Insightful Interfaces

Saiteja Bevara, Rithik Yelisetty and William Zhang

October 11, 2020

Motivation

- People have made various claims 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
- We want our predictions for the spread of COVID-19 to include demographic data and income in order to better predict the next hotspots in the state of Virginia

Background

- COVID-19 has brought the world to a halt
  - Infected over 68,500,000 individuals as of December 9
  - Killed over 1,500,000 people as of December 9
  - Closed borders for travel
  - Shut down economies
- Datasets
  - https://data.virginia.gov/Government/VDH-COVID-19-PublicUseDataset-Cases/bre9-aqqr/data
    - Cases, hospitalizations, and deaths (per locality)
  - https://data.ers.usda.gov/reports.aspx?ID=17828
    - Socioeconomic data: Unemployment rate and median household income (per locality)
  - https://www.cdc.gov/nchs/nvss/bridged_race.htm#Newest%20Data%20Release
    - Demographic data including distributions of age and race

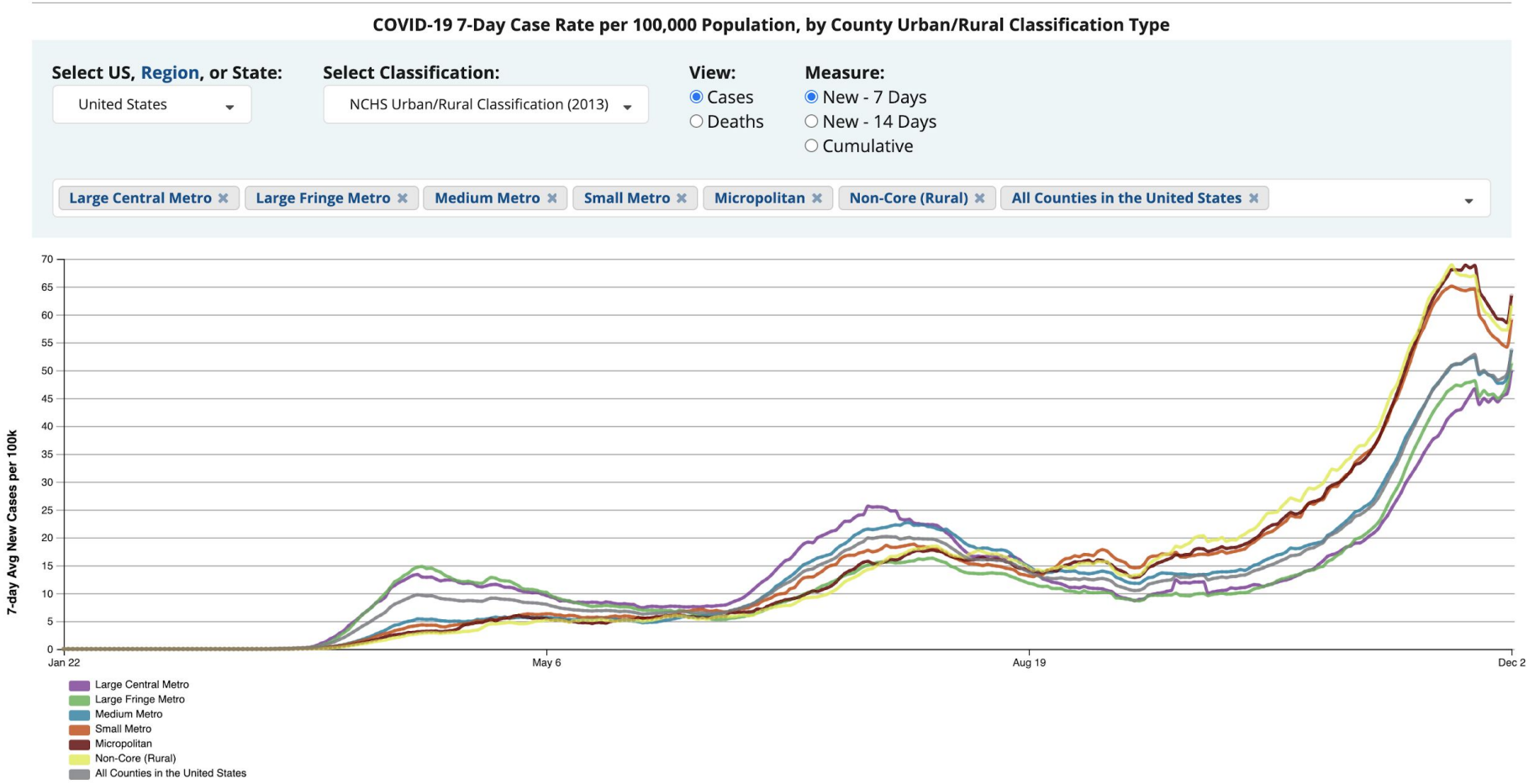
Related Work

- Many related projects have been completed on trying to associate COVID-19 case counts and demographic data
- https://www.kaggle.com/nightranger77/covid19-country-eda
  - Using metrics like 2018 GDP, Crime and Population, Smoking rate, Percent Female, Median Age and COVID-19 specific data, project found that countries with higher GDP, age population and smoking levels had more infections
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7258807
  - Considered environmental and economic factors as distribution indicators of COVID-19 using bivariate correlation and regression tests
  - Concluded that higher GDP was correlated with more COVID-19 and temperature had a reverse association
- https://www.kaggle.com/willshamma/covid-19-demographic-susceptibility
  - This notebook looked at how population age directly influenced the cases, hospitalizations, ICU admissions and deaths for COVID-19
  - The study found that the people admitted to the ICU across all age groups was roughly equal

