Predicting Flu Outbreak with Random Forest

Katie Will

Fall 2017

Abstract

Influenza takes a significant toll on the health of the United States each year. One technique that could help mitigate the impact of the virus is seasonal forecasting. Since 2013, the CDC has coordinated competitions to predict flu outbreak. In this paper we discuss the process of building random forest models for predicting seasonal flu activity. Other machine learning techniques, such as extreme gradient boosting and neural network, are compared. Models are built using six targets considered in the competition and then compared using cross validation error. To test the performance of the superior model for each target, predictions are made for the 2016-17 flu season.

Contents

1 Introduction

Influenza places a substantial burden on the health of the people in the United States. The Center for Disease Control and Prevention (CDC) estimates that between 9.2 and 60.8 million citizens become infected with the illness and approximately 12,000 to 56,000 deaths occur as a result each year [8]. Not only is there a considerable amount of variability in the number of people affected each year as suggested by the ranges reported by the CDC, there is also great variability in the influenza virus itself. Because the virus is constantly changing, surveillance is crucial throughout the flu season. Typically, the ”flu season” lasts through the fall and winter months with peak activity around late December to February.

To lessen the burden of the influenza virus in the United States, the CDC collects and analyzes information on the outbreak activity year-round, leading to a more thorough exploration of influenza surveillance [2]. Along with that, the CDC works with outside collaborators to improve seasonal influenza forecasts. Since 2013, the Influenza Division at the Centers for Disease Control and Prevention has coordinated seasonal influenza prediction challenges. The goal of these challenges has been to provide national and regional forecasts for the next four weeks and for the entire influenza season, each week during the influenza season [6]. Developing a more timely and forward outlook on the flu cycle could bring substantial benefits to national healthcare, such as providing insight when allocating resources for communications, disease prevention and control. These collaboration efforts are helping to improve the accuracy and consistency of influenza forecasting.

Other companies and independent researchers have taken it upon themselves to contribute to influenza surveillance and prediction. Trends on Google, Wikipedia and Twitter have and continue to be examined to support a greater understanding of the viral infection. Utilizing popular search engines and social media sites shows great potential in advancing the way influenza forecast are derived.

The goal of the analysis discussed in this paper is to build effective influenza prediction models using random forest and other machine learning techniques. We will combine CDC surveillance and Wikipedia search information to predict targets considered in the influenza prediction challenges coordinated by the Influenza Division at the Centers for Disease Control and Prevention with machine learning techniques. Finally, we test each model by making predictions on the 2016-17 flu season.

2 Data

The data used in this analysis was collected from various sources using an API created by members of DELPHI (Developing the Theory and Practice of Epidemiological Forecasting) [4]. A total of four unique data sets–Flu View, Wiki Access Logs, Google Flu Trends, and Health Tweets–provide estimates of seasonal influenza activity. Information from Flu View and Wiki Access Logs were used to created the prediction models constructed in this paper.

2.1 FluView

The Epidemiology and Prevention Branch in the Influenza Division at CDC creates a weekly U.S. influenza surveillance report by collecting, compiling and analyzing information on influenza activity throughout the year. This report, known as FluView, provides a national and regional picture of influenza activity each season. Many health departments, laboratories, providers, clinics, vital statistics offices and emergency departments partner with the CDC to make the FluView possible. One category of information that is emphasized in this paper is outpatient illness surveillance collected through the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). The surveillance network consists of more than 2,800 enrolled outpatient (outside the hospital) healthcare providers in the 50 states, Puerto Rico, the District of Columbia and the U.S. Virgin Islands reporting information on more than 39 million patient visits annually. Each week, the CDC receives data from approximately 2,000 providers about the total number of patients seen and the number of patients experiencing influenza-like illness. This information is consolidated and ILI estimates are created for the current week as well as for all weeks prior; these are known as 'lagged' estimates. A patient is classified as having Influenza-like illness (ILI) if they experience a fever (temperature of 100F or greater) along with a cough and/or a sore throat that cannot be explained by any other illness. The healthcare providers report to the CDC on a voluntary basis and can change/update their data at anytime throughout the season. [2]

Observations in the FluView data backdate to the 1997-98 season and are collected at a national and regional level. The regions found in the FluView data are equivalent to the U.S. Department of Health and Human Services (HHS) office regions identified in figure 1 below.

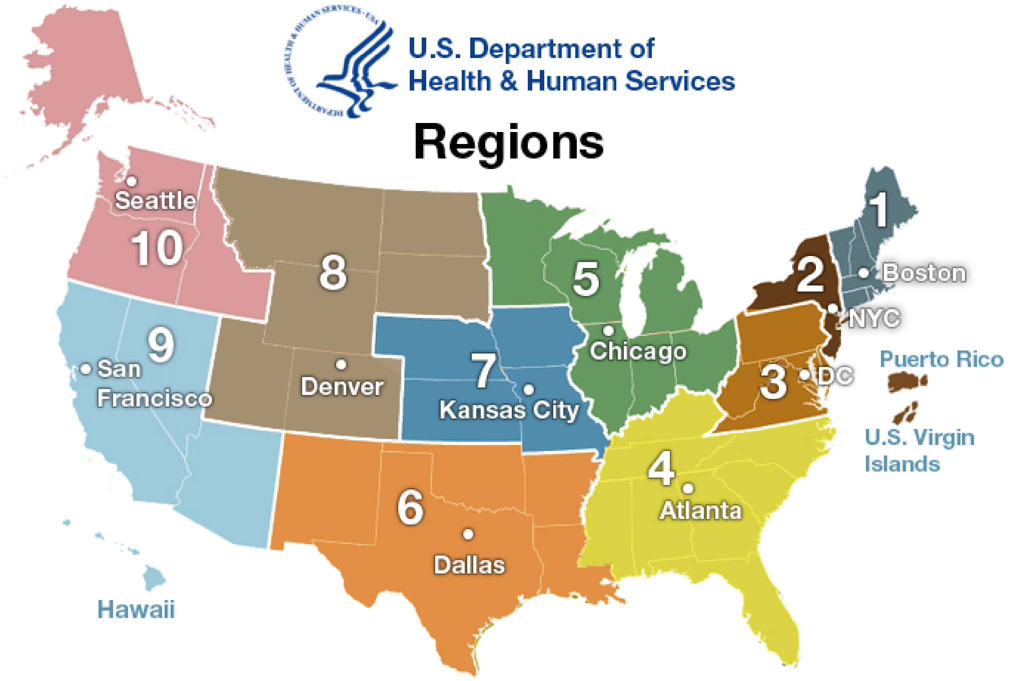

Figure 1: U.S. Department of Human and Health Services (HHS) office regions

The CDC identifies each unique observation in FluView with a region, year, and two epi weeks: the week of the ILI estimates and the week the estimates were issued. An epi week is a standardized method of classifying weeks whose values range from 1 to 52 or 53, depending on the year. Each epi week starts on a Sunday with the first epi week equating to the first week of the calendar year that has at least four days [3]. For this analysis, the most recent value for ILI each week is seen as the best estimate, rather than the value during the week the report was issued.

2.2 External Data Sources

Many sources outside of the U.S. Outpatient Influenza-like Illness Surveillance Network have been used to create an image of seasonal influenza activity. Search trends on Wikipedia have been analyzed to identify patterns in influenza activity since 2009. The Wiki Access Logs data set sourced using the

API includes number of page visits for influenza-related Wikipedia articles. Some influenza-related topics include different strains of flu (avian, swine, A, C, H1N1, etc.) and flu symptoms (cough, vomiting, fever, headache, shivering, etc.). In total, weekly page visits are collected on 54 influenza-related topics.

Other sources that are being utilized to examine flu activity are Google and Twitter. A Google web service, called Google Flu Trends, creates estimates of influenza activity based on volume of certain search patterns [4]. Regional flu estimates have been reported weekly since the 2003-04 flu seasons. An method was developed by the Johns Hopkins Social Media and Health Research Group that analyzes language patterns in Tweets to provide estimates of influenza activity [5]. The Health Tweet analysis was developed at a national level and began during the 2011-12 flu season. Both the Google Flu Trends and Health Tweets data are available via a private account, unlike Wiki Access Logs. For this reason, they were not considered in the creation of the prediction models.

2.3 Created Variables

Information from Wikipedia, past seasons, and past weeks was combined to build reliable predictive models. The raw data from Flu View and Wikipedia was reformatted to be used in a prediction setting. Each observation for prediction is one week in a flu season.

The target variables considered in this analysis needed to be extracted from Flu View. Those targets are the percent of influenza-like illness (the number of patients experiencing influenza-like illness divided by the total number of patients seen by providers) observed at the end of a flu season one week, two weeks, three weeks, and four weeks ahead from the week of the forecast. Also considered are the peak week and the peak intensity of a flu season. The predictors ili_week1, ili_week2, ili_week3, and ili_week4 were created by lagging the percent ILI estimates observed at the end of the season by 1, 2, 3, and 4 weeks. The other two targets, $peak\_week$ and $peak\_ili$, were found by finding the max ILI percent within each season. Onset of the season, defined as the week when the percent ILI exceeds the regional baseline value for three consecutive weeks, is usually another target of interest for the CDC competition. This target is not of interest in this analysis due to some flu seasons not experiencing an onset week in certain regions.

Data is collected on a total of 54 influenza-related Wikipedia articles.

For purposes of prediction, the number of page visits for each article topic is considered as a predictor. Since the Wikipedia data is collected at a National level, observations are repeated across each region.

Information about the flu season being forecasted and previous seasons is captured by variables that were also created from Flu View. A list of the variables and their descriptions can be found below in table 1. These variables were used to predict each of the target variables considered. The Flu View data is not complete in that there is lack of 'lag' observations for seasons before the 2010-11 flu season. For this reason, only seasons 2010-11 through 2016-17 were considered when building prediction models. Another variable that was created as a predictor was *season_week*. This variable, like *epiweek*, identifies the week of each observation. In the data set, *season_week* 1 is equivalent to *epiweek* 40 (first week in October) and the last *season_week* corresponds to *epiweek* 20 (mid-May). The maximum *season_week* in each flu season is 33 or 34, depending on the calendar year. This variable was used in the analysis instead of *epiweek* because it is continuous throughout the flu season and does not break at the end of the calendar year. Lastly, a variable indicating the region of each observation was considered as a predictor rather than building separate prediction models for each region. The 17 variables in table 1, *season_week*, *region*, and the 54 variables from Wikipedia were combined to create a prediction model for each of the 6 targets considered in this analysis.

Table 1: Variables created for prediction analysis

Variable	Description
$peak\_ili\_lastyr$	Peak percent ILI in prior season
$peak\_week\_lastyr$	Week in prior season that experienced peak percent ILI
ili_lag0	Percent ILI estimate for week of forecast
ili_lag1	Percent ILI estimate for 1 week prior to week of forecast
ili_lag2	Percent ILI estimate for 2 weeks prior to week of forecast
ili_lag3	Percent ILI estimate for 3 weeks prior to week of forecast
ili_lag4	Percent ILI estimate for 4 weeks prior to week of forecast
ili_lag01_diff	Difference between ili_lag0 and ili_lag1
ili_lag12_diff	Difference between ili_lag1 and ili_lag2
ili_lag23_diff	Difference between ili_lag2 and ili_lag3
ili_lag34_diff	Difference between ili_lag3 and ili_lag4
ili_lastyr_week0	Most recently observed percent ILI for the forecast week in the prior season
$ili\_lastyr\_lag0week0\_diff$	Difference between ili_lag0 and ili_lastyr_week0
$ili\_lastyr\_week10\_diff$	Difference between most recently observed percent ILI for the week prior to the forecast week in the prior season and ili_lastyr_week0
$ili\_lastyr\_week20\_diff$	Difference between most recently observed percent ILI for 2 weeks prior to the forecast week in the prior season and ili_lastyr_week0
$ili\_lastyr\_week30\_diff$	Difference between most recently observed percent ILI for 3 weeks prior to the forecast week in the prior season and ili_lastyr_week0
$ili\_lastyr\_week40\_diff$	Difference between most recently observed percent ILI for 4 weeks prior to the forecast week in the prior season and ili_lastyr_week0

3 Methods

Random forest models were built to predict the 6 targets for the 2017-18 flu season. A total of $p = 73$ predictors were used to build each model for the 6 different target variables. The assumption of independence required for random forest fails in this analysis because of the time series nature of the data. However, for the sake of this project, we are going to build random forest models assuming the criteria is met and check the credibility of the predictions.

Random forest is a tree-based method that is extremely effective when it comes to both classification and regression prediction. One reason why this technique is so powerful is its tendency to not overfit [1]. For regression, this algorithm builds a fixed number of decision trees (500 in this analysis) on bootstrapped training samples and averages their results to calculate the final predictions. Each time a split in a decision tree is considered, a random sample of *mtry* predictors is chosen as split candidates from the full set of p predictors. A new sample of *mtry* predictors is taken at each split and only one of *mtry* predictors is selected at each split [7]. The randomness implemented at this stage in the algorithm reinforces the prediction power for random forest [1].

3.1 Tuning Model Parameters

Two random forest models were created for each of the 6 targets, one model using the algorithm's default *mtry* and one model using a tuned *mtry*. The random forest models in this analysis were built using the `randomForest` package via the `train` function in the `Caret` package. The default *mtry* for the random forest algorithm in this package is $floor(p/3) = 24$. The `train` function was used to build the models since it enables tuning of model parameters using a grid search technique. Repeated 10-fold cross validation was implemented on seasons 2010-11 to 2015-16 for each *mtry* value in the grid, and root mean squared error (RMSE) was compared to determine the optimal *mtry* value in each model. A list of the tuned *mtry* values for each target model is found in table 2 below. Plots of the cross validation RMSE are shown in figure 2.

