University of Miami

Final Project Report:

Identifying COVID-19 Hotspots in Florida

Sushmita Persaud

Justin Prince

CSC 642: Statistical Learning with Applications

1. Introduction

a. Background of Disease

In 2019, a new disease emerged in society, causing distress and isolation in the world as leaders tried to identify a strategy to keep the virus contained. The first case of Coronavirus 2019 (COVID-19) emerged on December 29th, 2019 in Wuhan, China where it became a significant transport hub spreading the virus to other areas of the world [5]. The Coronavirus 2019, popularly known as COVID-19 is caused by a virus called severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) and is an airborne virus that is highly contagious to spread [6]. Several cases of COVID-19 were reported daily which resulted in the World Health Organization (WHO) announcing the cases as a pandemic on March 11th, 2020. [1].

It is spread through droplets which are very small particles that contain the virus [2]. When a person with the virus exhales the droplets, surrounding individuals can pick up the virus through inhaling the infected air [2]. The virus continues to be highly investigated as there is no specific treatment and researchers do not know the long-term effect of COVID-19 on the human body [6]. The virus can lead to severe chest symptoms and has reported deaths among all age ranges. However, there is a trend that the older population has a high mortality rate of the virus, especially when coupled with comorbidity diseases, thus indicating there is a susceptible population to the virus [6].

b. Prevention of COVID-19 Spread

Due to the rapid outbreak of the virus, world leaders had to impose strict traveling restrictions and curfews to prevent further spread of the virus [1]. Recommendations from WHO include to “interrupt human to human transmission” and implement early detection of the virus through rapid testing [6]. Decreased mobility and testing were highly encouraged to reduce the

spread of the virus [1]. In addition, increased hygiene practices were implemented such as heightening the importance of washing hands and wearing masks [6].

c. Research Question

As there are continued emergence of various strains of COVID-19, it is important to identify the risk factors that contribute to the spread of the virus to avoid another lockdown. The research project will focus on finding hotspots of COVID-19 in Florida. **The main question is on what are the hotspot counties that there is an increased risk of contracting the virus in Florida.** Through analyzing the main question, we will examine external factors, such as which genders are more susceptible to getting the virus and if population density plays a role in spreading COVID-19. We will investigate the following questions along with the main research question:

1. Can the models accurately predict whether a county in Florida is a hotspot?
2. Which attributes are important in distinguishing if a county should be considered a significant transport hub of the virus?
3. How well do the models perform in terms of accuracy and reliability?

d. Dataset

Observing external factors will help us filter the dataset to obtain optimal results for the main question posed. The field of data that will be utilized for this project will belong to different components from various backgrounds. It will include health components as the question is focused on a virus, COVID-19. The data needs to be collected from residents in Florida who have tested for the virus and the number of deaths from the virus. In addition, the dataset will include population density factors, such as the size of the county and the number of residents within that area. The dataset will be divided into counties. This is relevant as it allows the viewer

to understand how external factors can contribute to contracting the deadly virus. Also, the data will include income and employment status. This will play a role in determining if affluent neighbors play a heavy influence in reducing deaths and contractions of COVID-19. To answer this research question, we will explore various types of variables that contribute to making an area a hotspot for contracting the disease. The problem can be identified as a classification problem as the answer consists of dividing the dataset into various classes to help distinguish which county is the most prone in contracting and spreading COVID-19.

2. State of Art

a. Databases

A research paper provides an overview of the various datasets and breaks down the components that are analyzed to make decisions regarding COVID-19. It hones in on open resources data and analyzes the variables that are considered for identifying reliable sources of datasets for data-driven models [1]. It was concluded the important variables were correlated to the susceptibility of an individual of contracting the virus [1]. The variables were divided into two groups: demographics and health system [1]. Demographics included “age, gender, prevalence of diabetes, high blood pressure, obesity, and other risk factors” while the health system included “availability of artificial respiration equipment, ICU, specialized medical surveillance and treatments, etc. [1].

b. Previous Methods

The datasets related to COVID-19 have been used in various ways. It has been utilized for predictions of infected individuals, economic impact, analysis of seasonal behaviors, impacts on healthcare system, and potential forecast on resource availability [1]. Map visualization is a

popular tool used to display the data collected by the datasets to show centers of major disbursement of the disease [5].

