Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines

Problem Statement

Vaccines provide immunization for individuals, and enough immunization in a community can further reduce the spread of diseases through "herd immunity."As of the launch of this competition, vaccines for the COVID-19 virus are still under development and not yet available. The competition will instead revisit the public health response to a different recent major respiratory disease pandemic. Beginning in spring 2009, a pandemic caused by the H1N1 influenza virus, colloquially named "swine flu," swept across the world. Researchers estimate that in the first year, it was responsible for between 151,000 to 575,000 deaths globally. A vaccine for the H1N1 flu virus became publicly available in October 2009. In late 2009 and early 2010, the United States conducted the National 2009 H1N1 Flu Survey. This phone survey asked respondents whether they had received the H1N1 and seasonal flu vaccines, in conjunction with questions about themselves. These additional

questions covered their social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and behaviors towards mitigating transmission. A better understanding of how these characteristics are associated with personal vaccination patterns can provide guidance for future public health efforts.

Research Question

- Predict how likely individuals are to receive their *H1N1 and seasonal flu vaccines*.

Objectives

- To find the best ROC AUC score in the Dataset
- To discover which model used can produce a high prediction score.

Performance Metric

Performance will be evaluated according to the area under the receiver operating characteristic curve (ROC AUC) for each of the two target variables. The mean of these two scores will be the overall score. A higher value indicates stronger performance.

Data Understanding

Data Source

The dataset used for this project was obtained from Drivendata website.

Data Description

Our dataset was in csv format and contained data grouped in columns of where the Predictors were in training_set_features.csv whereas the Target variables were in training_set_labels.csv.

Data Preparation

Loading the data

At the beginning of the process, the necessary libraries were imported and then the dataset was loaded onto the jupyter notebook.

Reading and checking the data

Read the data and familiarize with it.

Cleaning the data

The data was analyzed and cleaning was done by checking the missing values. Inspect the outliers.

Exploratory Data Analysis

Visualizations were part of this process with helped in answering some of the questions asked earlier on. Did compare some features with the target variable which made the data familiar i.e Bivariate and Univariate Analysis.

Modeling

Several models were used in the dataset i.e Logistic Regression, Random Forest, Decision Tree Classifier just to mention a few so that a high prediction score could be achieved.