Table 2: Random forest tuned model parameters

	ili_week1	ili_week2	ili_week3	ili_week4	$peak\_week$	$peak\_ili$
mtry	22	39	40	38	40	40

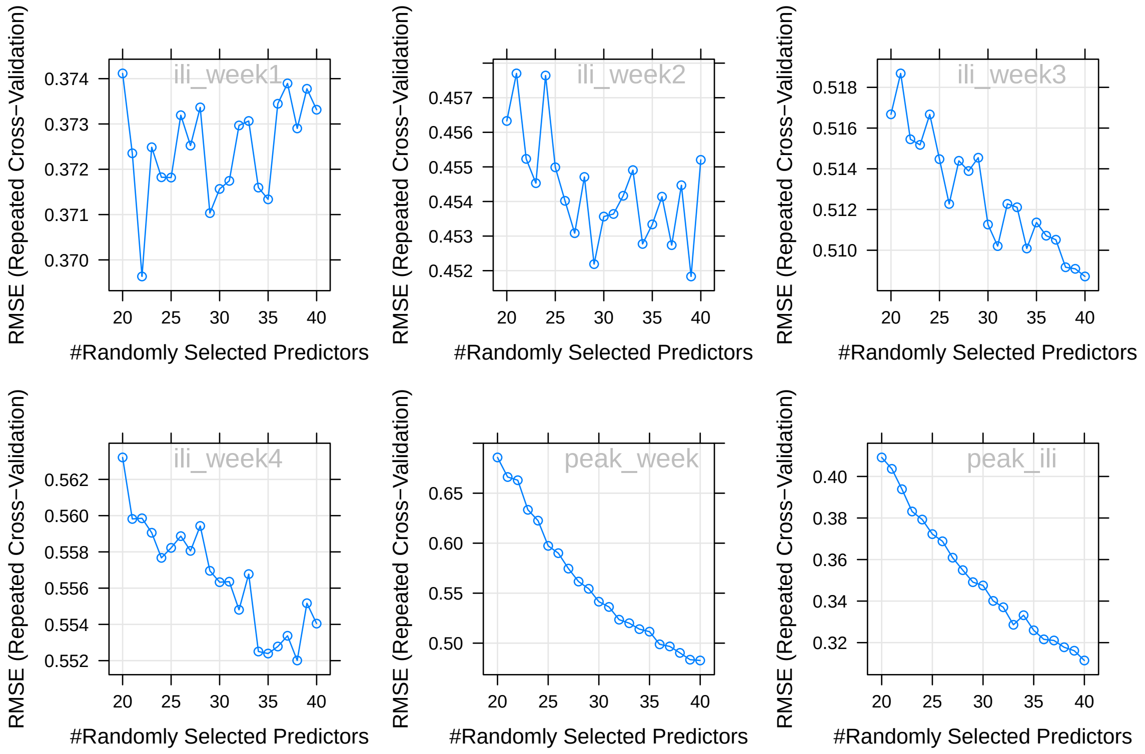

Figure 2: Random forest cross validation RMSE for tuning *mtry*

3.2 Leave-one-season-out Cross Validation

Performance of the tuned and default random forest (RF) models was compared for each target using leave-one-season-out cross validation on seasons 2010-11 to 2015-16. The mean absolute percentage error (MAPE), mean absolute error (MAE) and root mean squared error (RMSE) for each fold of the cross validation was calculated and averaged across all iterations for comparisons. Figure 3 below displays the cross validation averaged errors for each random forest (RF) model that was considered. We can see that all three errors are extremely similar between the tuned and default random forest models for each target. This tells us that tuning *mtry* is not adding much prediction power to the random forest models; the default value of *mtry* is sufficient.

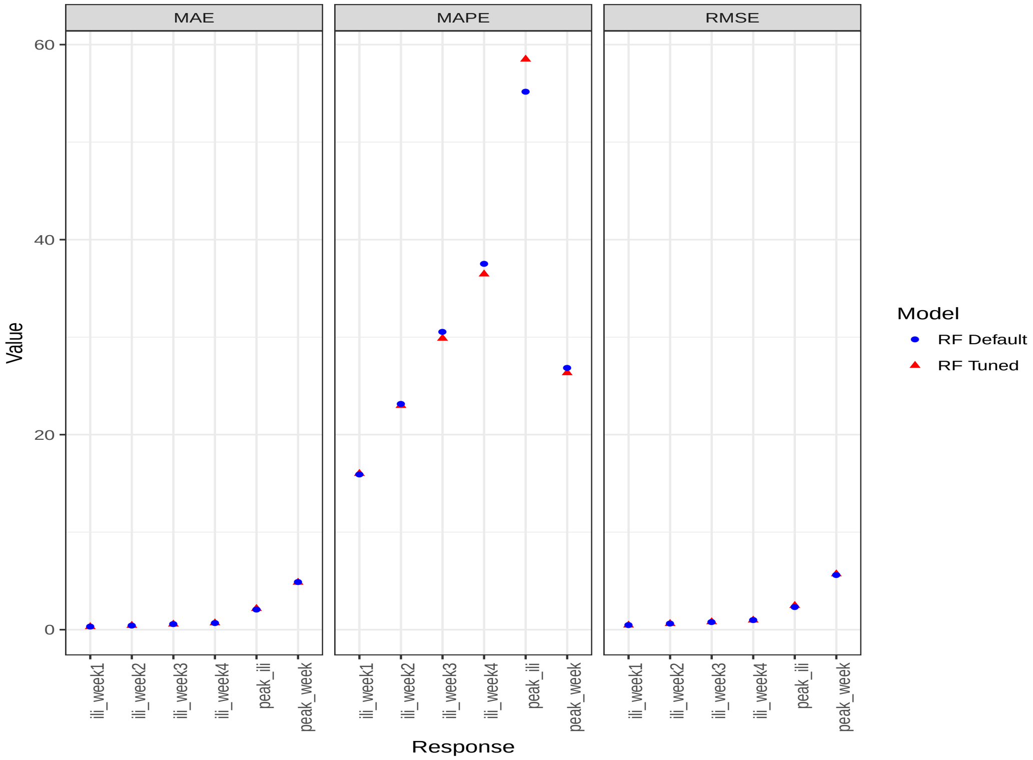

Figure 3: Leave-one-season-out cross validation average errors for random RF models

3.3 2016-17 Predictions

The last step in validating a prediction model is examining how well the random forest models predict on the season that was left out of the cross validation steps, the 2016-17 flu season. Predictions with the most recent flu season give us an adequate indication of how well the model will predict on the 2017-18 season. Figure 4 below shows the MAE, MAPE, and RMSE for the predictions of each model. Once again, we can see that all three errors are extremely similar between the tuned and default random forest models for each target. The predictions produced from each model can be seen in figures 5 through 10 .

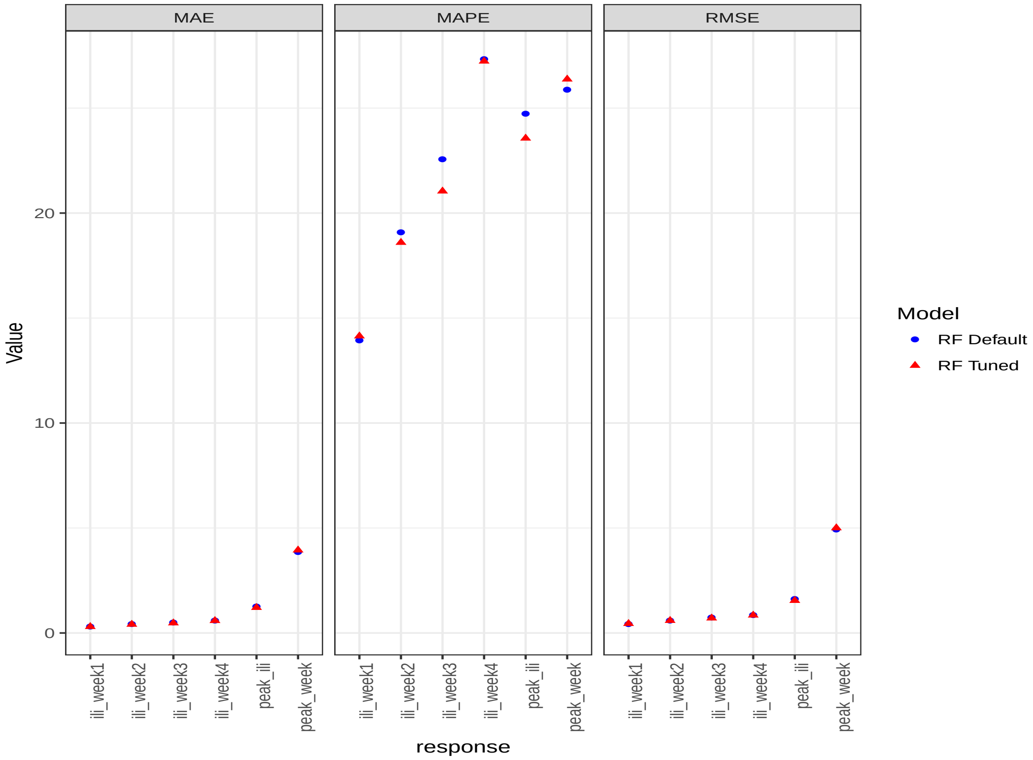

Figure 4: Random forest prediction errors for the 2016-17 flu season

For all 6 targets, the default and tuned random forest models return similar predictions. This is a direct reflection of the similarity in cross validation errors previously displayed in figure 3. Not surprisingly, the best predictions appear with the percent ILI for one week ahead and the worst for four weeks ahead. Observing less variability in prediction error with more immediate predictions is consistent with traditional time series models. We also notice more variability and and uncertainty in the peak week and peak ILI predictions which could be due to the nature of the predictions. For every week in the season, the same value is predicting using the same predictors as the week(s) ahead models. Creating predictors that capture information about the peak of a season could help with more accurate predictions. Further, the random forest models do not do the best with extreme regions for every target. For regions that experienced unusually low or unusually high influenza activity in the 2016-17 season, the predictions tend to over or under estimate what was actually observed. However, this is expected since the predictors in the model only considered information from the current and the prior flu season.

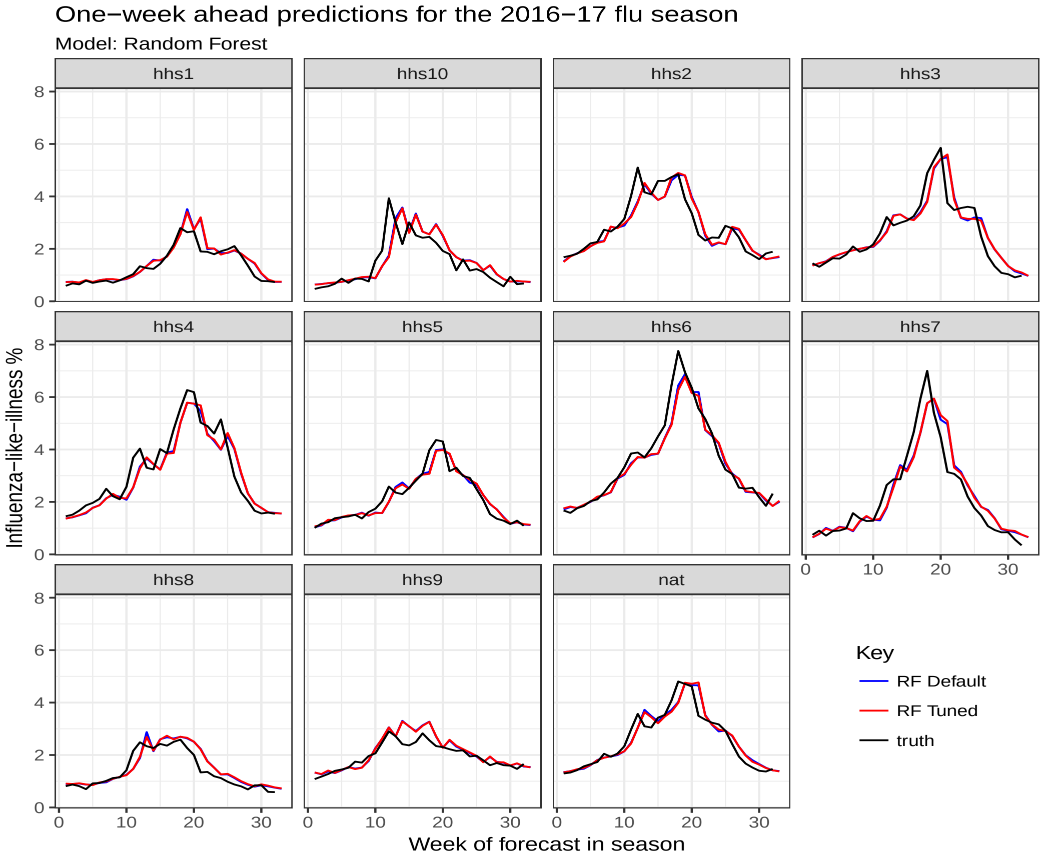

Figure 5: Random forest predictions for the percent ILI one week ahead from the date of forecast for the 2016-17 flu season

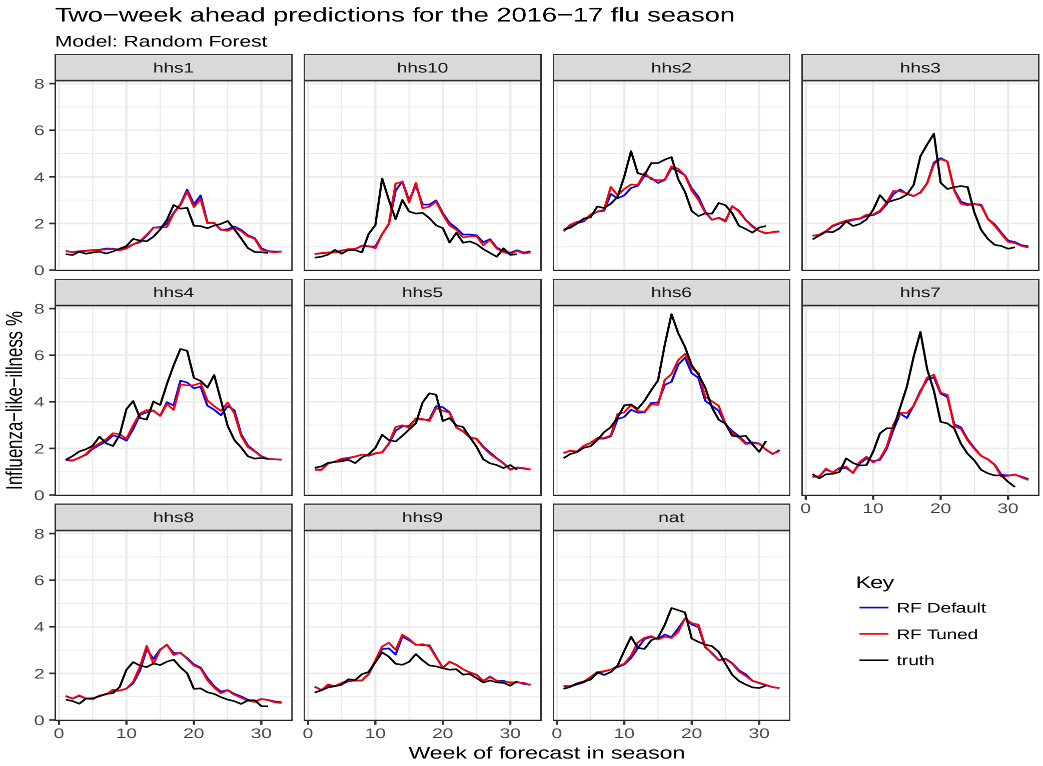

Figure 6: Random forest predictions for the percent ILI two weeks ahead from the date of forecast for the 2016-17 flu season

