Project:P1330White **PI:**Alice White
Prepared By: David Weitzenkamp & Caroline Ledbetter
Date: 11/29/2018

Introduction

Methods

Data

Data were collected from NORS 1998-2016 from NORS.

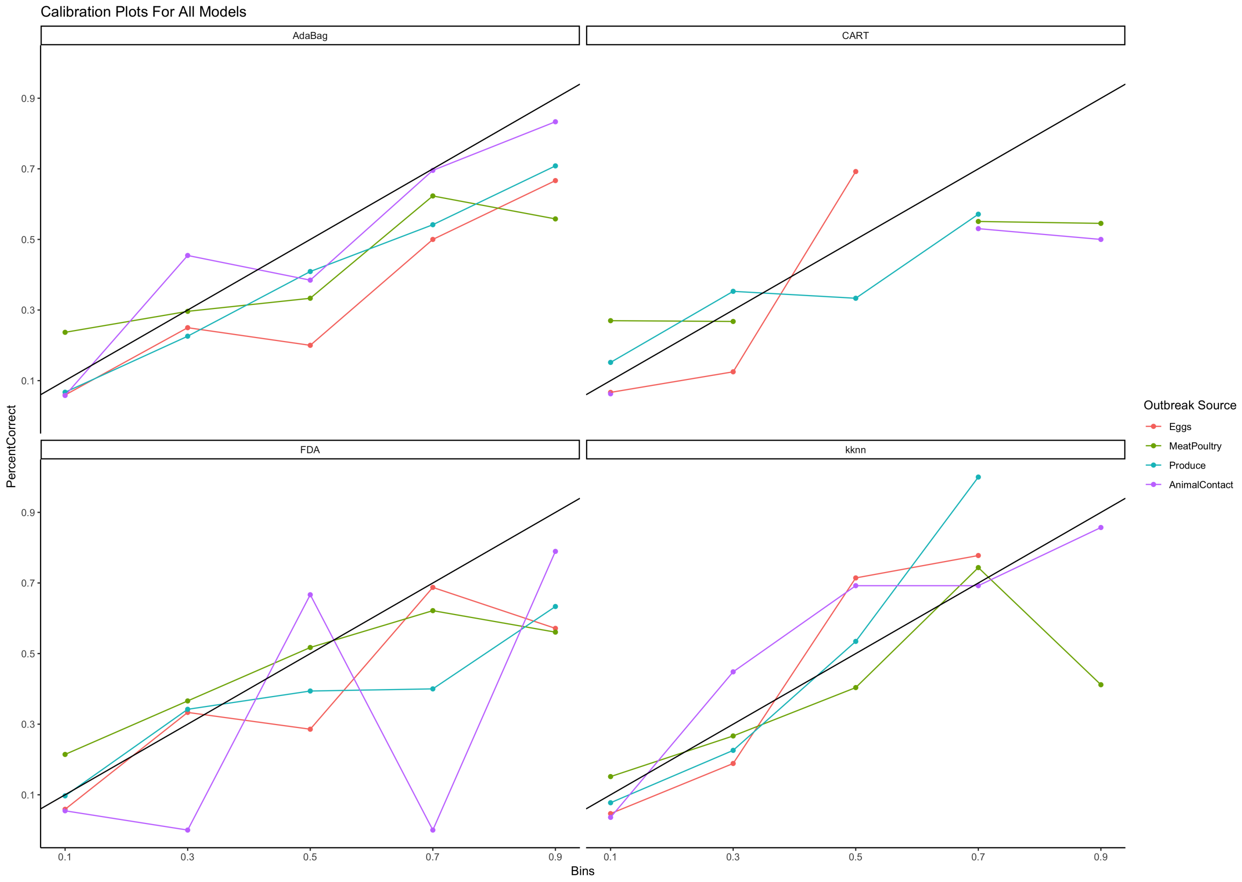
Model Selection

Foodborne outbreaks were grouped into categories based on there food source as identified in NORS. Outbreaks missing IFSAC information, those caused by multiple sources, unnclassifiable outbreaks, outbreaks of undetermined source and outbreaks from a source other than animal or plant were removed. Food sources that were rare (less than 100 outbreaks) were removed. We also included non foodbourne outbreaks caused by animal contact.

Data were split into a training set (75%) and a testing set (25%). Outbreaks from rare sources were left not removed until after the split so they could be left in the testing set. In order to more accurately reflect actual usage, foodbourne outbreaks of other origin were included in the testing set, outbreaks with multiple, unclassifiable and no identified food sources were not. The number of total cases, the season the outbreak started, the geography of the outbreak (multistate, single state - multicounty, single state - single county), the agent (STEC or Salmonella Serotype), the percentage of female and male cases, the percentage of people hospitalized, and the percentage of cases in each age group (Under 1 year, 1-4 yrs, 5-9 yrs, 10-19 yrs, 20-49 yrs, 50- 74 yrs, 75 yrs and older) were used as predictors. We selected four algorithmic methods for predicition based on their ability to predict mutliple class probabilities well - Adaptive bagging, classification and regression trees (CART), weighted k nearest neighbots (knn), and flexible discriminant analysis (FDA). The final model was chosen based on Brier Scores (a measure of the difference in the predicted probability and the actual event). In order to more accurately reflect actual usage, foodbourne outbreaks of other origin were included in the testing set, outbreaks with multiple, unclassifiable and no identified food sources were not. Parameter selection was performed using the Caret package.

Results

1261 outbreaks missing IFSAC information, 479 outbreaks caused by multiple sources, 51 unnclassifiable outbreaks, 145 outbreaks of undetermined source and 12 outbreaks from a source other than animal or plant were removed. 79 dairy, 19 fish, 9 game, 10 grains-beans, 18 nuts-seeds, 1 oils-sugars, and 35 Aquatic Animals outbreaks were removed.
All four models performed well, the Adaptive bagging model had a brier score of 0.145, CART of 0.146, weighted k nearest neighbors of 0.125 and FDA of 0.143. Calibration curves based on the testing data set are shown inn Fig 1.


Calibration Plots For All Models
AdaBag
CART
FDA
kknn
PercentCorrect
0.1
0.3
0.5
0.7
0.9
Bins
0.1
0.3
0.5
0.7
0.9
Outbreak Source
Eggs
MeatPoultry
Produce
AnimalContact