Flu Shot Learning: Prediction H1N1 and Seasonal Flu Vaccines

Anders Poirel
University of California, Santa Cruz

October 25, 2020

1 Introduction

This paper describes our approach in the Flu Shot Learning on Driven Data [1] The goal of the competition is to predict how likely individuals are to receive their H1N1 and seasonal flue vaccines. Specifically, participants are asked to predict a probability for each vaccine. Competition ranking is based on the ROC AUC of predictions on a hold-out test set.

2 Model

Our approach uses a logistic regression model. Our pipeline was developed using `tidyverse`[4] and `tidymodels`[3]. Numerical features were standardized and imputed using the mean. Categorical features were one-hot encoded and imputed with "missing" flags.
We tuned a single hyper-parameter for this pipeline, logistic regression's regularization parameter C, using cross-validation on 5 folds.

3 Results

3.1 Cross-validation

Cross-validation AUC scores suggest that the model does not over-fit for any choice of C. Indeed, performance degrades for low values of C (stronger regularization). The best tested value is $C = 0.010$ for both the `h1n1_vaccine` and `seasonal_vaccine` models. Selecting higher values of C resulred in inconsistent model performance.

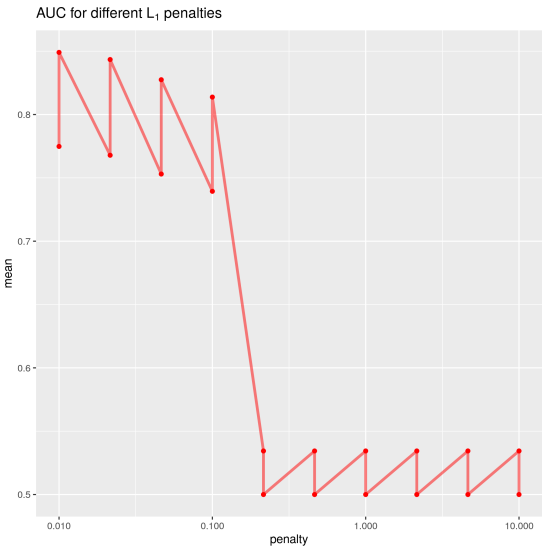

Figure 1: Cross validation mean AUC for `seasonal_vaccine`

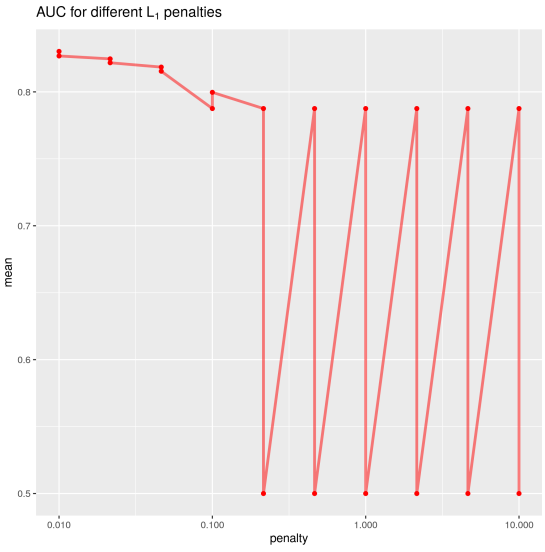

Figure 2: Cross validation mean AUC for `h1n1_vaccine`

3.2 Hidden test set

After submission to the competition website, the model's predictions scored 0.8342 ROC AUC on the hidden test set, enough to beat the organizer's benchmark (.8185). At time of writing, this score places 181st out of 948 on the competition leaderboard.

4 Future Work

Our model's validation curve shows no sign of over-fitting. As such, it is likely that higher scores can be achieved by using a more flexible model (e.g. gradient boosted trees). Furthermore, model stacking will generally improve results [2], in particular in this kind of machine learning competition, where test sets are guaranteed to be sourced from the same distribution as training data.

References

[1] Peter Bull, Isaac Slavitt, and Greg Lipstein. Harnessing the power of the crowd to increase capacity for data science in the social sector, 2016.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[3] Max Kuhn and Hadley Wickham. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*, 2020.

[4] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.