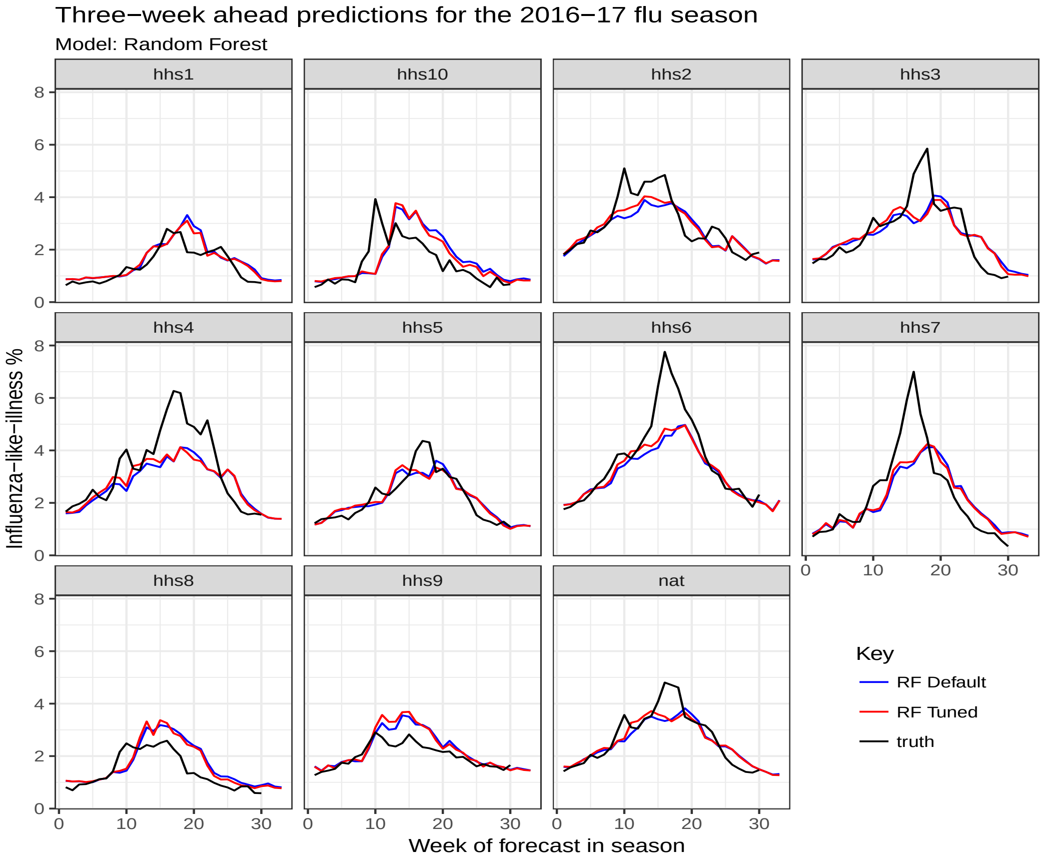

Figure 7: Random forest predictions for the percent ILI three weeks ahead from the date of forecast for the 2016-17 flu season

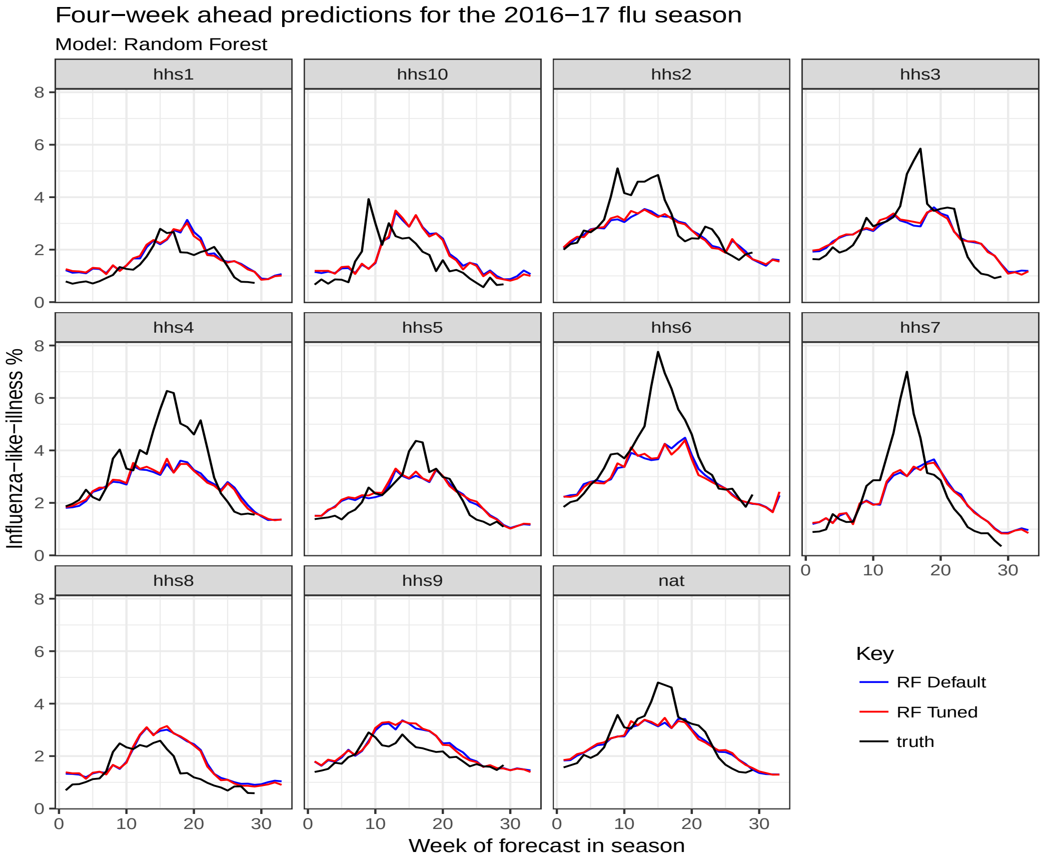

Figure 8: Random forest predictions for the percent ILI four weeks ahead from the date of forecast for the 2016-17 flu season

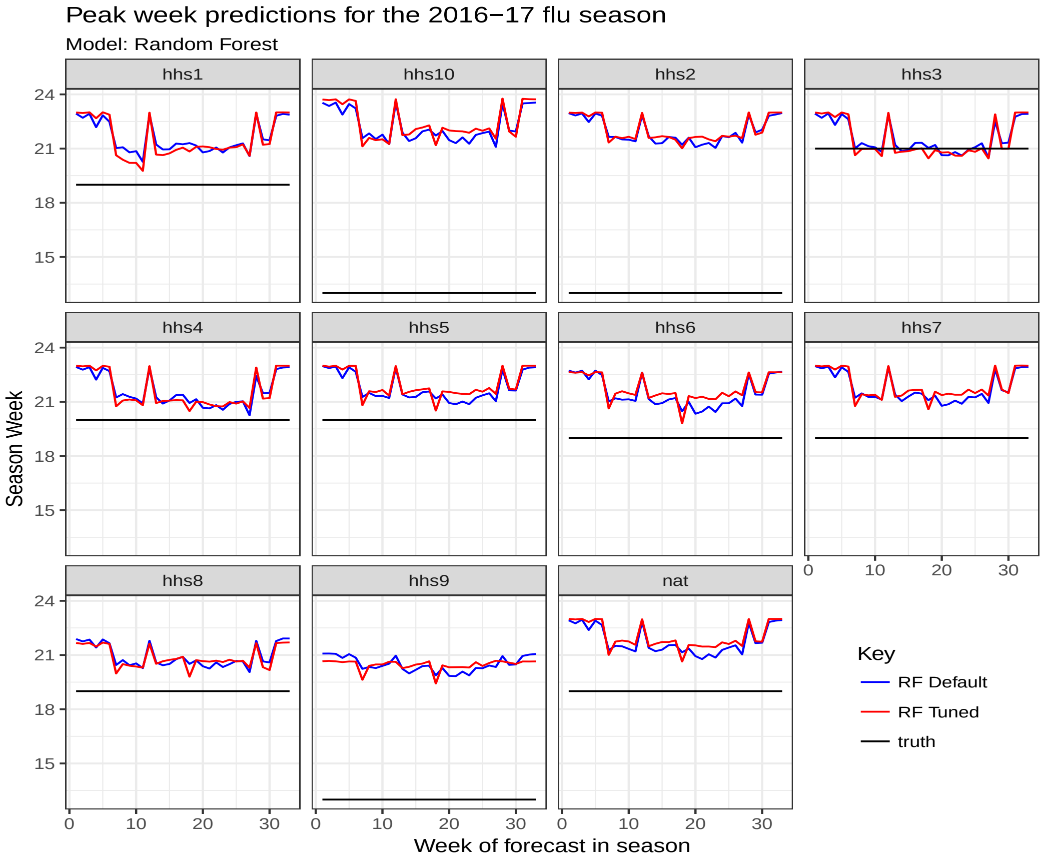

Figure 9: Random forest predictions for the peak week in the 2016-17 flu season

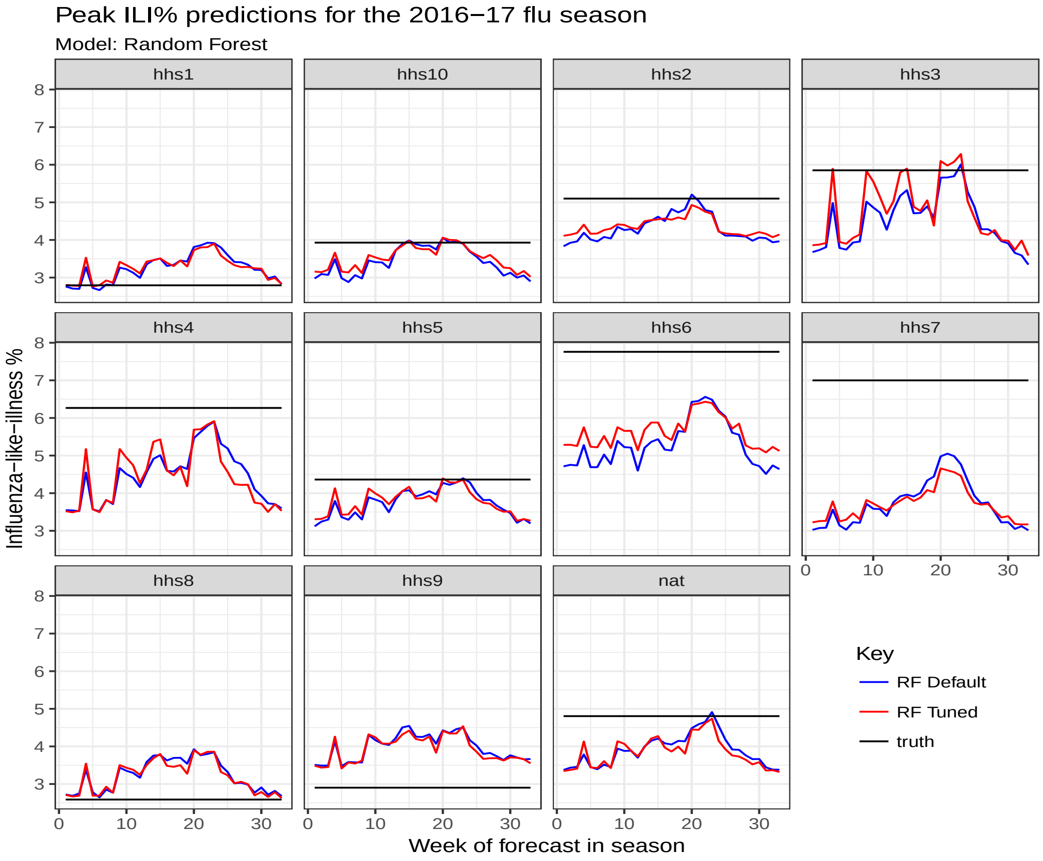

Figure 10: Random forest predictions for the peak percent ILI in the 2016-17 flu season

3.4 Other Models

Two other machine learning techniques were examined in see if random forest predictions could be outperformed. Extreme gradient boosting and neural networks models were constructed and validated for each of the 6 targets in a similar process as the random forest models. The figures in the Appendix show the cross validation error, prediction error and predictions for each of these techniques.

The performance of the extreme gradient boosting models is extremely similar to random forest when predicting influenza activity for the 2016-17 flu season. However, on average the extreme gradient boosting models did better at predicting three and four week ahead as well as peak week and peak percent ILI. The models still struggled to predict the extreme observations for the three and four week ahead targets; they consistently under-predicted regions that experienced unusually high peaks. The neural net models performed worse than both the random forest and extreme gradient boosting models, especially when predicting four weeks ahead. The underlying reason for this could be that the neural net models were built without considering region as a predictor. The reasoning behind this decision was that neural nets have troubles managing categorical predictors; more precise techniques, such as one-hot encoding, need to be implemented when building a neural network model with one (or multiple) categorical predictor. Overall, the cross validation and prediction errors from the extreme gradient boosting and neural network models are comparable to those from the random forest models.