Claim / Target Task

- Predict future cases of COVID-19 based on past confirmed case count as well as the population and demographics of the area in Virginia.
- The factors / model features that will be considered will be case count, along with demographic data such as race distribution, age group distribution, median income and unemployment rate
- The output or label assigned will be predicted number of new cases expected for the next day

An Intuitive Figure Showing WHY needed



Proposed Solution

- Combine demographic data sets for race, age group, unemployment rate, median household income by VA locality
- Match the demographic data by FIPS code to VA COVID-19 cases
- Implement an MLP model to predict the number of COVID-19 cases the next day
- Train the MLP model on the combined and matched dataset
- Tune hyperparameters to find the best model architecture and parameters for training

Implementation

- All data from training set is shuffled and 90% is used for training and 10% is used for testing
  - For validation testing, the team used the Stratified K-Fold method with n_classes = 10.
- MLP-based Sequential Model created using Keras
- 20 hidden layers between input and output layer with a hidden dimension of 200 nodes and “relu” activation function
- Dropout layer between hidden layers and output to remove all extraneous data
- Output layer returns 1 number (case count) with the “linear” activation function
- Adam optimizer and Mean Squared Logarithmic Error Loss Function is used for the model
- Via hyperparameter training, the model uses a learning rate of 0.000001 and is run for 10 epochs

Contributions

By combining VA demographic data and COVID-19 case count data, we:

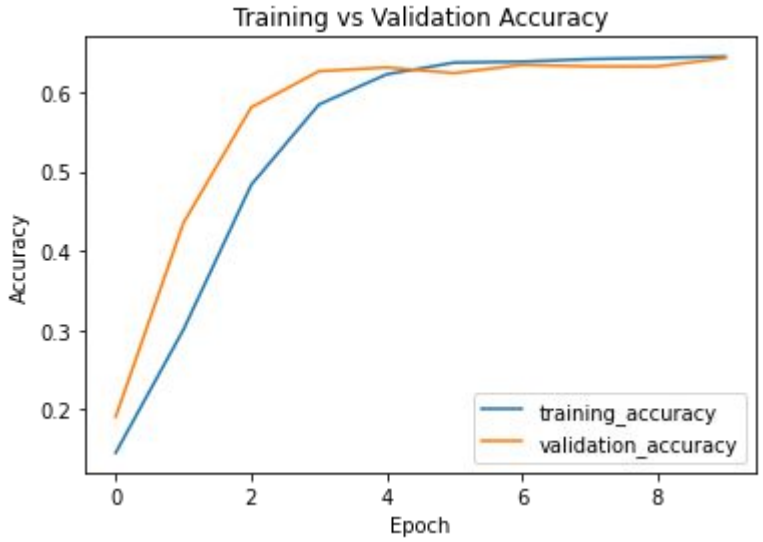
- Provided a better picture of the correlation between demographic factors and case counts
- Provided a better indication of which demographic categories impacted case counts in VA
- Provided predictions for case counts at a county level based on demographic factors and past case counts

