Flu Shot Learning

Predicting H1N1 and Seasonal flu vaccination

Jorge Pais, José Baptista e Pedro Duarte

MsC Electrical and Computer Engineering

Final project of the Machine Learning course M.EEC006

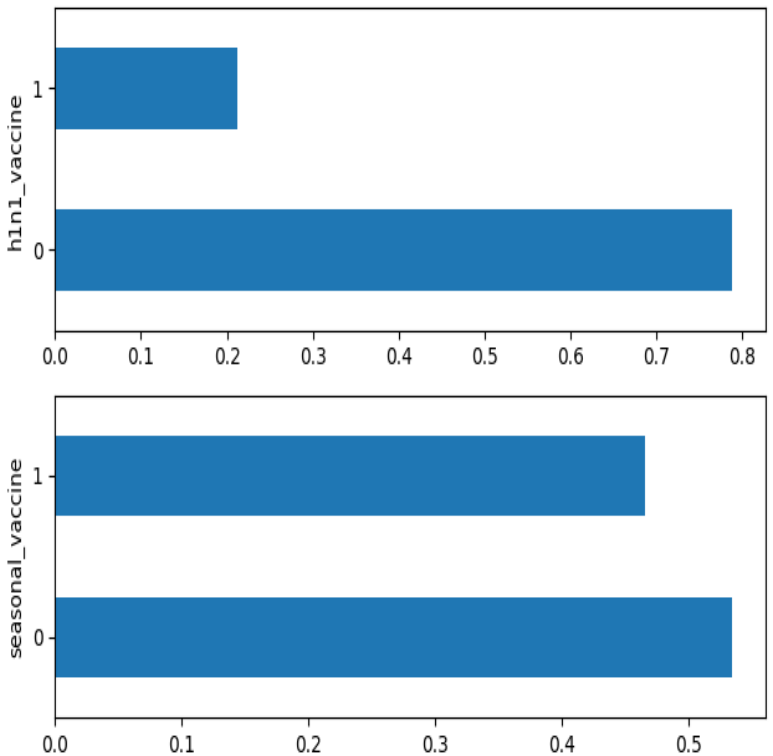
Problem Description

Context and Objectives

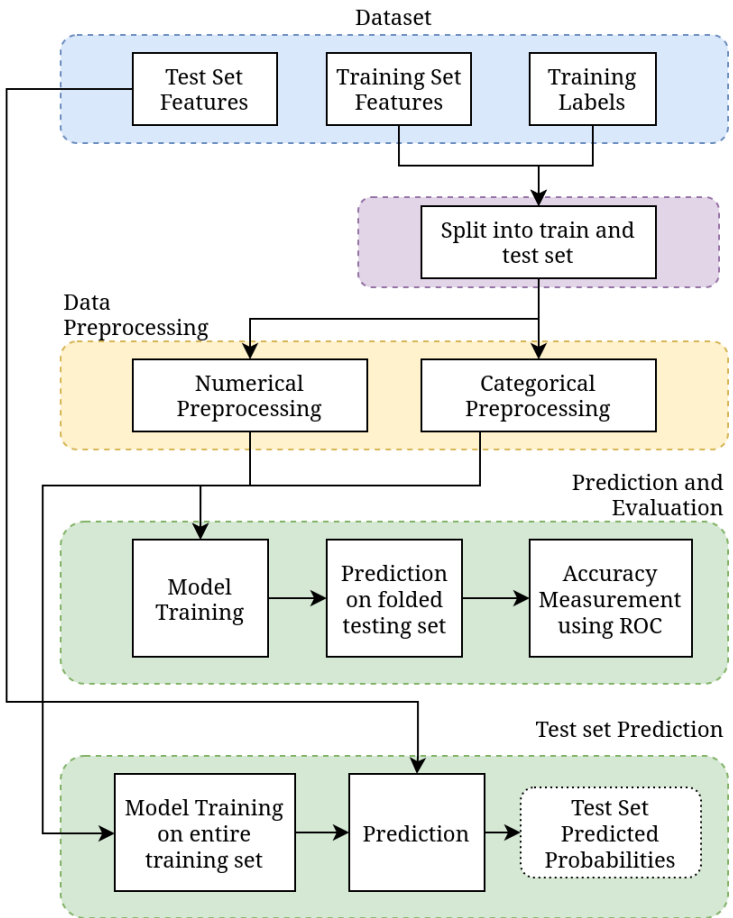
- Competition hosted by DrivenData
- The objective is to, based on socio-economical background, opinions, concerns, e.t.c, predict if whether a individual has taken the H1N1 or Seasonal Flu vaccines
- Performance measurements by using Receiver Operating Characteristic (ROC)

