Project:P1330White **PI:**Alice White
Prepared By: David Weitzenkamp & Caroline Ledbetter
Date: 12/12/2018

Introduction

Methods

Data

Data were collected from NORS 1998-2016 from NORS.

Foodborne outbreaks were grouped into categories based on the 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. Non foodbourne outbreaks caused by animal contact were included.

Data were split into a training set (75%) and a testing set (25%). Outbreaks from rare sources were 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.

Outbreaks which started between Jan and Mar were classified as winter, between Apr and Jun as spring, between Jul and Sept as summer and between Oct and Dec as fall.

Model Selection

We selected seven algorithmic methods for predicition based on their ability to predict mutliple class probabilities well - bagged adaptive boosting classification trees, classification and regression trees (CART), weighted k nearest neighbots (knn), flexible discriminant analysis (FDA), weighted subspace random forest, Naive Bayes, and rule-based classifier. A non-informative model that uses no information from predictors was also generated for comparison purposes. The final model was chosen based on Brier Scores (a measure of the difference in the predicted probability and the actual event). Outbreaks with other origins were included in the brier score calculations. All analysis was done in R version 3.5.1 (2018-07-02) with the Caret Package v(6.0.81. Parameter selection was performed using the Caret package.

	Eggs N = 155 N(%)	MeatPoultry N = 536 N(%)	Produce N = 282 N(%)	AnimalContact N = 187 N(%)	Other N = 215 N(%)
Season					
Winter	27(17)	68(13)	45(16)	65(35)	41(19)
Spring	37(24)	165(31)	88(31)	51(27)	54(25)
Summer	59(38)	186(35)	83(29)	43(23)	74(34)
Fall	32(21)	117(22)	66(23)	28(15)	46(21)
Agent					
Braenderup	1(1)	8(1)	7(2)	6(3)	1(0)
Enteritidis	103(66)	70(13)	20(7)	7(4)	19(9)

Heidelberg	7(5)	25(5)	1(0)	2(1)	6(3)
I 4,[5],12:i:-	0(0)	12(2)	2(1)	9(5)	4(2)
Javiana	0(0)	5(1)	14(5)	0(0)	3(1)
Montevideo	0(0)	6(1)	2(1)	9(5)	4(2)
Multiple Serotypes	2(1)	12(2)	9(3)	10(5)	9(4)
Newport	0(0)	25(5)	33(12)	0(0)	8(4)
Saintpaul	1(1)	4(1)	13(5)	1(1)	3(1)
Salm unk sero	6(4)	117(22)	53(19)	23(12)	38(18)
STEC	23(15)	121(23)	64(23)	48(26)	49(23)
Typhimurium	5(3)	46(9)	14(5)	28(15)	28(13)
Paratyphi B	0(0)	0(0)	1(0)	3(2)	4(2)
NonSpecific Sero group	1(1)	22(4)	20(7)	8(4)	9(4)
Primary Plant Sero group	0(0)	1(0)	14(5)	3(2)	3(1)
Primary Animal Sero group	5(3)	40(7)	3(1)	20(11)	8(4)
Rare	1(1)	22(4)	12(4)	10(5)	19(9)
Geography					
MultiState	4(3)	69(13)	131(46)	71(38)	40(19)
MultiCounty	12(8)	63(12)	47(17)	25(13)	38(18)
SingleCounty	137(88)	401(75)	100(35)	87(47)	136(63)
Missing	2(1)	3(1)	4(1)	4(2)	1(0)
	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)
% Male	39(28)	43(28)	33(21)	37(25)	43(26)
% Female	40(29)	43(29)	55(26)	50(29)	45(26)
% Sex Unknown	21(41)	13(34)	12(32)	13(32)	12(32)
% Age Under1	0.059(0.41)	0.32(2.37)	0.44(1.78)	7.3(12.17)	0.74(5.02)
% Age 1to4	0.63(3.2)	2.7(10.2)	2.5(8.9)	23(26.8)	6(15.5)
% Age 5to9	0.96(6.2)	2.4(9.2)	3(7.7)	12(19.4)	4.5(12.6)
% Age 10to19	2.3(9.3)	4.4(11.9)	5.8(11.9)	14(20.8)	6.5(14.8)
% Age 20to49	10(25)	16(27)	20(27)	16(20)	16(25)
% Age 50to74	5(13)	6.6(14)	10(17)	7.9(14)	6.7(16)
% Age 75plus	1.6(8.5)	1.7(7.2)	2.9(7.8)	2(6.6)	1.4(4.6)
% Age Unknown	58(49)	53(49)	44(49)	10(27)	46(49)
Proportion Hospitalized	0.16(0.21)	0.24(0.27)	0.22(0.20)	0.22(0.24)	0.24(0.24)
Missing(N%)	20(13)	65(12)	32(11)	9(5)	17(8)
	Median(IQR)	Median(IQR)	Median(IQR)	Median(IQR)	Median(IQR)
Total Cases	10(4-27)	10(4-22)	19(10-46)	8(3-31)	12(5-32)

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 classified as other.

Characterstics of outbreaks in the analysis are given in table 1. Egg outbreaks were more likely to tabke place in the summer, and animal contact outbreaks were more common in winter. Meat/Poultry and Produce outbreaks were both uncommon in the winter. Egg outbreaks were most likely to be cause by Salmonella Enteritidis and be single county. Egg outbreaks were also more likely to have unknown proportions of male

vs female and ages. Animal Contact outbreaks were more likely to have higher percentages of young children, with a mean percentage of 23% for 1 to 4 year olds. Produce outbreaks had higher percentages of people over 20 (20-49 yr olds, 50-74 yr olds, and those over 75) than any other outbreak source. Produce outbreaks also tended to be larger.