4 Assessing Backfill

The healthcare providers report to the CDC on a voluntary basis and can change/update their data at anytime throughout the season, as stated in section 2.1. This results in the issue of backfill; a constant update of the weekly ILI estimates throughout the season. This is an issue for prediction modeling because the percent of outpatient visits experiencing influenza-like illness (ILI) reported during the forecasting week (within season estimates) differs from the percent of outpatient visits experiencing influenza-like illness (ILI) for the forecasting week reported at the end of the season (end of season estimates). Therefore, the prediction model that is predicting the end of the season estimates cannot be updated each week of the current flu season to include new information. To address the severity of this issue, default random forest models were built on flu seasons 2010-11 to 2015-16 for the one, two, three, and four week ahead targets that were created using the within season percent ILI estimated instead of the end of season estimates. Figure 11 below displays the leave-one-season-out cross validation average errors for each of these models. These models are comparable to their respective models shown in figure 3 when examining cross validation errors. The backfill observed throughout a flu season does not seem to have a significant impact on the models built to predict season end estimates for a flu season. Instead of building a prediction model using the end of season estimates, a using the estimates observed within a season could be built for prediction. This model could be updated weekly throughout a forecasting season to include the newly observed ILI estimates.

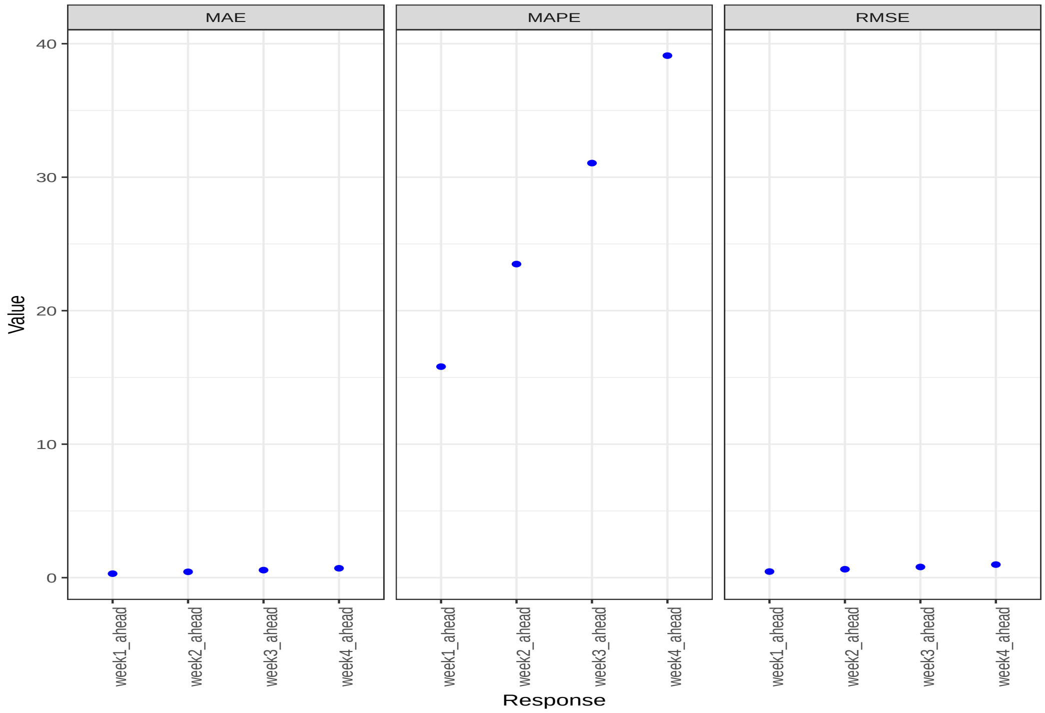

Figure 11: Leave-one-season-out cross validation average errors for models built on within season percent ILI estimates

Predictions were also made for the 2016-17 flu season with models built on flu seasons 2010-11 to 2015-16 for the one, two, three, and four week ahead targets created using within season percent ILI estimates instead of end of season estimates. Figure 12 shows the errors between those predictions and the end of season percent ILI estimates. We see that the errors produced are comparable to those calculated with the models considered in section 3.3 shown in figure 4. Again, this indicates that backfill is not having a significant impact when it comes to predicting the end of season percent ILI estimates.

Figures 13 through 16 show the predictions for the respective targets and regions, along with the within season and end of season estimates.

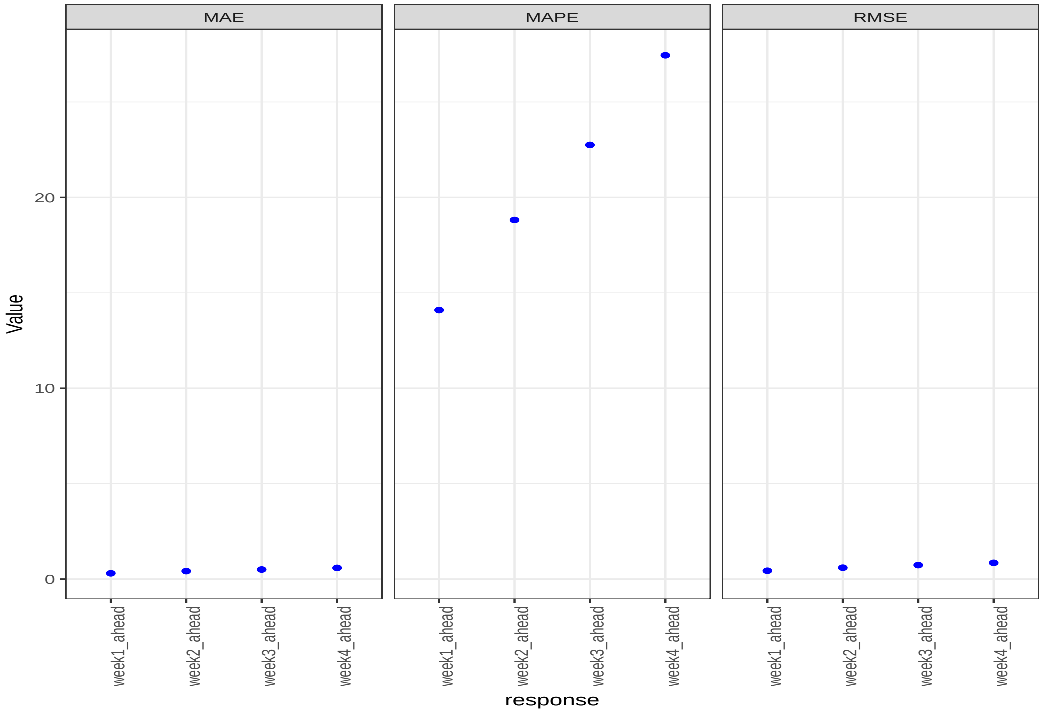

Figure 12: Random forest backfill prediction error for the 2016-17 flu season

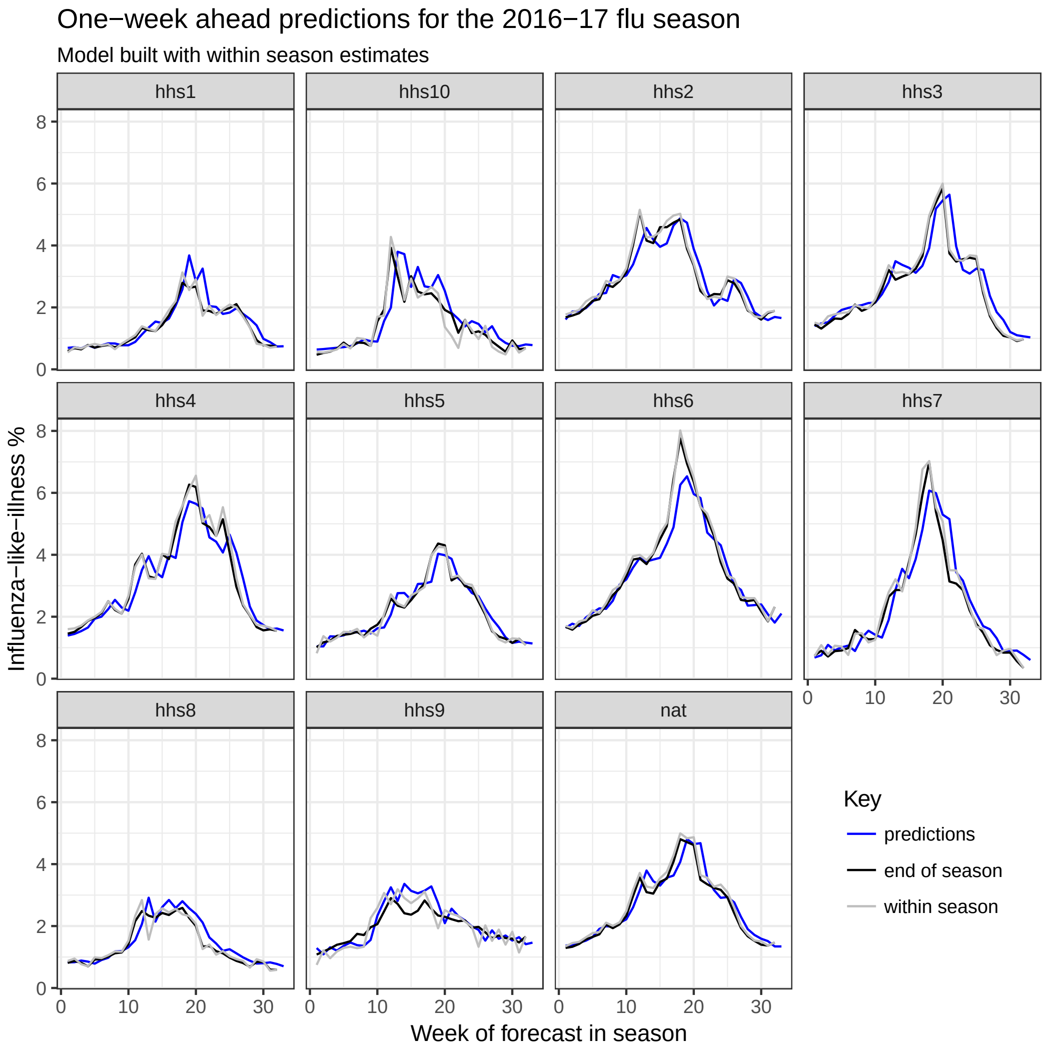

Figure 13: Predictions for the percent of outpatient visits experiencing influenza-like illness (ILI) one week ahead from the date of forecast with model built with within season estimates

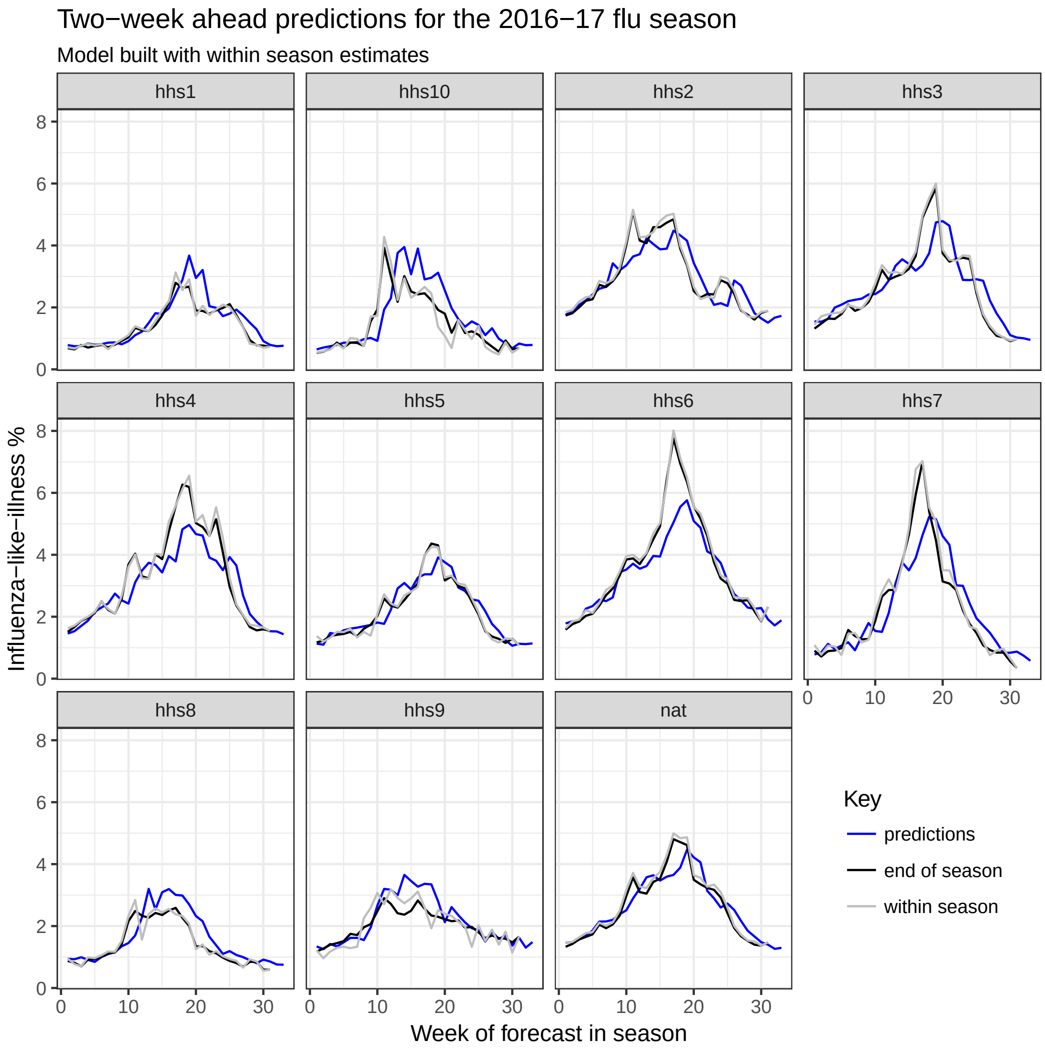

Figure 14: Predictions for the percent of outpatient visits experiencing influenza-like illness (ILI) two weeks ahead from the date of forecast with model built with within season estimates

