PHASE 3 PROJECT
FLU SHOT LEARNING

Business Understanding

An NGO in the health sector, wants to learn about the trends in the vaccination space.In the year 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.The Gates foundation wants to find out if they can forecast vaccination of an individual based on specific parameters.

Research Question

Can one predict whether a person got seasonal flu vaccine using information they shared about their backgrounds, opinions, and health behaviors?

Objectives

Main objective

This project aims at getting to know whether a person has received the seasonal flu vaccine using machine learning techniques.

Specific Objectives

- To identify which features are significant in creating the model.
- To provide insights on who who is most likely to get seasonal flu vaccine.
- Creating a model with a model accuracy of atleast 80%

Data Understanding

Data source

The data was sourced from Driven Data, but was initially scraped from National Centre for Health Statistics

Data Description

The data in use is from Datadriven made up of 26707 rows and 36 columns(12 categorical columns and 24 are numerical.) Nanely:

'respondent_id'- Unique id
'h1n1_concern'- the concern one has about the virus.
'h1n1_knowledge'- knowledge they have about the H1N1 virus.
'behavioral_antiviral_meds'- If they believe in anti-vaccination.
'behavioral_avoidance'-do they avoid roaming in public.
'behavioral_face_mask'- do they wear a face mask.
'behavioral_wash_hands'- do they regularly wash their hands.
'behavioral_large_gatherings'- do they tend to be in gatherings.
'behavioral_outside_home'- are they usually outdoors.
'behavioral_touch_face'- do they touch their faces often.
'doctor_recc_h1n1'-
'doctor_recc_seasonal',
'chronic_med_condition',
'child_under_6_months',
'health_worker',
'health_insurance',
'opinion_h1n1_vacc_effective',
'opinion_h1n1_risk',
'opinion_h1n1_sick_from_vacc',
'opinion_seas_vacc_effective',
'opinion_seas_risk',
'opinion_seas_sick_from_vacc',
'age_group'- their age group.
'education'- level of education
'race'- their race
'sex' - their gender
'income_poverty'-
'marital_status'- whether they are married or not.

'rent_or_own'- if they rent or own a house.
'employment_status'- whether they are employed
'hhs_geo_region',
'census_msa'- geographical region
'household_adults'-number of adults in the house.
'household_children'-number of children in the house.
'employment_industry'-industr of employment.
'employment_occupation'- what they do for a living.

Data Preparation

1. Loading the data.

The data set was loaded into the notebook.A data frame was then created and displayed to show content of the data as well as how the variables relate to each other.

2. Cleaning data.

The data was analyzed,checked for duplicates and missing values.Missing values were dropped and irrelevant columns were dropped as well.

Exploratory Analysis

Visualizations were created to show distribution of the variables in our data set.Multivariate analysis was also done to show relationships between some variables.

Modelling

At first, a dummy model was created,it had an accuracy of 52.2%. Thereafter,several models were generated with default parameters. The top 3 models were picked,namely: Random Forest model,Logistic regression model and Decision tree classifier.

The picked models went through hyperparameter tuning and picking best hyperparameters to work for the model.

The best performing model was the Random forest model,with an accuracy of 80.9%.

Conclusions and Recommendations

Conclusions about the model were made and recommendations made to the stakeholders.