In an article published by NIH, they used R-programming to visualize the hotspots in the United States (Rimal Y, 2021). The research utilizes multiple libraries such as "*(leaflet), (tidy verse), (ggmap), (htmltools), (leaflet. extras), (maps), (ggplot2), (mapproj), (mapdata), (spData)*" [5]. They were able to filter through the dataset and identify clusters throughout the states and identify the hotspots through red dots on the map to separate the clusters. Another study used map visualization using the following libraries from R programming: "janitor", "ggplot2", "lubridate", "dplyr" and "tidyverse" [7]. They examined the relationship between confirmed cases of COVID-19 and death cases in India using regression models [7]. The correlation between death cases and confirmed COVID-19 cases in India was 0.91, indicating the value is statistically significant [7]. The researchers used the model to assess the prediction of the number of COVID-19-positive cases on morbidity which resulted in the training model to performed at 89% accuracy [7]. It was concluded that this application can be used in other regions and states.

3. Materials and Methods

a. Data Collection and Integration

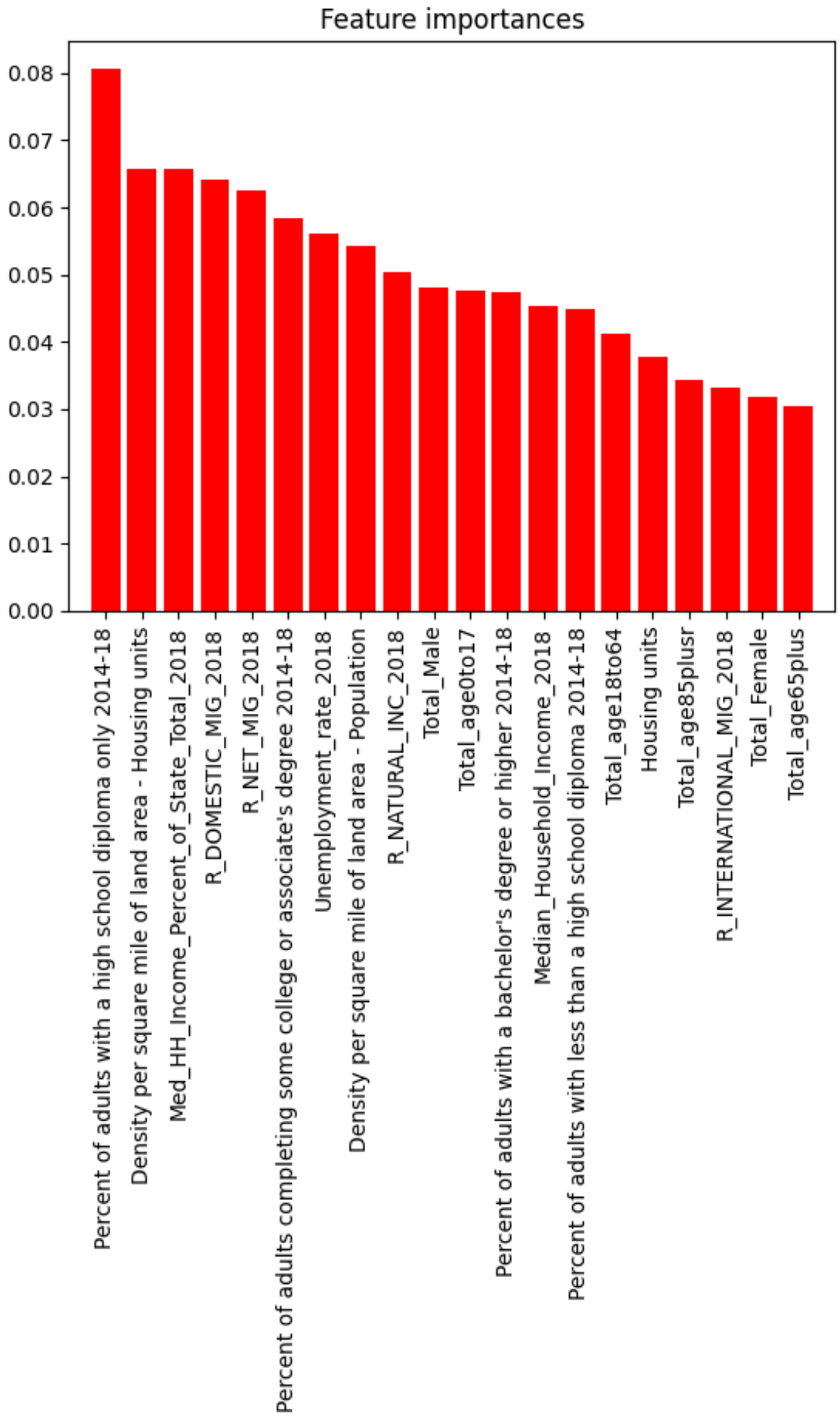
In this work, we integrated two datasets. The datasets that were selected contain demographic, socioeconomic, health care, education and transit data for each county in the 50 states and Washington DC. The first dataset is from Florida Covid Cases dataset which was sourced from ArcGIS Hub and archived by the University of South Florida Libraries [4]. It contained information on COVID-19 instances and testing metrics for every Florida county.