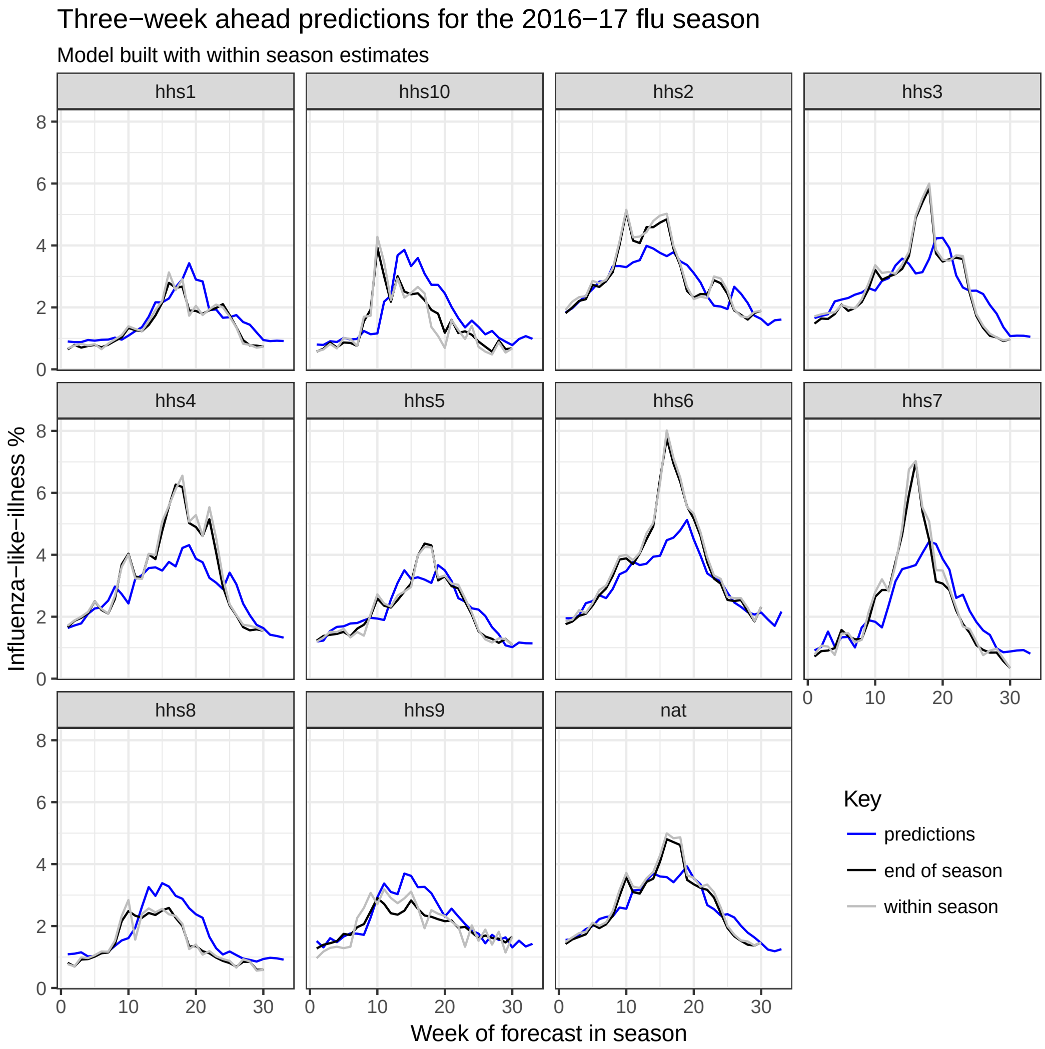

Figure 15: Predictions for the percent of outpatient visits experiencing influenza-like illness (ILI) three weeks ahead from the date of forecast with model built with within season estimates

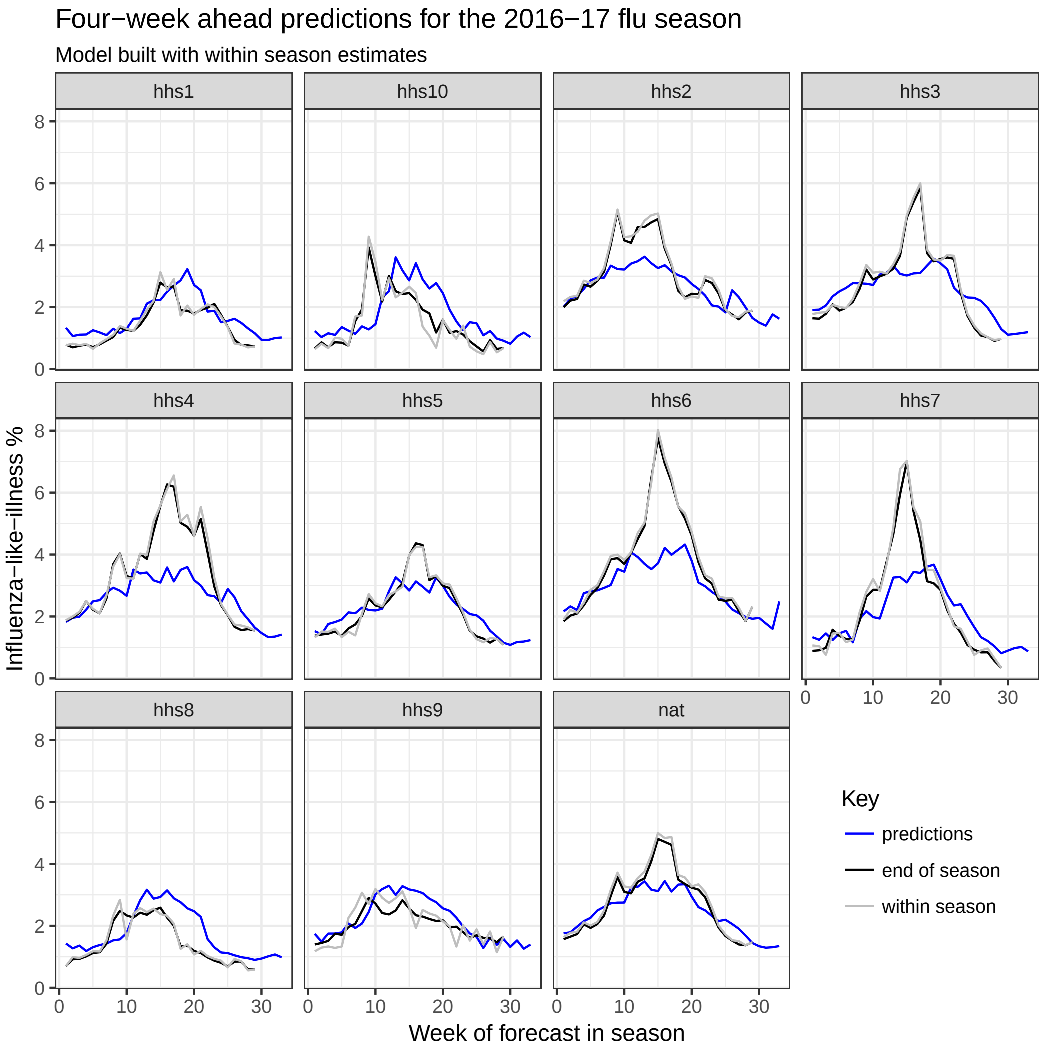

Figure 16: Predictions for the percent of outpatient visits experiencing influenza-like illness (ILI) four weeks ahead from the date of forecast with model built with within season estimates

5 Discussion

Overall, random forest as well as extreme gradient boosting and neural net techniques show great promise with prediction of the 2017-18 flu season activity, especially when predicting percent ILI one and two weeks ahead. Not only are the machine learning techniques powerful, but building prediction models that utilize the information collected by Wikipedia and information about past influenza activity reinforce the prediction power. The most important variable in predicting the week-ahead targets turned out to be the percent of outpatient visits experiencing influenza-like illness (ILI) for the forecasting week. This is expected since this estimate is very similar to the end of season weekly estimates that were used to build the prediction models. Other specific variables that have great importance in the models for those targets are the forecasting season week, region, and the number of Wikipedia page visits for the influenza A virus, shivering or chills. For the peak intensity and peak week targets, some variables of high importance are region, peak intensity of the previous flu season, and peak week of the previous flu season.

Though the current models show substantial promise, many things could be done to improve these models further. The most discrepancy in predictions happens when predicting peak week and peak percent ILI. This could be due to the nature of the target as well as the structure of the predictors. The predictors in the model were catered toward the one, two, three and four week ahead targets. One improvement could be to create more predictors that would capture information about the trend of the season and whether or not the peak of the season has already passed. All models discussed also had a difficult time predicting extreme influenza activity. Adding more outside variables like whether, influenza strain specific information, or vaccination information may help capture unusually high or low influenza activity experienced within a region. Lastly, predictions could be improved if information from more seasons were considered in the construction of the models.

6 Appendix

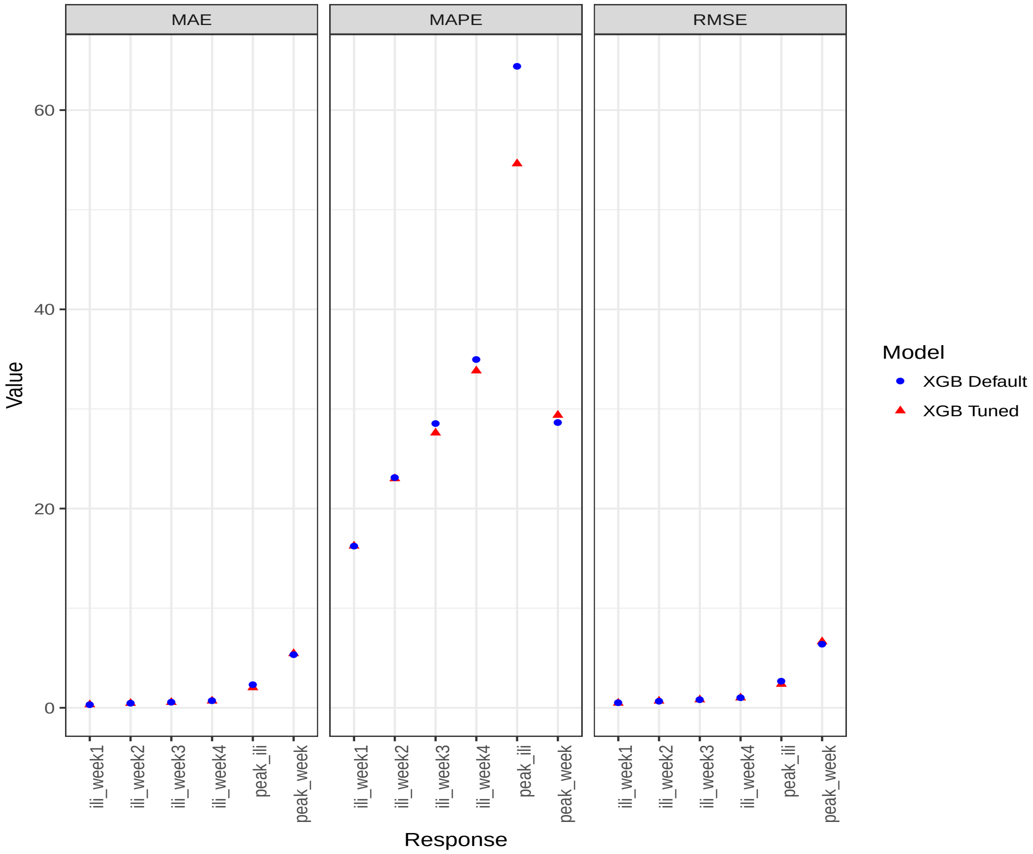

Figure 17: Leave-one-season-out cross validation average errors for XGB models

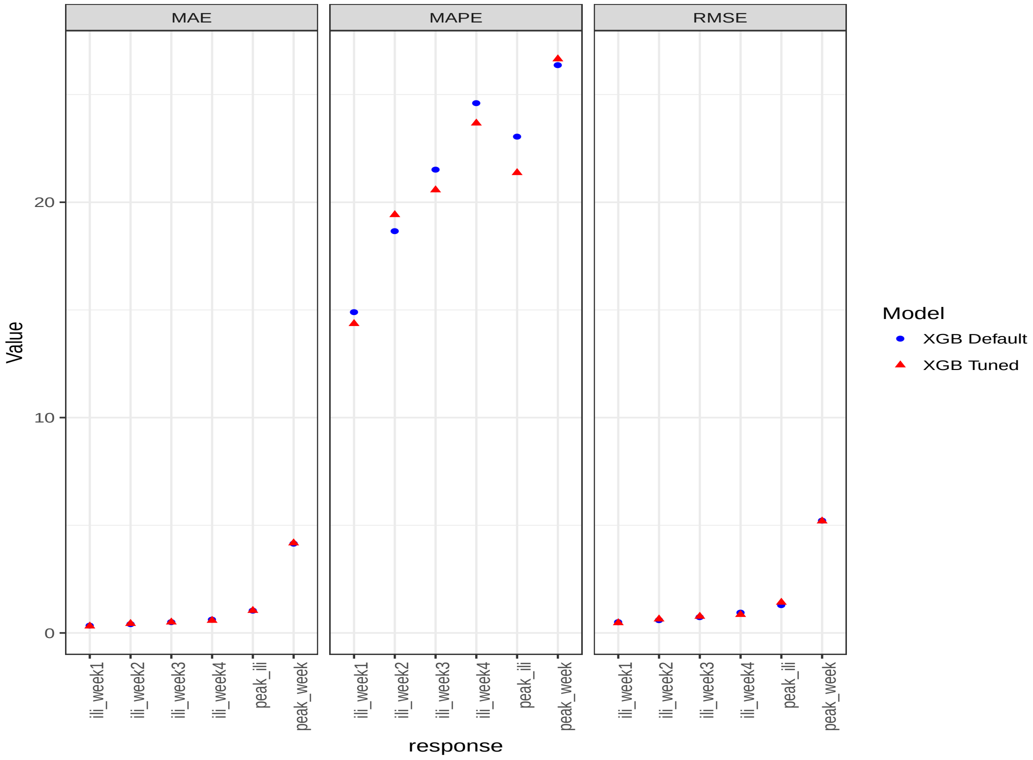

Figure 18: Extreme gradient boosting prediction errors for the 2016-17 flu season

