which represented the entire population consisting of this age group. This was repeated for each age group, and finally this test data was used to generate predictions. These predictions revealed whether having a larger portion of the population belonging to a single age group would lead to a higher number of predicted cases. This process was also repeated in a similar manner for the race distribution. For the unemployment percentage, the value was varied between 0.0 and 10.0. For the median household income, the baseline income value from which the test sample was generated was multiplied and divided by a factor of 5 to create the upper and lower range.

The results displayed that a lower unemployment rate and lower median household income resulted in a higher number of predicted cases from the model. Similarly, a larger portion of the population belonging to the age group of 65-85 or a larger portion of the population being black also led to a higher number of predicted cases. For age and race, the results are summarized in the figure 6, with a higher rank indicated a higher number of predicted cases. However, it is important to note this analysis was simply to determine which categories or values were correlated with a higher number of cases. There is no strict determination of causality in this scenario. In that respect it is similar to a Granger causality with the demographic factors presenting data, albeit not time series data as is expected by the Granger causality test, which provides a different prediction for the number of cases. This prediction we project is more accurate holistically with the other demographic factors than predictions based solely on previous case counts.

As this project provides a model to study demographic data and its correlation and use in predicting COVID-19 cases, it provides a novel application which studies this information at a locality-level within Virginia. However, it can be applied to other scenarios, such as at a higher level studying states or countries rather than localities.

CONCLUSION AND FUTURE WORK

Overall, the model performance was relatively good. Considering the largest range of case counts, the values ranged from 0 cases to 496 cases in Fairfax County. Therefore, using a baseline of random guessing for the number of cases, we can

expect an accuracy of around 0.2%. Furthermore, the accuracy of the baseline model without demographic data was about 30.45%. The model thus outperforms both these scenarios.

The project was also able to determine for each of the demographic factors taken into account, which of the categories or values correlated with a higher number of predicted cases. This provided insight into the relationship between demographic data and the correlation to COVID-19 cases.

It should be noted that demographic data does not imply causation and can only reasonably be correlated with daily COVID-19 case counts. As a result, it makes sense that the model does not have extremely high accuracy. In order to achieve more accurate predictions, the model would need to take many more external factors into account that could have a much larger impact on predictions. Further work could, for example, consider a larger number of external factors such as the population density, smoking rate, and mask-wearing compliance in a locality. Any of these factors could have a much larger impact than demographics on the final result. Overall, a larger number of factors and amount of data could be used to further improve the accuracy of this model.

ACKNOWLEDGEMENTS

Thank you to Professor Yanjun Qi and teaching assistants Zhe Wang, Jack Lanchantin and Arshdeep Sekhon for their help and support on this project.

REFERENCES

[1] Centers for Disease Control and Prevention. 2019. U.S. Census Populations With Bridged Race Categories. (2019). (https://www.cdc.gov/nchs/nvss/bridged_race.htm#Newest%20Data%20Release).

[2] Centers for Disease Control and Prevention. 2020. CDC COVID Data Tracker. (2020). (https://covid.cdc.gov/covid-data-tracker/#pop-factors_7daynewcases).

[3] United States Department of Agriculture Economic Research Service. 2020. Unemployment Rate 2019. (2020). https://data.ers.usda.gov/reports.aspx?ID=17828.

[4] Virginia Open Data Portal. 2020. VDH-COVID-19-PublicUseDataset-Cases. (2020). https://data.virginia.gov/Government/VDH-COVID-19-PublicUseDataset-Cases/bre9-aqqr/data.

[5] Night Ranger. 2020. COVID19-Country-EDA. (2020). (https://www.kaggle.com/nightranger77/covid19-country-eda).

[6] Kazemi Moghaddam V. Sarmadi M, Marufi N. 2020. Association of COVID-19 global distribution and environmental and demographic factors: An updated three-month study. *Environ Res.* (2020). DOI: http://dx.doi.org/10.1016/j.envres.2020.109748

[7] William Shamma. 2020. Covid-19 Demographic Susceptibility. (2020). (https://www.kaggle.com/willshamma/covid-19-demographic-susceptibility).