CFA CONCLAVE

IIT INDORE'S

FLUXUS 2021

A CELESTIAL EUPHORIA

11TH EDITION

CFA CONCLAVE

IIT INDORE'S FLUXUS 2021 A CELESTIAL EUPHORIA 11TH EDITION

Analytica X

A Data Science Challenge

Background

Swine flu, often known as swine flu, is caused by a type of flu virus (influenza) known as H1N1. This is an influenza A virus, and H1N1 is one of numerous flu viruses that can cause seasonal flu. The symptoms of H1N1 flu are similar to those of seasonal flu. In the spring of 2009, scientists identified a strain of the H1N1 flu virus. This virus is a mash-up of viruses that cause disease in people from pigs, birds, and humans. H1N1 caused a respiratory tract infection in humans known as swine flu during the 2009–2010 flu season. Because so many individuals have become ill, the World Health Organization (WHO) proclaimed the H1N1 virus a pandemic in 2009. WHO declared the outbreak over in August 2010. Following the end of the pandemic, the H1N1 flu virus became one of the strains that cause seasonal flu. Fortunately, flu vaccines are now available that can help protect against H1N1 flu (swine flu). Seasonal flu vaccines, including those produced in 2020 and 2021, include the H1N1 flu virus strain. The COVID-19 pandemic struck the world in 2019. Following the findings of several studies on H1N1 and SARS, experts concluded that COVID-19 is likewise categorized as a seasonal flu disease. The sickness is also delaying attempts to produce a vaccine to combat seasonal flu.

CFA CONCLAVE

IIT INDORE'S FLUXUS 2021 A CELESTIAL EUPHORIA 11TH EDITION

Goal

Your goal is to anticipate how likely people are to get H1N1 and their yearly flu vaccine. You would specifically forecast two probabilities: one for vaccine_h1n1 and one for vaccine_seasonal. Each row in the data set represents one individual from the 2009 National H1N1 Flu Survey (NHFS) by CDC.

Needs Must Be Done:

1. Exploratory Data Analysis
2. Research Design

Evaluation:

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. In Python, you can calculate this using sklearn.metrics.roc_auc_score for this multilabel setup with the default average="macro" parameter..

CFA CONCLAVE

IIT INDORE'S FLUXUS 2021 A CELESTIAL EUPHORIA 11TH EDITION

Dataset

Training_set Features

Training set Labels

Test Set Features

Dataset Description

For this competition, there are two target variables:

- h1n1_vaccine - Whether respondent received H1N1 flu vaccine.
- seasonal_vaccine - Whether respondent received seasonal flu vaccine.

Both are binary variables: 0 = No; 1 = Yes.

Some respondents didn't get either vaccine, others got only one, and some got both.

This is formulated as a multilabel (and not multiclass) problem.

Features

You are provided a dataset with 36 columns.
The first column respondent_id is a unique and random identifier. The remaining 35 features are described below.
For all binary variables:
0 = No; 1 = Yes.

- h1n1_concern - Level of concern about the H1N1 flu.
  0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.
- h1n1_knowledge - Level of knowledge about H1N1 flu.
  0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.
- behavioral_antiviral_meds - Has taken antiviral medications. (binary).
- behavioral avoidance - Has avoided close contact with others with flu-like symptoms. (binary)
- behavioral_face_mask - Has bought a face mask. (binary)
- behavioral_wash_hands - Has frequently washed hands or used hand sanitizer. (binary)
- behavioral_large_gatherings - Has reduced time at large gatherings. (binary)
- behavioral_outside_home - Has reduced contact with people outside of own household. (binary)
- behavioral_touch_face - Has avoided touching eyes, nose, or mouth. (binary)
- doctor_recc_h1n1 - H1N1 flu vaccine was recommended by doctor. (binary)
- doctor_recc_seasonal - Seasonal flu vaccine was recommended by doctor. (binary)

● chronic_med_condition - Has any of the following chronic medical conditions: asthma or an other lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness. (binary)
● child_under_6_months - Has regular close contact with a child under the age of six months. (binary) health_worker - Is a healthcare worker. (binary) health_insurance - Has health insurance. (binary)
● opinion_h1n1_vacc_effective - Respondent's opinion about H1N1 vaccine effectiveness. 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.
● opinion_h1n1_risk - Respondent's opinion about risk of getting sick with H1N1 flu without vaccine. 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
● opinion_h1n1_sick_from_vacc - Respondent's worry of getting sick from taking H1N1 vaccine. 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.
● opinion_seas_vacc_effective - Respondent's opinion about seasonal flu vaccine effectiveness. 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.
● opinion_seas_risk - Respondent's opinion about risk of getting sick with seasonal flu without vaccine. 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
● opinion_seas_sick_from_vacc - Respondent's worry of getting sick from taking seasonal flu vaccine. 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried. ● age_group - Age group of respondent.

- education - Self-reported education level.
- race - Race of respondent.
- sex - Sex of respondent.
- income_poverty - Household annual income of respondent with respect to 2008 Census poverty thresholds.
- marital_status - Marital status of respondent.
- rent_or_own - Housing situation of respondent.
- employment_status - Employment status of respondent.
- hhs_geo_region - Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.
- census_msa - Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.
- household_adults - Number of other adults in household, top-coded to 3.
- household_children - Number of children in household, top-coded to 3.
- employment_industry - Type of industry respondent is employed in. Values are represented as short random character strings.
- employment_occupation - Type of occupation of respondent. Values are represented as short random character strings.

Submission Format

You need to submit the following files:

1. Sample Submission File

The format for the submission file is three columns: respondent_id, h1n1_vaccine, and seasonal_vaccine. The predictions for the two target variables should be float probabilities that range between 0.0 and 1.0. Because the competition uses ROC AUC as its evaluation metric, the values you submit must be the probabilities that a person received each vaccine, not binary labels. As this is a multilabel problem, the probabilities for each row do not need to sum to one.

2. Research Design- You can use any data visualization tools for that. E.g - Tableau etc. 3. Report in the form of pdf not exceeding 10 pages