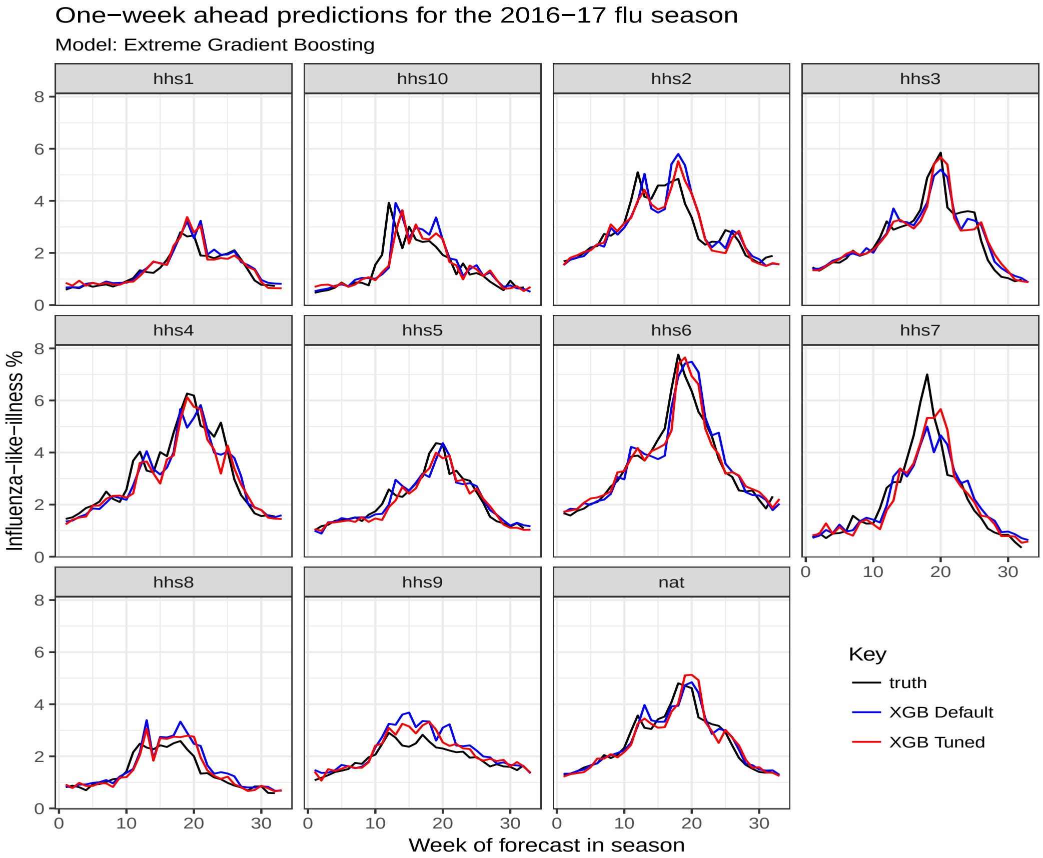

Figure 19: Extreme gradient boosting predictions for the percent ILI one week ahead from the date of forecast for the 2016-17 flu season

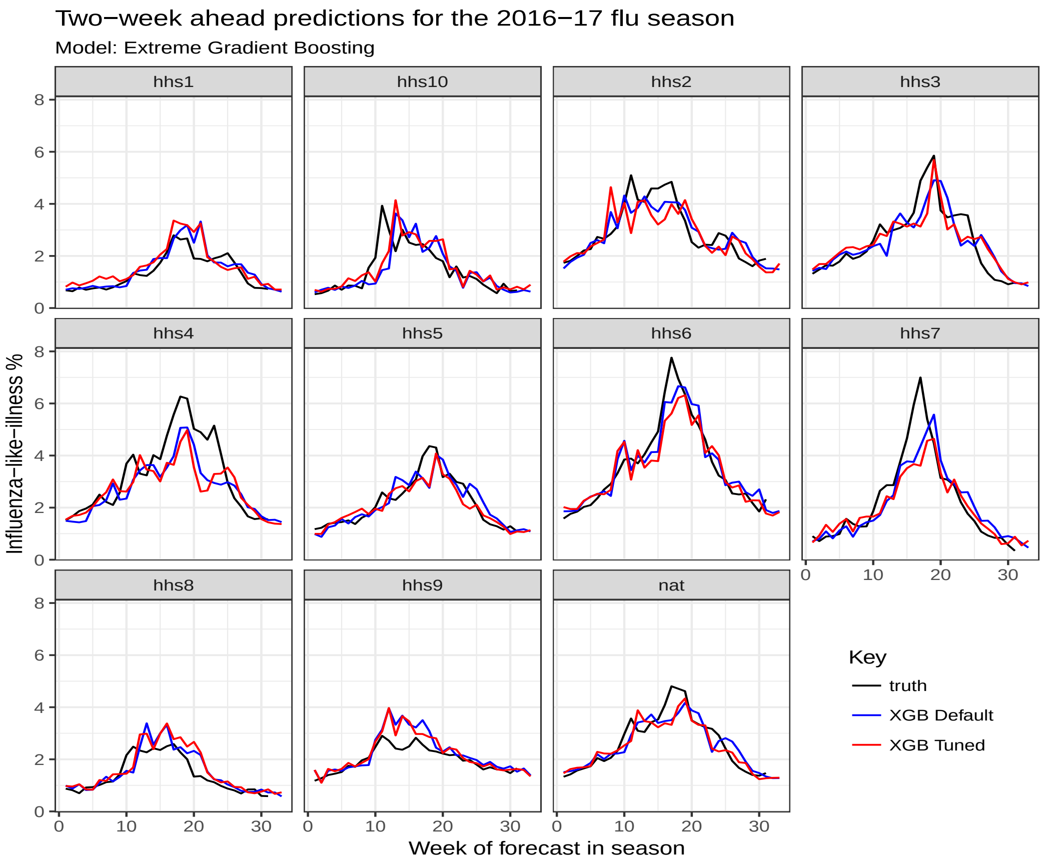

Figure 20: Extreme gradient boosting predictions for the percent ILI two weeks ahead from the date of forecast for the 2016-17 flu season

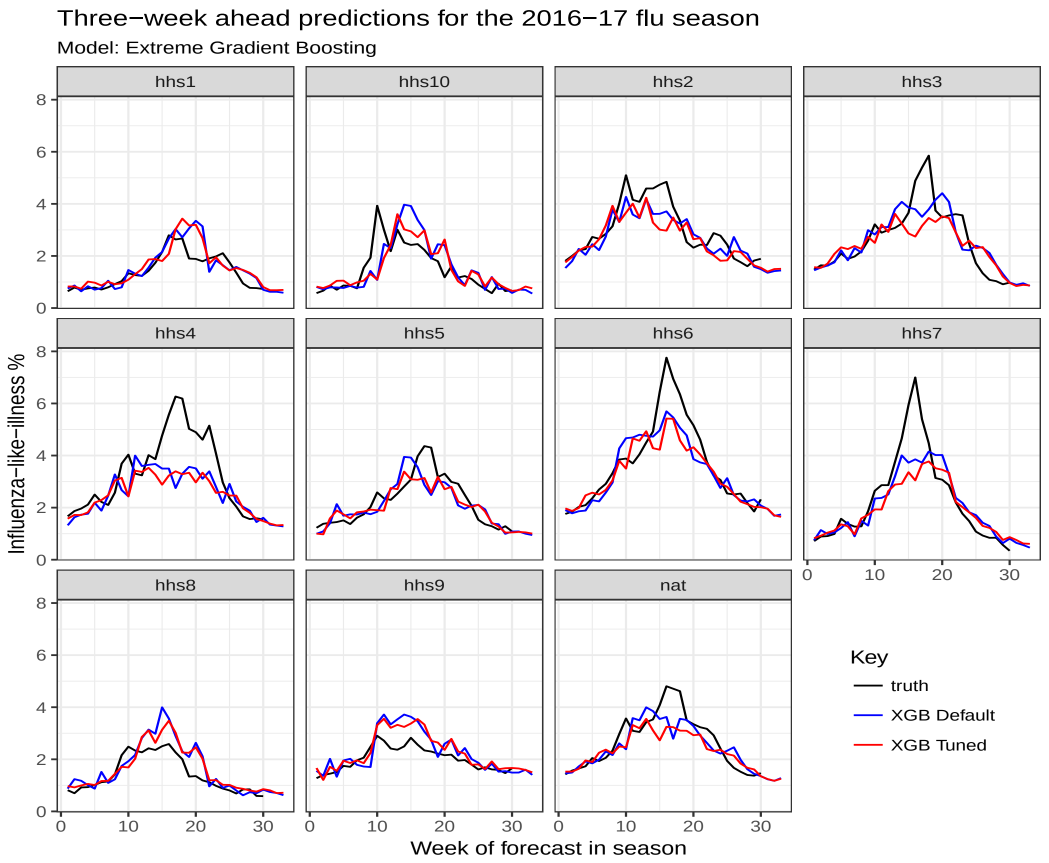

Figure 21: Extreme gradient boosting predictions for the percent ILI three weeks ahead from the date of forecast for the 2016-17 flu season

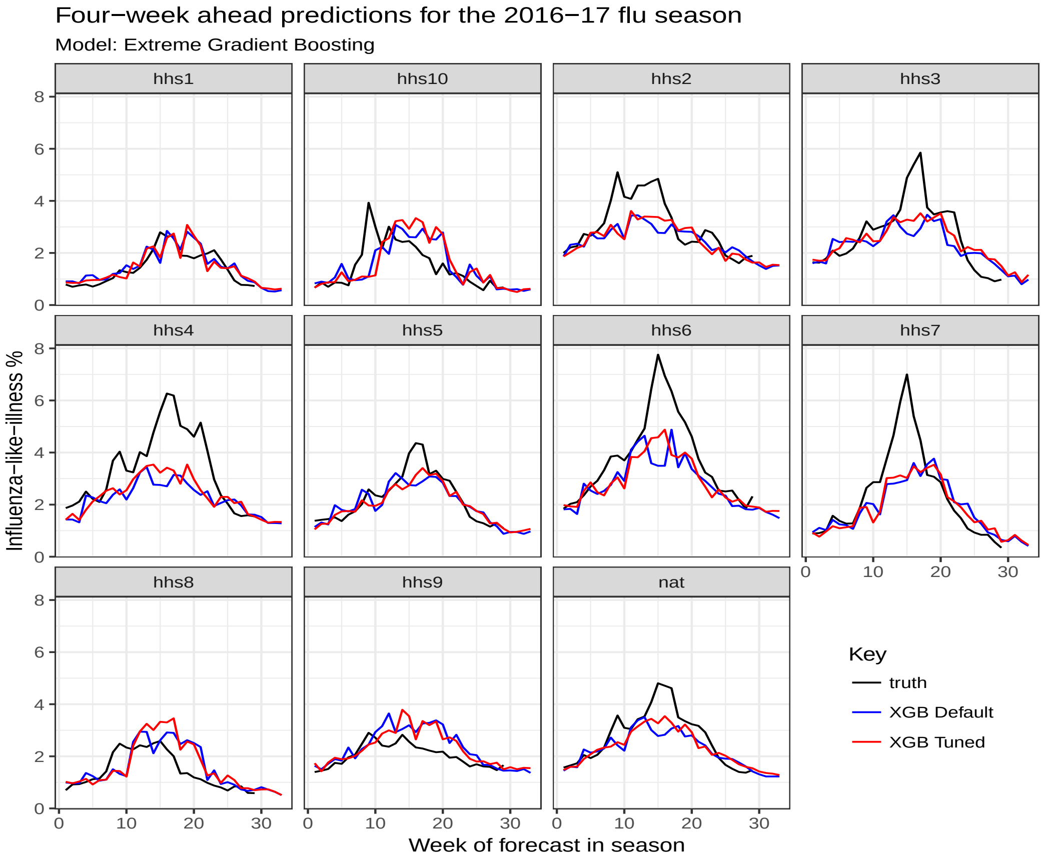

Figure 22: Extreme gradient boosting predictions for the percent ILI four weeks ahead from the date of forecast for the 2016-17 flu season

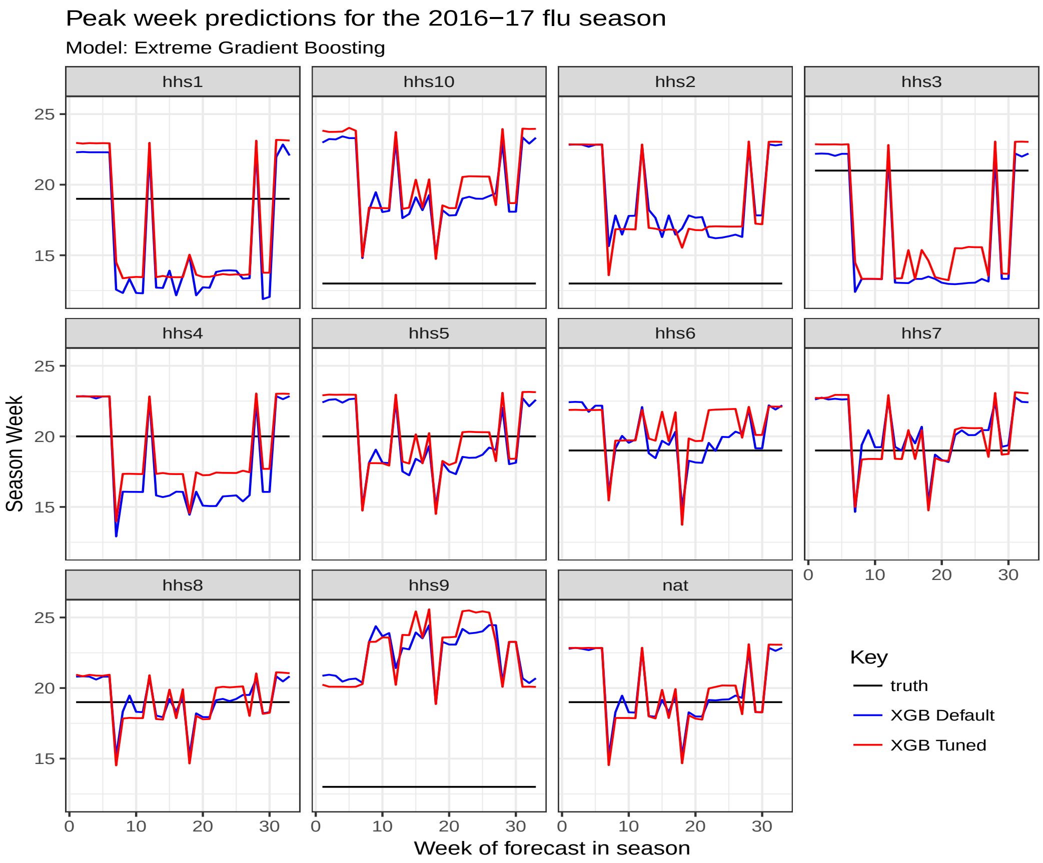

Figure 23: Extreme gradient boosting predictions for the peak week in the 2016-17 flu season