The second dataset was a machine-readable collection of socioeconomic factors which affect the spread of outbreaks, particularly COVID-19 [3]. The information is sourced from the Covid-19 Data Repository by the Center for Systems Science (CSSE) and Engineering at Johns Hopkins University (JHU) [3]. There are 348 features available for the 3143 county records however we will only focus on a subset of those records which pertain to the Florida counties [4]. The dataset does unfortunately contain some missing values though they are typically for the number of households in a county which is not a feature we aim to analyze. Therefore, the data for the number of households can be omitted.

The variables we are considering are the population density of each county, median household income, age profiles of each county (how many people there are in different age ranges), number of COVID-19 tests and the number of COVID-19 related deaths. Additionally, the data is numerical in nature for ease of analysis.

As the main goal of our research is to identify which Florida counties are pandemic hotspots, classification methods will be used to highlight high-risk areas. First, a target feature for high risk areas will be created by finding the number of positive cases in each Florida county per capita. Counties in the 80th percentile or above for cases per capita will be classed as high risk and this feature will be appended to the aforementioned Florida Covid Cases dataset. Then, the features from the JHU dataset will be analyzed to see which of them have the highest relevance in correctly predicting the risk status of a county.

Table 1. Variables Used in the Analysis


Feature importances
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Percent of adults with a high school diploma only 2014-18
Density per square mile of land area - Housing units
Med_HH_Income_Percent_of_State_Total_2018
R_DOMESTIC_MIG_2018
R_NET_MIG_2018
Percent of adults completing some college or associate's degree 2014-18
Unemployment_rate_2018
Density per square mile of land area - Population
R_NATURAL_INC_2018
Total_Male
Total_age0to17
Percent of adults with a bachelor's degree or higher 2014-18
Median_Household_Income_2018
Percent of adults with less than a high school diploma 2014-18
Total_age18to64
Housing units
Total_age85plusr
R_INTERNATIONAL_MIG_2018
Total_Female
Total_age65plus

After sorting the data and determining which variables had the highest correlation, we ran models using the following categorical variables:

- Percent of adults with a high school diploma only 2014-18
- Density per square mile of land area- Housing units
- Median household Income as a Percent of State Total 2018
- Rate of domestic migration 2018
- Rate of net migration 2018
- Percent of adults completing some college or associate's degree 2014-18
- Unemployment rate 2018
- Density per square mile of land area- Population
- Rate of natural population increase 2018
- Total Male population
- Total age 1 to 17
- Percent of adults with a bachelor's degree or higher 2014-18
- Median Household Income 2018
- Percent of adults with less than a high school diploma 2014-18
- Total age 18 to 64
- Housing units
- Total age 85 plus
- Rate of INTERNATIONAL MIGRATION 2018
- Total Female Population
- Total age 65 plus

b. <u>Methods and Evaluation</u>

The main classification methods we will use include Support Vector Machines (SVM) and Decision Tree classifiers. Leave One Out Cross Validations will also be used to improve model performance by optimizing the parameters. Support Vector Machines (SVM) is mainly used for classification and regression. In this case, it will be used to find a hyperplane in a multi-dimensional space that distinctly classifies the data points. The dataset will be split into two sets: training set and test set. The Decision Tree is used to separate the dataset into tree-like models based on different conditions. This will used to separate the external factors in helping see which ones are highly associated with large cases of COVID-19 cases. Leave-One-Out Cross-Validation will be used to enhance these models and provide more accurate representation in training the dataset to match the actual database.

