**Project: Predict H1N1 and Seasonal Flu Vaccines**

Website: <https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/>

**Introduction**

This project is based on the competition Driven Data published related to H1N1 flu and the Seasonal one. It is about 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 [1].

Data has been collected on some aspects related to this problem, in general the data can be divided in a survey, demographics and some information related to the use of face mask, frequently washing hands, people contact. Etc.

It is a very unusual the data set because it has two targets, one for H1N1 and other for Seasonal flu, in both probabilities are wanted. The goal is to predict how likely individuals are to receive their H1N1 and seasonal flu vaccines.

This document analyses the elements that, eventually affect patients who gets these kinds of flu. One of the problems in this dataset is the high quantity of missing values in almost all the features, this situation limits the obtaining of reliable results.

**Dataset description**

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 [1].

**Exploratory Data Analysis**

The exploratory data analysis was performed by analyzing the context of the problem and the characteristics of the variables as observed in the directory of the project, in this case “Exploration.ipynb” in the notebooks directory of the project.

Starting with the feature label corresponding to two targets “*h1n1\_vaccine*” and “*seasonal\_vaccine*”, each one with two possible values “0” or “1”. It had been observed that the features *h1n1\_vaccine* is imbalanced, and the other one *seasonal\_vaccine* is balanced.

It had been mentioned above that this data may be divided in a survey information, demographics and some information related to the use of face mask, frequently washing hands, people contact. Etc. The exploratory analysis will consider the mentioned division.

It is important to mention that, in general, there are lots of features that do not need further processing because its distribution statistics metrics tell us that are ok for the analysis. But the main problem in this dataset was the quantity of missing values in almost every feature, and in some cases much more that 60%. This problem is something that limit the obtaining of reliable results.

The following features are related to some general characteristics of people behavior, like the use of face mask, the avoidance of large gatherings, touching face, if the person is a health worker, etc. All of them are binary features that are, kind of balanced, then it did not need any further processing.

* *behavioral\_antiviral\_meds*
* *behavioral\_avoidance*
* *behavioral\_face\_mask*
* *behavioral\_wash\_hands*
* *behavioral\_large\_gatherings*
* *behavioral\_outside\_home*
* *behavioral\_touch\_face*
* *doctor\_recc\_h1n1*
* *doctor\_recc\_seasonal*
* *chronic\_med\_condition*
* *child\_under\_6\_months*
* *health\_worker*
* *health\_insurance*

A couple of interesting features are related to people concern and knowledge about these kinds of flu, the following are the features:

* *h1n1\_concern*
* *h1n1\_knowledge*

The following features are related to the mentioned survey that took place. These features obtain the perceptions of the people about some relevant aspects. In all these cases there is a response option graded, in general, from high to low. There were also not found feature distributions that need any further processing.

* *opinion\_h1n1\_vacc\_effective*
* *opinion\_h1n1\_risk*
* *opinion\_h1n1\_sick\_from\_vacc*
* *opinion\_seas\_vacc\_effective*
* *opinion\_seas\_risk*
* *opinion\_seas\_sick\_from\_vacc*

The following features are related to demographic information, most of these features are categorical, in this case, it had been used one hot encoding for the processing, which needed to change some of those features values not to have spaces in them.

* *age\_group*
* *education*
* *race*
* *sex*
* *income\_poverty*
* *marital\_status*
* *rent\_or\_own*
* *employment\_status*
* *household\_adults*
* *household\_children*
* *employment\_industry*
* *employment\_occupation*

A correlation analysis had been done for continuous and categorical features, in both cases there were not found any high correlation to get rid of some features.

**Performance Metric**

The performance metric is evaluated using the area under the receiver operating characteristic curve (ROC AUC) for each of the two target variables. In the case of the feature *h1n1\_vaccine* was imbalanced that is why some preprocessing to balance this feature had been done.

**Modeling**

The dataset had been divided according to the two targets it has, every technique used was applied to the same dataset using target *h1n1\_vaccine* and then *season\_vaccine*. The analysis had included the use blending and stacking process to try different options to get good results. Also, grid search had been used for hyperparameter tuning. It had been a “Trial an error” process for identifying good combinations of techniques and processes that generate great predictions.

**Prediction Results**

The characteristic of the dataset having two targets needed a processing to get together both prediction for the submission. The project had been built in python using vscode®, and submitted to Data Driven®, getting a result of 0.8426 that is among 14% of the ranking.

**Bibliography**

[1] <https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/>