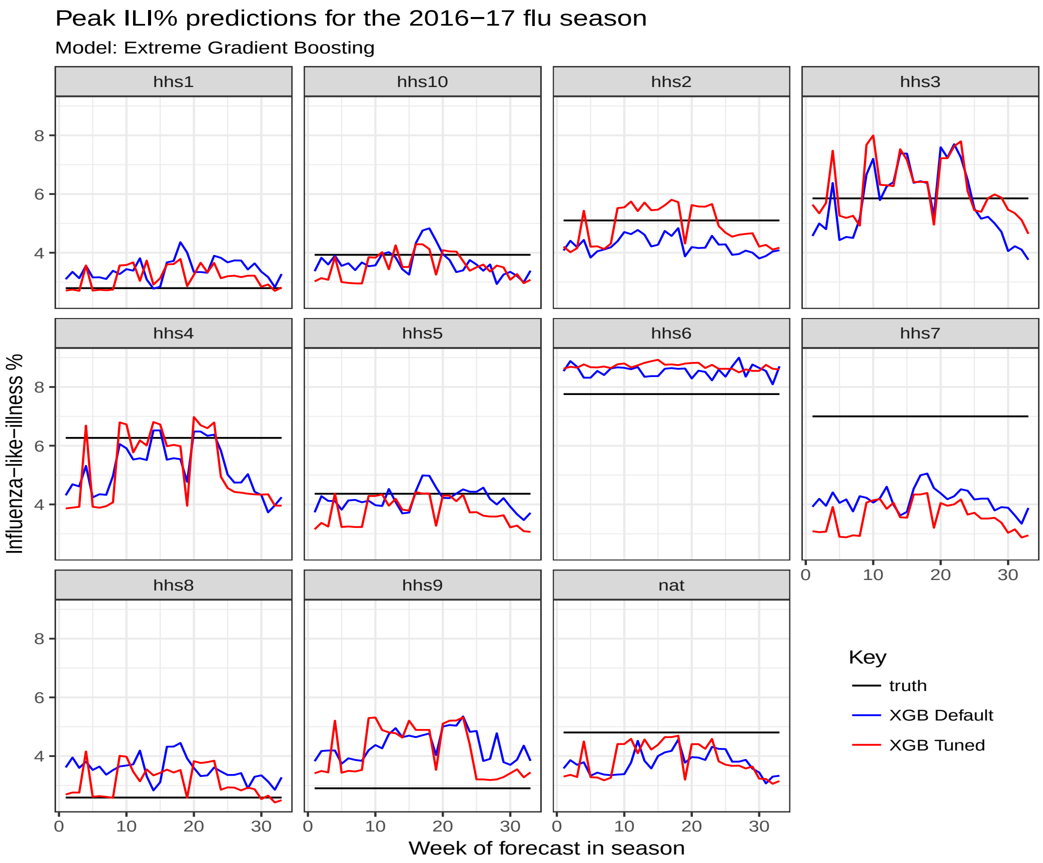

Figure 24: Extreme gradient boosting predictions for the peak percent ILI in the 2016-17 flu season

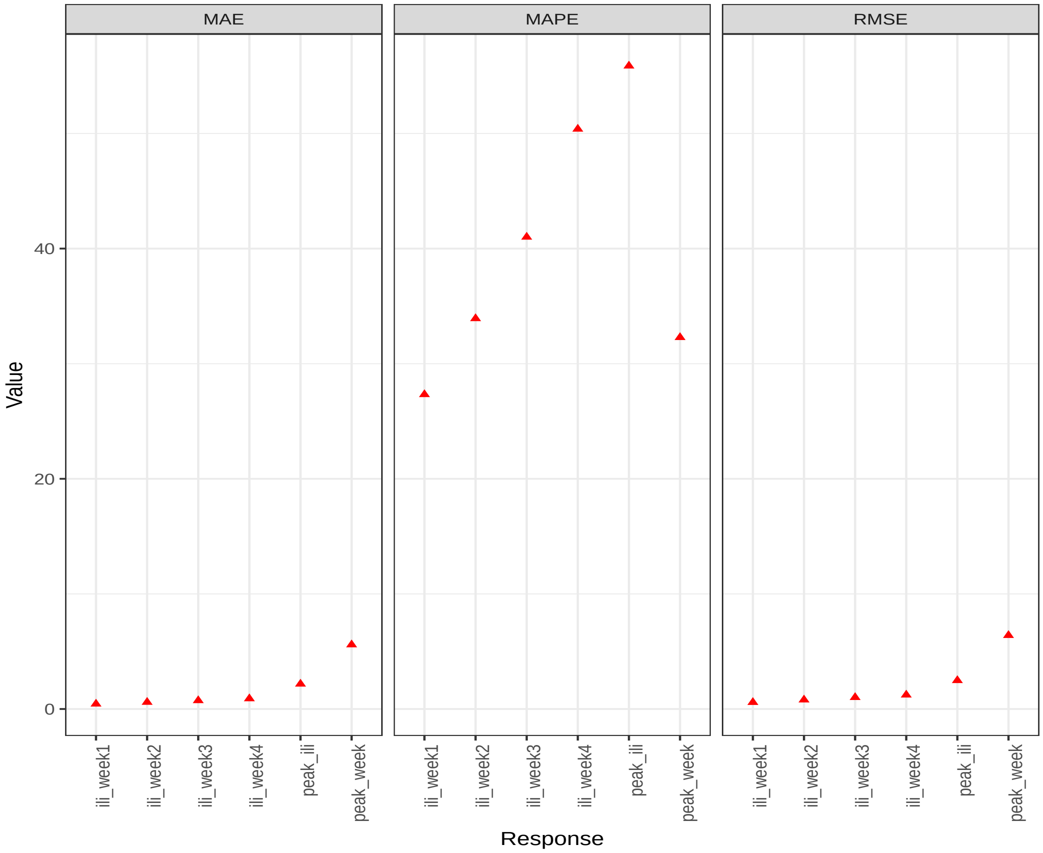

Figure 25: Leave-one-season-out cross validation average errors for NN models

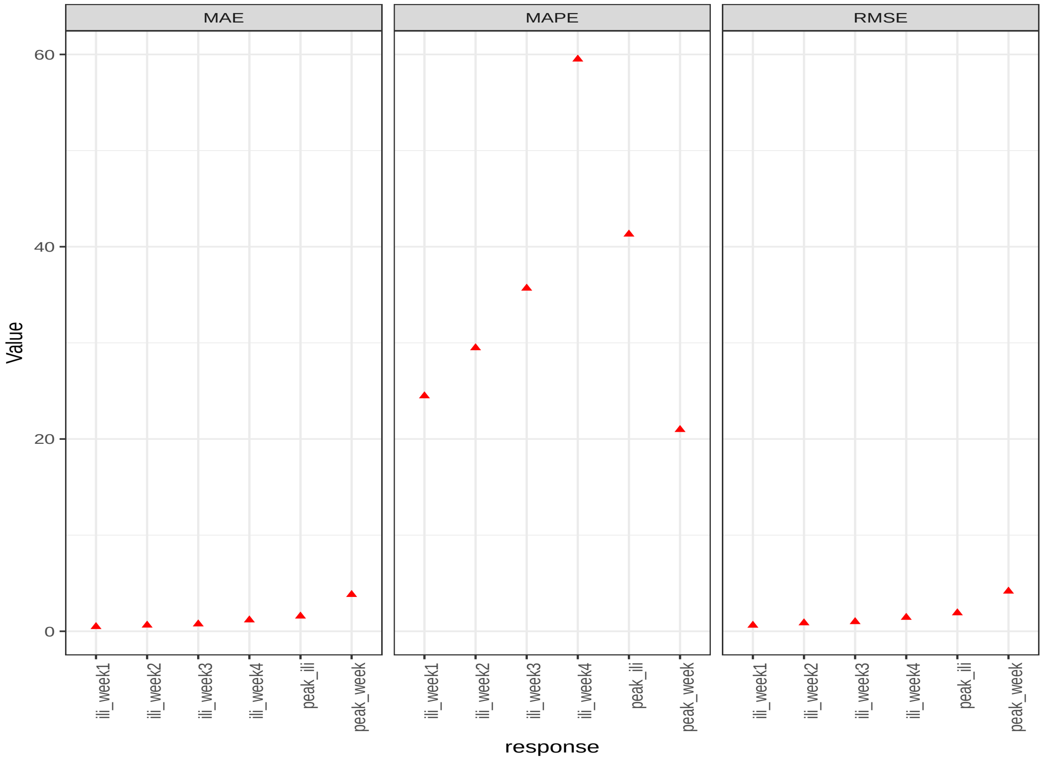

Figure 26: Neural net prediction errors for the 2016-17 flu season

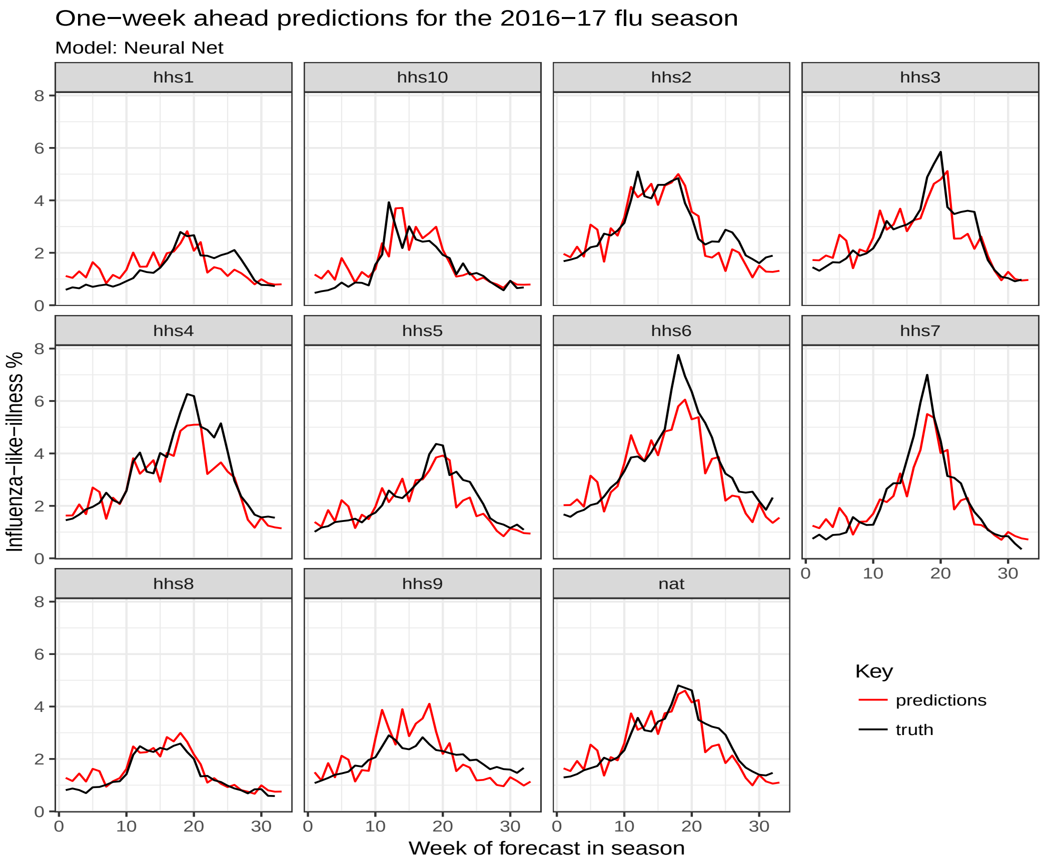

Figure 27: Neural net predictions for the percent ILI one week ahead from the date of forecast for the 2016-17 flu season

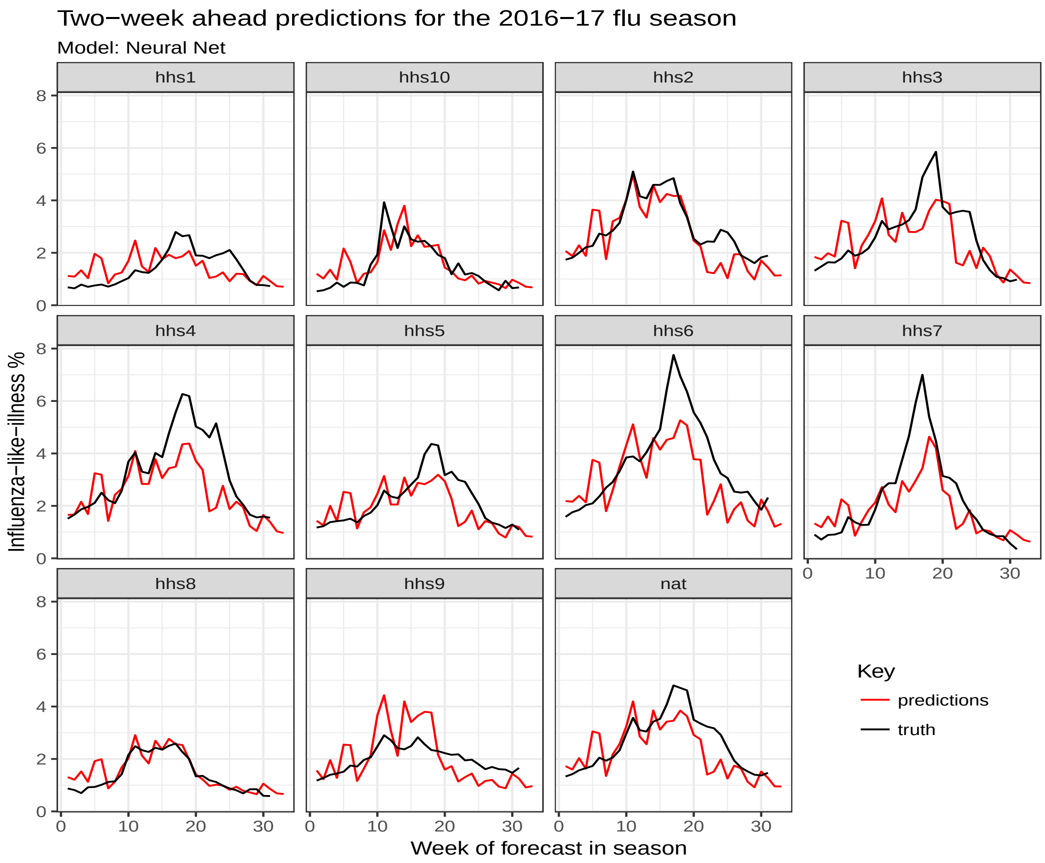

Figure 28: Neural net predictions for the percent ILI two weeks ahead from the date of forecast for the 2016-17 flu season