Data Set

- The dataset was the 2009 National H1N1 Flu Survey, consisting of 35 attributes with both numerical (ordinal/binary) and categorical features.
- Two-labels to predict: `h1n1_vaccine` and `seasonal_vaccine`
- Two sets of 26707 respondents for the training data and evaluation, including and excluding labels, respectively.



Implementation



Several standard classification models were utilized:

- Logistic Regression
- Naive Bayes Classifiers
- Decision Trees
- K-Nearest Neighbors

Along with some gradient boosting algorithms:

- *XGBoost*
- *CatBoost*

Results

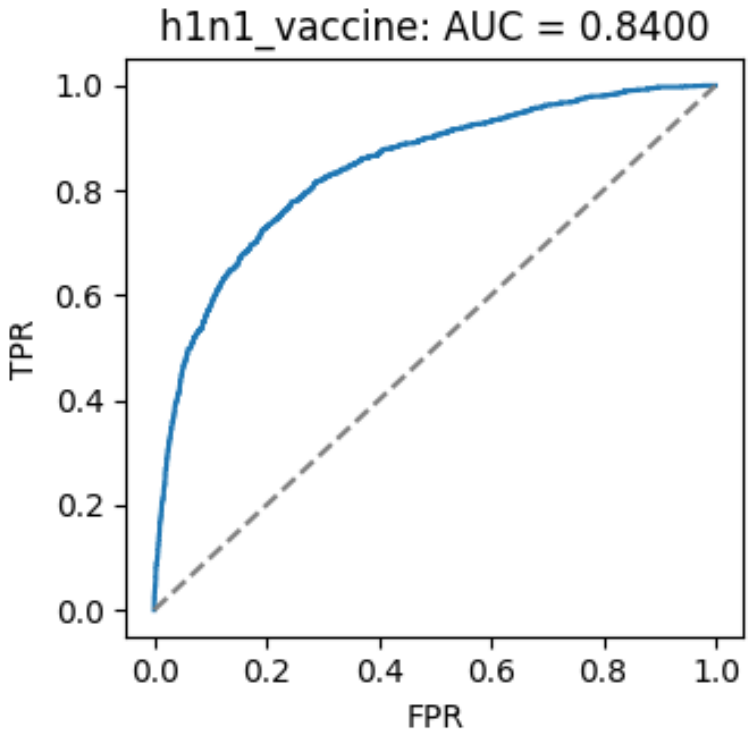
For the standard models, the results on folded training data:

Model	ROC AUC score
Logistic regression (C = 0.01)	0.84555
Logistic regression (C = 0.1)	0.84655
Logistic regression (C = 1)	0.84640
Logistic regression (C = 10)	0.84635
Gaussian Naive Bayes	0.72947
Multinomial Naive Bayes	0.79467
Decision Tree	0.66738
KNearestNeighbors (k=169)	0.82478

And for gradient boosting:

Model	ROC AUC score
XGBoost (Default parameters)	0.83732
CatBoost with One-Hot encoding	0.85069
CatBoost with specified cat_features	0.86907

Example of ROC plot:



Results

Competition

Model	**ROC AUC score**
Logistic Regression (C=0.1)	0.8356
K-Nearest Neighbors (k=169)	0.8122
CatBoost with One-Hot encoding	0.8430
CatBoost with specified cat_features	0.8616

Final place on the leaderboard: 285th

Thank you for listening!