4. Results

The three chosen models had high yet varying degrees of accuracy when determining which counties had high infection rates. Thus, the models show that the variables are highly correlated with increased cases of COVID-19 in the regions. Apart from population density, the most significant factors appear to be socioeconomic conditions such as education level, median income, and unemployment rate.

After running the models, the following counties were identified as the top four COVID-19 hotspots in Florida: Dade, Broward, Palm Beach, and Hillsborough.

These results suggest that areas wealthier than the states average with unemployment rates and education attainment levels lower than average tend to be Covid hotspots. This matches the profile of counties with high wealth disparity. Interestingly, the number of residents under the age of 18, and the working aged population to a slightly lesser degree, have a far higher correlation with high Covid rates than the number of residents over 65. This makes sense as

Covid would be spread more rapidly amongst people who are part of education institutions or companies that necessitate interaction with others in person.

Using grid search to optimize parameters allowed the SVM to adapt to the data without underfitting or overfitting however, the Logistic Regression Model (LRM) performs worse with cross-validation which suggests that it may not generalize well beyond the original train-test-split

The best prediction model was RandomForest with parameters:
{'max_depth': 3, 'min_samples_leaf': 10, 'min_samples_split': 2, 'n_estimators': 50}

Model	**Tuning**	**Score**
Random Forest	Grid Search	92.9% (93% after GS)
SVM	Grid Search	78.6 % (86% after GS)
Logistic Regression	Cross Validation	85.7% (75% after CV)

5. Conclusions and future work

After running these models, we should see which areas are highly susceptible to getting COVID-19-positive cases. Future applications can be implemented to see if higher incidences of testing lead to higher COVID-19-positive cases and can be used to differentiate the best solution for containing the virus and reducing the spread of COVID-19. In addition, we can use the template and structure of the models to run predictions in other states to help identify hotspots in other regions of the United States. The following questions can be asked to explore other areas of interest similar to the spread and nature of the virus:

- Is this research applicable to other airborne viral infections?

- To what extent does the vector of a disease determine the regions where it spreads most rapidly?

References

[1] Alamo, T., Reina, D. G., Mammarella, M., & Abella, A. (2020). Covid-19: Open-data resources for monitoring, modeling, and forecasting the epidemic. *Electronics*, *9*(5), 827.

[2] Centers for Disease Control and Prevention. (2024, April 9). *About covid-19*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/your-health/about-covid-19.html#:~:text=COVID%2D19%20(coronavirus%20disease%202019,%2C%20the%20flu%2C%20or%20pneumonia.

[3] CSSEGISandData. (n.d.). *CSSEGISANDDATA/covid-19: Novel coronavirus (COVID-19) cases, provided by JHU CSSE*. GitHub. https://github.com/CSSEGISandData/COVID-19

[4] *Florida COVID19 07042020 byzip CSV*. ArcGIS Hub. (n.d.). https://hub.arcgis.com/datasets/016470ec61c34cc8bdf4a32b7ca43aea/about

[5] Rimal Y, Gochhait S, Bisht A. Data interpretation and visualization of COVID-19 cases using R programming. Inform Med Unlocked. 2021;26:100705. doi: 10.1016/j.imu.2021.100705. Epub 2021 Aug 30. PMID: 34485681; PMCID: PMC8404394.

[6]] Tanno, L. K., Casale, T., & Demoly, P. (2020). Coronavirus disease (COVID)-19: world health organization definitions and coding to support the allergy community and health professionals. *The Journal of Allergy and Clinical Immunology: In Practice*, *8*(7), 2144-2148.

[7] Vashisth, R., Tripathi, S., Goel, H., & Srivastava, P. (2022, November). Visualization of Covid-19 Pandemic Data: An Analysis. In *2022 3rd International Conference on Computation, Automation and Knowledge Management (ICCAKM)* (pp. 1-6). IEEE.