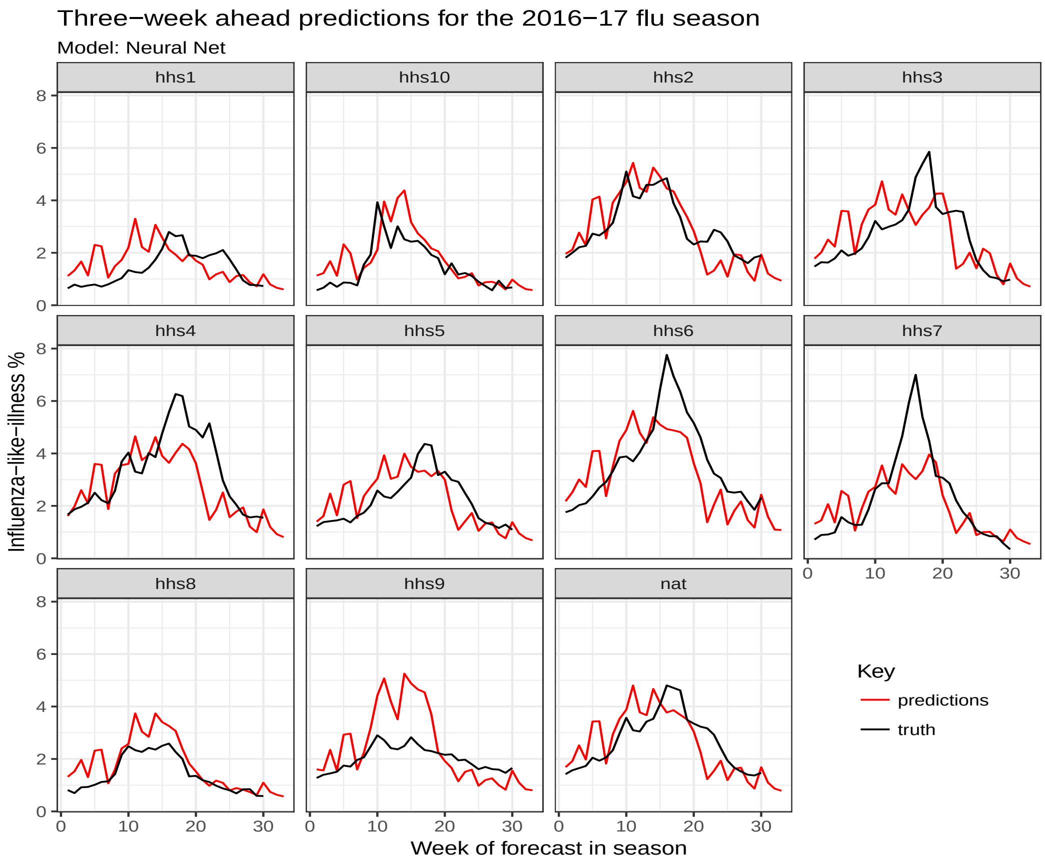

Figure 29: Neural net predictions for the percent ILI three weeks ahead from the date of forecast for the 2016-17 flu season

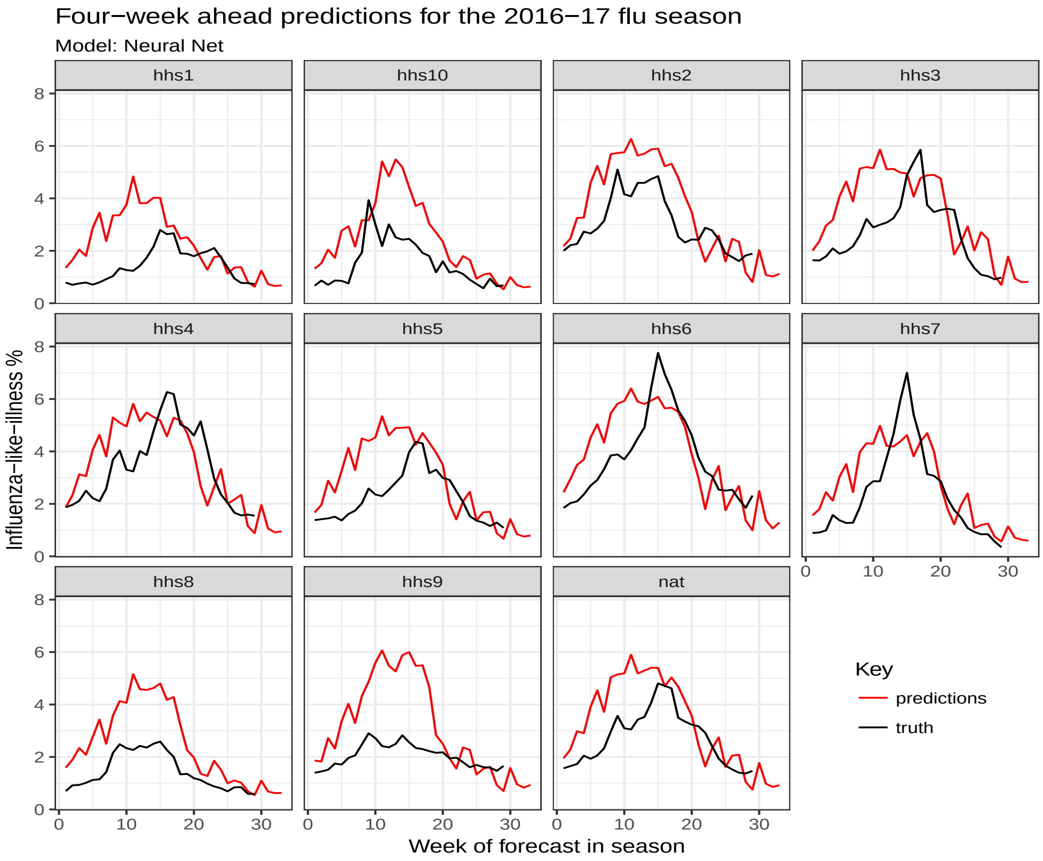

Figure 30: Neural net predictions for the percent ILI four weeks ahead from the date of forecast for the 2016-17 flu season

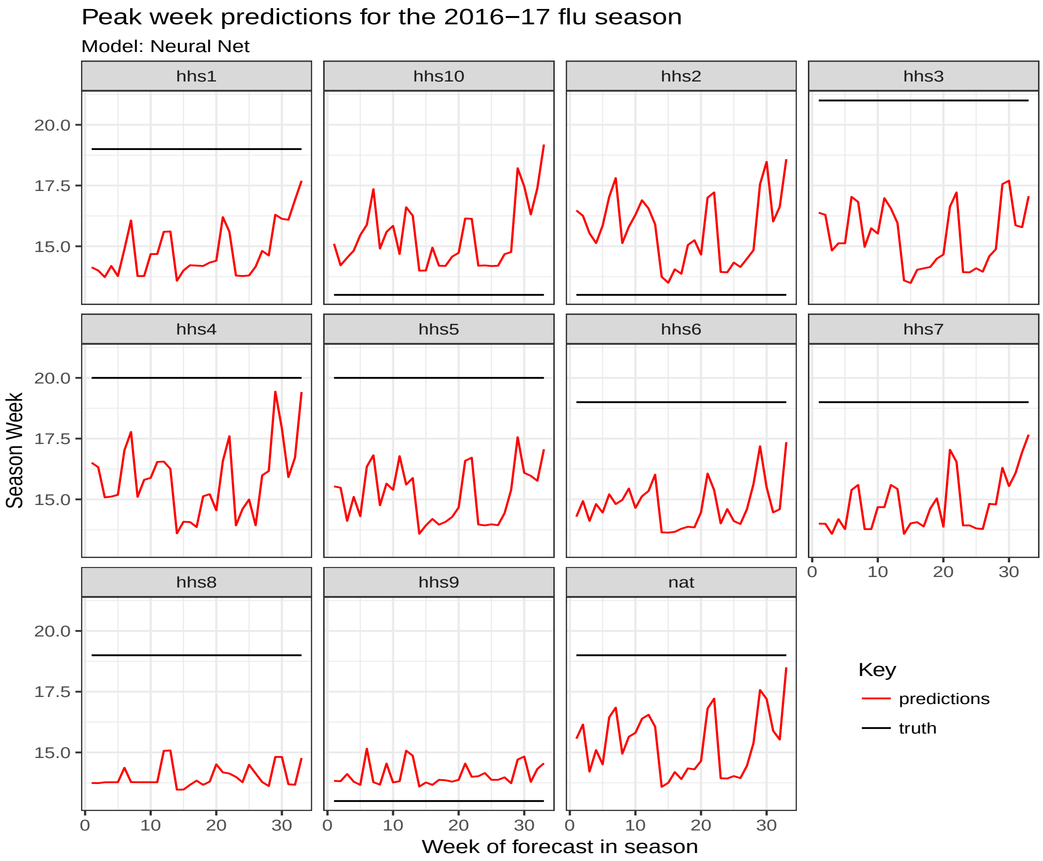

Figure 31: Neural net predictions for the peak week in the 2016-17 flu season

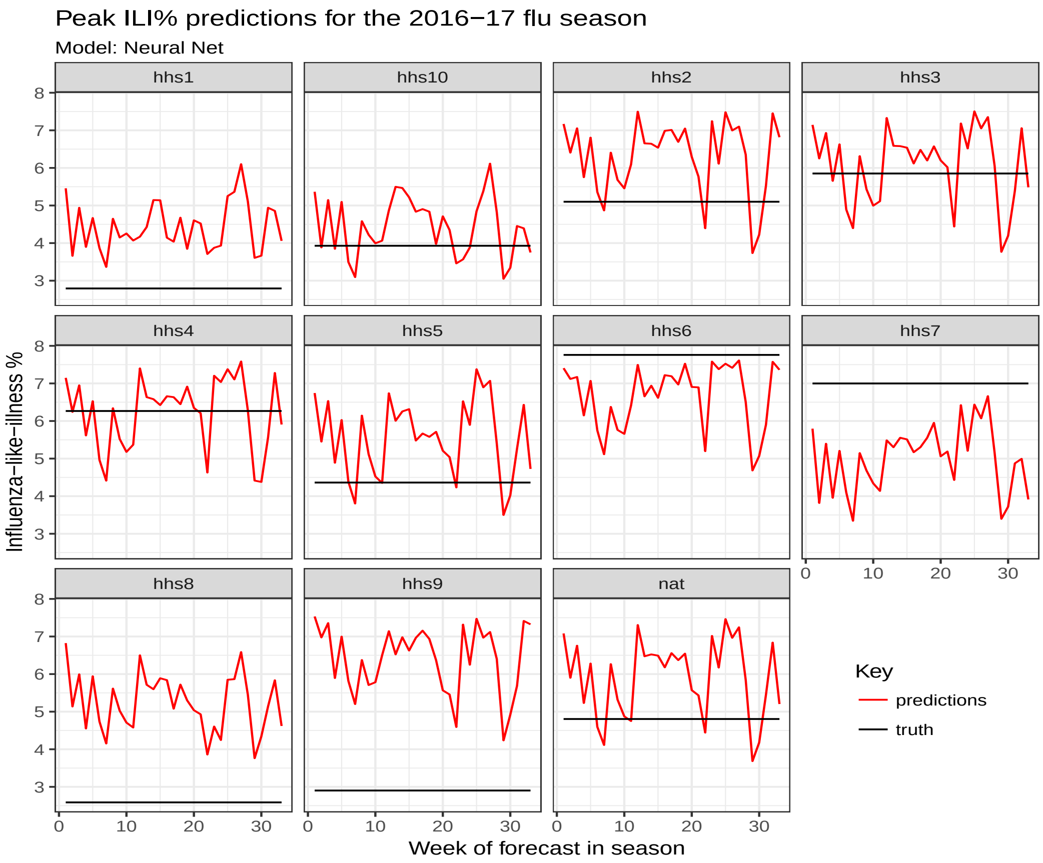

Figure 32: Neural net predictions for the peak percent ILI in the 2016-17 flu season

References

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] Overview of influenza surveillance in the united states. Last accessed on November 24, 2017.

[3] Mmwr weeks. Last accessed on November 25, 2017.

[4] Carnegie Mellon University DELPHI. Delphi epidata, 2017.

[5] Mark Dredze, Renyuan Cheng, Michael J Paul, and David Broniatowski. Healthtweets. org: a platform for public health surveillance using twitter. In *AAAI Workshop on the World Wide Web and Public Health Intelligence*, pages 593–596, 2014.

[6] Flusight: Seasonal influenza forecasting. Last accessed on November 28, 2017.

[7] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[8] Rolfes MA, Garg S Foppa IM, Flannery B, Brammer L, and Singleton JA. Estimated influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination in the united states, 2016. Last accessed on November 28, 2017.