Prediction of H1N1 and Seasonal Flu Vaccinations

1. Introduction

Influenza (flu) is an illness that causes respiratory problems and can lead to hospitalisation or even death in some cases as a result of its complications. People over the age of 65, young children and people with pre-existing conditions are at a higher risk of flu-related complications. An annual seasonal flu vaccine can help mitigate the effects of the flu by reducing its severity. The vaccine causes antibodies to develop after two weeks which provide protection against the flu. Everyone over 6 months of age is advised to take the vaccine according to the Center for Disease Control and Prevention.

The H1N1 pandemic began in 2009 and the influenza A virus first emerged in the United States and led to the deaths of approximately 151,700-575,400 people worldwide. The flu primarily affected young children and middle aged adults but surprisingly, people over 60 years old were less affected, presumably because they had antibodies against this virus, likely from exposure to an older H1N1 virus earlier in their lives. A vaccine for the flu became available to the public in October 2009 and consequently, the World Health Organisation (WHO) declared an end to the pandemic in August 2010. However, the virus continues to circulate as a seasonal flu virus.

1.1. Problem Statement

Vaccine hesitancy is the unwillingness to receive vaccines due to fears of the vaccine's effects on one's body and is one of the biggest threats to global health. People's views on vaccination are influenced by their personal and social circumstances and they can change over time. Unfortunately, the effects of vaccine hesitancy are more pronounced during pandemics and pose huge risks to the wider community.

It is up to scientists and public health practitioners to identify the reasons behind this and work towards mitigating them. The first step could be using information people's socioeconomic status and demographic backgrounds, beliefs as well as their opinions on vaccine effectiveness and risks in order to identify potential reasons for vaccine hesitancy thus working on public awareness campaigns and creating new public health policies thus reducing the risks involved with vaccine hesitancy for future pandemics.

1.2. Main Objective

The goal of this project is to create a model that can predict seasonal flu vaccine uptake based on a person's background and behavioral patterns.

1.3. Data Understanding

The datasets used for this project were downloaded from Driven Data. The original data source is the National 2009 H1N1 Flu Survey (NHFS) and it contains information on the social,economic and demographic backgrounds of the respondents as well as their opinions on the H1N1 and seasonal flu vaccines. The datasets have been divided into the training set features and the training set labels. The training data has 26707 rows and 36 columns.

2. Data Assessment

The data was loaded into the jupyter notebook and the necessary libraries imported. The data was then assessed so as to better understand it.

2.1. Exploratory Data Analysis

Exploratory data analysis was conducted so as to better understand the distribution of various features in the data as well as their relationships with the target variable. A number of observations were made including:

- People identifying as white dominated the dataset at around 80%.
- Most of the respondents were employed and college graduates.
- People's opinions about the effectiveness of the vaccine and the risks associated with not taking it greatly influenced their decision to take it.

3. Data Preprocessing

The data was preprocessed so as to make it fit for modeling. This was achieved by checking for collinearity between for features, dealing with missing values and encoding categorical variables. Before anything was done, the dataset was split into the training and test portions at a ratio of 70%-30%.

3.1. Checking for Multicollinearity

This was achieved by creating a dataframe displaying the collinearity between features. Since the highest collinearity was around 58%, no columns were dropped as it was decided that none of them were too correlated with each other.

3.2. Dealing with Missing Values

This was done in two ways:

- Replacing missing values.
- Dropping rows with missing data.

With regards to replacing, some columns were replaced with strings communicating that the information was missing while others were replaced with the mode of the columns.
The SimpleImputer class from scikit-learn was used to facilitate this process and it was fitted on the training data then transformed on the test data. This process was done on the columns with the most missing data; around 50% of the dataset.

After this, any remaining columns had the rows with missing data dropped as they were not that many and would not affect the model performance. This was also done on the test data.

3.3.Encoding Categorical Variables

A couple of columns were encoded with random strings, presumably to anonymise the data.
The OrdinalEncoder class was used to assign them numbers for the sake of readability.
The datatypes of the dataset were then changed to categorical format after which they were one-hot encoded.

