Proposal

The data selected for the project is from the DriveData project: Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines.

From DrivenData's website:

About DrivenData

DrivenData works on projects at the intersection of data science and social impact, in areas like international development, health, education, research and conservation, and public services. We want to give more organizations access to the capabilities of data science, and engage more data scientists with social challenges where their skills can make a difference.

In pursuit of these goals, DrivenData runs online machine learning competitions with social impact and works directly with mission-driven organizations to drive change through data science and engineering. We also maintain a number of popular open-source projects for the data science community, and have shared the prize-winning solutions from past competitions openly on GitHub for anyone to learn and build from.

Overview

Predict whether people got H1N1 and seasonal flu vaccines using information shared about backgrounds, opinions, and health behaviors.

In this challenge, we will take a look at vaccination, a key public health measure used to fight infectious diseases. 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.

Data Source

The data for this competition comes from the National 2009 H1N1 Flu Survey (NHFS) and is provided courtesy of the United States National Center for Health Statistics.

In their own words:

The National 2009 H1N1 Flu Survey (NHFS) was sponsored by the National Center for Immunization and Respiratory Diseases (NCIRD) and conducted jointly by NCIRD and the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). The NHFS was a list-assisted random-digit-dialing telephone survey of households, designed to monitor influenza immunization coverage in the 2009-10 season.

The target population for the NHFS was all persons 6 months or older living in the United States at the time of the interview. Data from the NHFS were used to produce timely estimates of vaccination coverage rates for both the monovalent pH1N1 and trivalent seasonal influenza vaccines.

The NHFS was conducted between October 2009 and June 2010. It was one-time survey designed specifically to monitor vaccination during the 2009-2010 flu season in response to the 2009 H1N1 pandemic. The CDC has other ongoing programs for annual phone surveys that continue to monitor seasonal flu vaccination.

Data Restrictions

The source dataset comes with the following data use restrictions:

- The Public Health Service Act (Section 308(d)) provides that the data collected by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), may be used only for the purpose of health statistical reporting and analysis.
- Any effort to determine the identity of any reported case is prohibited by this law.
- NCHS does all it can to ensure that the identity of data subjects cannot be disclosed. All direct identifiers, as well as any characteristics that might lead to identification, are omitted from the data files. Any intentional identification or disclosure of a person or establishment violates the assurances of confidentiality given to the providers of the information.

Therefore, users will:

- Use the data in these data files for statistical reporting and analysis only.
- Make no use of the identity of any person or establishment discovered inadvertently and advise the Director, NCHS, of any such discovery (1 (800) 232-4636).
- Not link these data files with individually identifiable data from other NCHS or non-NCHS data files.
- By using these data, you signify your agreement to comply with the above requirements.

Data Labels

The goal is to predict how likely individuals are to receive their H1N1 and seasonal flu vaccines. Specifically, predicting two probabilities: one for h1n1_vaccine and one for seasonal_vaccine.

Each row in the dataset represents one person who responded to the National 2009 H1N1 Flu Survey.

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

Provided is 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.

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.

Additional (Personal) Comments

I selected this dataset as I am keenly interested in the use of data science and artificial intelligence for social good. DrivenData's values align with my own personal beliefs.

This dataset is relevant to the medical/health domain and contains a lot if interesting features, including many protected characteristics (race, sex, age, education, income) for which I would like to DALEX's fairness and unbiasdness tools.

As this comes from a competition, the data is assured a decent level of quality/cleanliness and is well documented (both metadata and features). Furthermore, there is a benchmark blog post providing additional information on how to get started.

I would like to apply a variety of different models to this project and compare how they differ in performance and fairness. This provides the opportunity to use DALEX's arena module, and `ExplainerDashboard()` for multiple models.

Some initial questions:

- Is there any bias in our dataset?
  - The topic of vaccines has become very politicised and can be a polarising topic. There are a lot of opinion based features in this dataset. It will be interesting to see if there is any 'baked-in' bias.
- How accurate can we predict the likelihood of an individuals receiving their H1N1 and seasonal flu vaccines?
  - Specifically using whitebox vs blackbox models
    - Interpretability of AI tools to assist medical desicions needs to be clearly interpretable, understandable, and explainable.
- Which features are most important in our predictions?
  - In particular, protected characteristics, and these opinion based features.
    - It will be interesting to compare initial expectations on how influential features will be, vs their *actual* importance.
  - Can we do some form of feature engineering?
- How "fair" are our models, and can this be improved?
  - Excluding certain features (such as protected characteristics, perhaps)