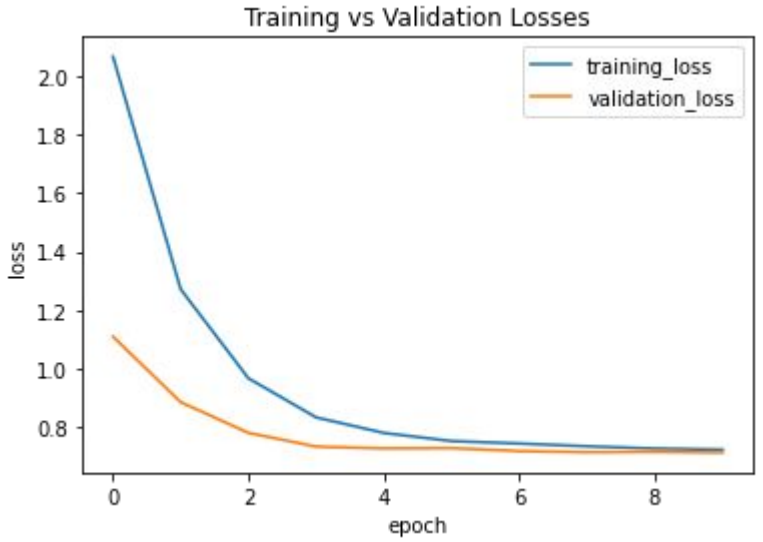
Data Summary

- Three distinct datasets were used to gather information about demographics (age, race, unemployment percent and median household income) and COVID-19 case counts
- For the age and race data, each group is measured as a percent of the total population. Once percentages are computed, all data is added to a SQL table.
- For the unemployment percent and median household, US census data is used to insert all relevant data into a separate SQL table.
- COVID-19 case counts for the past 10 days are also added to a SQL Table for each FIPS district along with the current day.
- All data is combined on FIPS district using the SQL NATURAL JOIN command. This data is then read by the pandas library and outputted to a csv format to be ingested by the ML model.

Experimental Results

- Best Performing MLP Model
  - Batch_size: 8, epoch: 10, learning_rate: 0.000001, hidden_dim: 200, hidden_layer: 20, optimizer: adam
- Validation Accuracy: ~65%
- Test Accuracy: ~65%





Experimental Analysis

- With the given dataset, the best accuracy for all models and parameters was around ~0.65
  - More data for each county or other demographic factors could increase accuracy
- Certain input features were more indicative of a higher number of predicted cases
  - Lower unemployment rate and lower median household income

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

Conclusion and Future Work

- Model performance was relatively good
  - Largest range of daily new COVID-19 case counts in Fairfax County
  - Range from 0 to ~500
  - Baseline accuracy of < ~0.2% (random guessing)
  - 65% accuracy relatively high in comparison
- Determined for each of the demographic factors taken into account, which age group and race in a locality had the most impact on the next day's case count
- Further work could consider a larger number of demographic factors and take more factors into account (i.e. smoking, population density)

References

- https://data.virginia.gov/Government/VDH-COVID-19-PublicUseDataset-Cases/bre9-aqqr/data
- https://data.ers.usda.gov/reports.aspx?ID=17828
- https://www.cdc.gov/nchs/nvss/bridged_race.htm#Newest%20Data%20Release
- https://www.kaggle.com/nightranger77/covid19-country-eda
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7258807
- https://www.kaggle.com/willshamma/covid-19-demographic-susceptibility
- https://covid.cdc.gov/covid-data-tracker/#pop-factors_7daynewcases