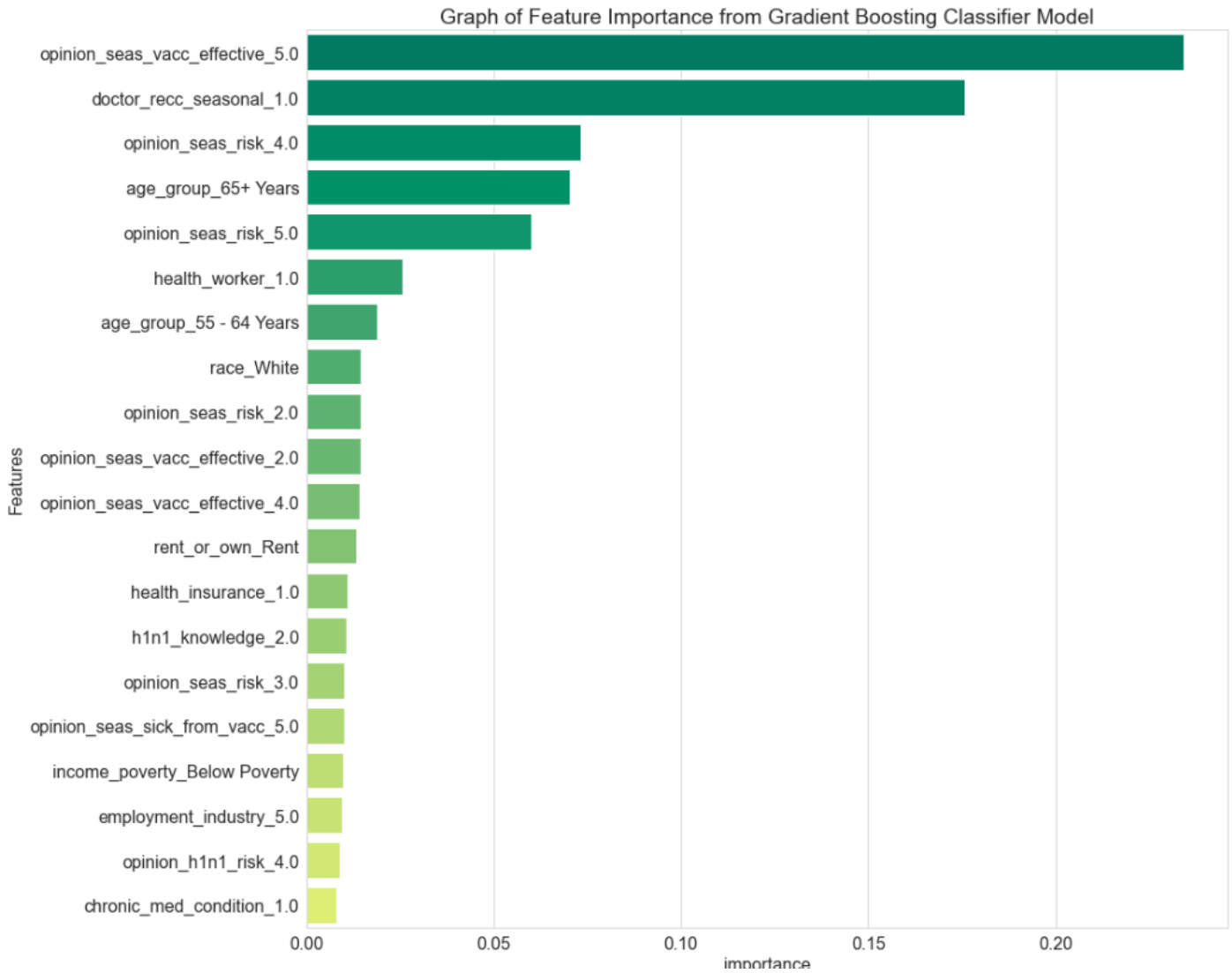
4. Modeling

For the modeling portion, a logistic regression model was first generated as the baseline model with no parameters set. As it was not able to converge, a solver was defined for large datasets and the test data accuracy was found to be 78.22%. Recursive Feature Elimination was employed in order to remove unnecessary features and reduce model complexity.

After this, a Random Forest Classifier was used. Instantiated without tuning hyperparameters, it was found to be overfitting. GridSearchCV was used to look for the best hyperparameters and after tuning, the classifier returned a test accuracy of 77.56%. This was still lower than the baseline model so other alternatives were sought.

Gradient Boosting models were looked at next, specifically XGBoost Classifier and GradientBoosting Classifier. XGBoost Classifier returned a test accuracy of 77.49% with default hyperparameters and after manual tuning based on insights from the documentation and a lot of research, the classifier had a test accuracy of 78.62%. This brought the light back to the researcher’s eyes.
The GradientBoosting Classifier was used next. It returned a test accuracy of 78.13% with default hyperparameters and after a lot of tuning, the test accuracy was determined to be 78.99%. This is roughly 79%. Oh what sweet sounds of joy.

Feature importance was also assessed for all the tested models. The following graph shows the most important features of this dataset:



5. Evaluation

The best model was the one from the GradientBoostingClassifier with an accuracy of 78.98% and a training accuracy of 81.36%. Cross validation was also undertaken 10 times to return a mean accuracy of 78.95%. This means that the model was able to correctly classify 78.98% observations.

Some of the features that were most crucial in predicting the uptake of seasonal flu vaccine were:

- The respondent's opinion on the effectiveness of the vaccine - if a person thought the vaccine was very effective, they were very likely to take it.

- Doctor's recommendation to take the vaccine.
- The respondent's opinion on the risks involved with not being vaccinated - if a person thought they were at high risk of getting sick they were more likely to take the vaccine
- People older than 65 and the general older population.

These feature importances also dominated all the tested models.

While the targeted 80% accuracy was not reached, the project was still considered a success because a lot of time was taken to tweak these models and reduce the noise as much as possible. Any errors past this point can surely be declared irreducible.

6. Challenging the Solution

While this dataset provided some interesting insights, it was quite biased. 80% of the respondents were white people. In addition to that, more than half the sample consisted of college graduates, employed people and people that earn between the poverty line and $75000 annually. This made the models generated biased towards this group of people and this may not be the general situation. This was already visible in the final model as one of the important predictors was identified as being white. It would be unfair to assume that white people were more likely to get vaccinated when the other races were not adequately sampled. Better insights would be generated if the data was more balanced.

It would also be efficient to include all this code in a pipeline and pickle the pipeline so as to encompass the whole preprocessing and modeling part in fewer lines of code.

7. Recommendations

From the models it can be concluded that people's opinions have a huge impact on their likelihood to take vaccines. In light of this, the following recommendations can be made:-

- Public awareness campaigns should be made regarding the effectiveness of the seasonal flu vaccine as well as the risks associated with the flu.
- It would help to emphasise the safety of the vaccines for use by the public.
- Older people are more likely to take the seasonal flu vaccine. The younger population could, therefore, be targeted for such campaigns.