**Business Understanding**

Despite being a common occurrence every year, influenza, also known simply as the flu, is one of the leading causes of death in the United States and around the world. By depressing the immune system of the host, and effecting children and elderly in particular, influenza paves the way for more dangerous infections to cause pneumonia and was associated with 55,672 deaths in the U.S. in 2017, making it the 8th leading cause of death that year. Although flu vaccines are an important tool available to reduce the public health impact of influenza, because of the slow process with which these vaccines are manufactured they have to be made months ahead of the flu season. The long lead times force manufacturers to develop the vaccine based on an estimate of the influenza virus strain that will be dominant in the upcoming season, which can result in the vaccine being less effective if the guess is wrong. Because of this high uncertainty, the Center for Disease Control has made it a national priority to track the spread of influenza by maintaining a weekly national database of key metrics related to influenza. The two most important of these metrics are the influenza-like-illnesses, which tracks the number of patients who presented themselves to the healthcare system with symptoms of influenza, and a count of the influenza cases that were confirmed by laboratory tests. The goal of the CDC is to use this data to better predict the spread of the flu in order to make better plans about resource allocation and public engagement in the vaccination campaign with the ultimate aim of reducing the number of infections and deaths. Although these metrics are reported to the hospitals on a weekly basis, there is a two-week lag between the hospitals collecting the data and the aggregate results being reported to them because one week is required to laboratory test and submit the data to the CDC and another week is necessary for the CDC to compile and analyze the data that they’ve received. It is a critical mission for researchers to fill this two-week knowledge gap with an effective forecasting model because the incubation period for the flu is typically one to four days. If the CDC were able to forecast a surge in influenza cases two weeks before it happens, it would have enough time to warn the public and get people vaccinated before they get sick, as well as allow the healthcare system to prepare their emergency rooms for this surge. Previously attempts have been made to improve the forecasting model with weather data, as temperature and precipitation are important to the flu incubation process. In addition, data science tools have been used to improve the model with the inclusion of search and social media data. However, many of these attempts are at the location-insensitive national level or only focus on one county. We develop a model for all NY state counties and attempt to improve the model with a unique recently-released county-level economic dataset based on modern public health research showing the importance of social determinants of health such as income and employment.

**Data Understanding**

The CDC

**Data Preparation**

<https://www.medicalnewstoday.com/articles/282929.php#influenza-and-pneumonia>

<https://www.cdc.gov/flu/weekly/flusight/why-flu-forecasting.htm>