H1N1 and Seasonal Flu Vaccine Prediction

Data Science Intro Project

February 16, 2021

Abstract

In this project I try to predict a subject's probability of taking the H1N1 and seasonal flue vaccines from the input data provided by the competition. I use an iterative approach based on gathering insight from data using EDA, performing the preprocessing of the data, selecting the best model to be used with the data, tuning that same model (or models) and then joining every step to make my final predictions. The final model was chosen as a stacking classifier of different models which performed in the top 14% of participants with an AUROC of 0.8442.

Contents

1 Introduction

The aim of this report is to provide insight into the framework I used to achieve a top 14% entry to the competition Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines.

I began by looking at the competition problem description.

1.1 Competition Objective

The objective of the competition is to predict the probability of an individual to take their H1N1 and seasonal flu vaccines. This puts the problem as a multilabel classification one where we have to predict two labels: *h1n1_vaccine* and *seasonal_vaccine*.

It is important to notice that both labels are binary and any combination of the two for each individual is possible.

1.2 Dataset Features

The dataset is composed of 36 columns in which one of them is the *respondent_id* which identifies the individual and the remaining 35 are features we can use in modeling.

Among these features we have multiple regarding the individual's opinion, behaviour, etc. regarding the deseases as well as some socio-economic features. I will not get into detail on each feature here since I will explain some of them further in Sec. 3.

For now, it is sufficient we know that there are both numerical and categorical variables - which will need to be encoded before modeling.

1.3 Evaluation Metric

The metric used to evaluate the result of the model predictions is the area under the receiver operating characteristic curve (AUROC).

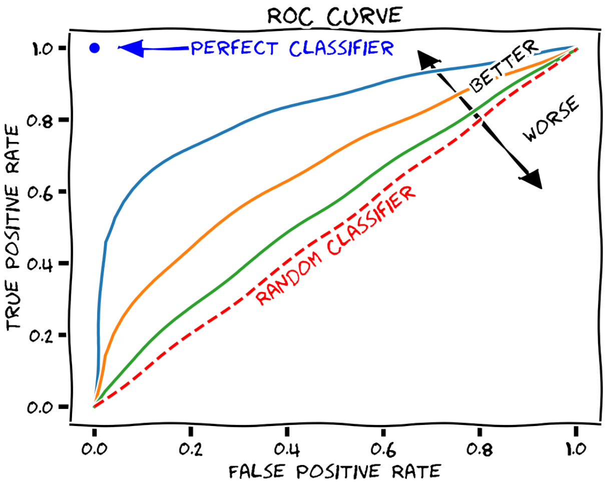

Figure 1: AUROC evaluation metric. [1]

This metric calculates the area under curves such as the ones in Fig. 1.3. The curves are a plot of the model true positive rate (TPR) vs its false positive rate (FPR). Thus, the curve for a random classifier is a straight line at a slope of 1 and the perfect classifier is a single point with $TPR = 1$ and $FPR = 0$. As such, the AUROC for a random classifier is 0.5 and the AUROC of a perfect classifier is 1. Taking this into account we see that we want to increase our model's AUROC, to increase its accuracy.

2 Solution Approach

3 Exploratory Data Analysis

4 Model Selection

5 Model Tuning

6 Model Fitting and Predictions

7 Conclusions and Future Work