The testing data set consiseted of 11% egg, 39% meat-poultry, 21% produce, 13% animal contact, and 16% other outbreaks (table 2).

	Eggs	MeatPoultry	Produce	AnimalContact	Other
test	38	134	70	46	53
train	117	402	212	141	117

Model performance varied, two models(Naive Bayes and rule-based classifier) had a Brier Score worse than the non informative model. Weighted k nearest neighbors and weighted subspace random forest performed the best with Brier Scores of 0.125 and FDA of 0.127, respectively. Calibration curves based on the testing data set are shown in Fig 1. Weighted k nearest neighbors was chosen for the final model.

AdaBag	CART	kknn	FDA	NonInf	wsrf	NaiveBayes	PART
0.1505	0.1423	0.1246	0.1572	0.1562	0.1267	0.1585	0.1995

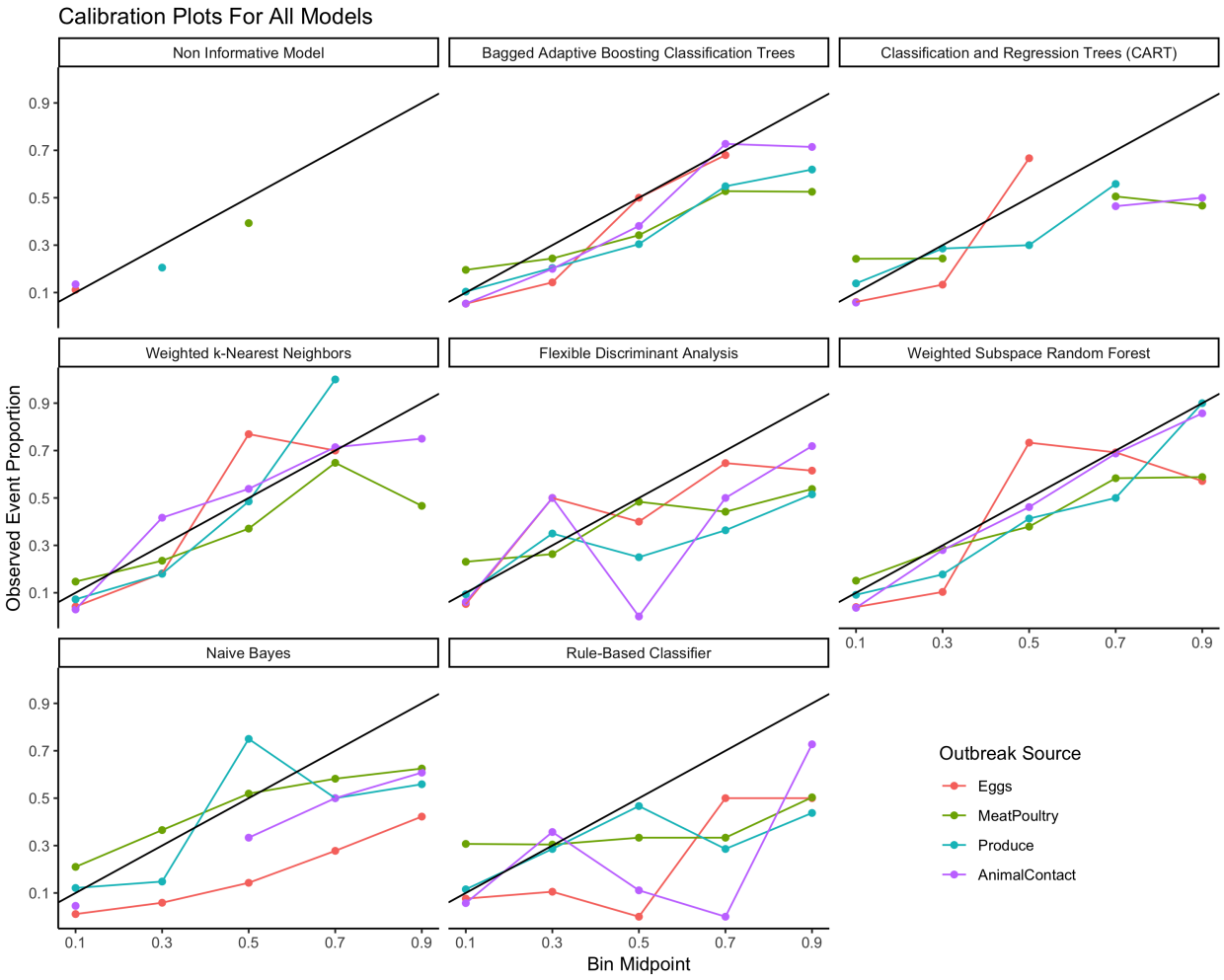
With knn, a source with a predicted probability from 0 to 20% was correct 5% of the time and a source with a predicted probability from 80 to 100% probabilies was correct 56.5% of the time (table 3).

Table 4: Percent of Time Correct

	[0,0.2]	(0.2,0.4]	(0.4,0.6]	(0.6,0.8]	(0.8,1]
Eggs	4	18	77	70	NA
MeatPoultry	15	24	37	65	47
Produce	7	18	49	100	NA
AnimalContact	3	42	54	71	75
All	5	23	44	68	57

Discussion

Outbreaks with eggs as source were much more likely to have missing information (higher percent sex unknown, higher age unknown, and a higher percentage of missing hospitalization information). As there is a probability that this information will not be unknown to a epidemiologist using the tool and represents reporting bias in the training data, this could lead to less accurate results in practice.


Calibration Plots For All Models
Non Informative Model
Bagged Adaptive Boosting Classification Trees
Classification and Regression Trees (CART)
Weighted k-Nearest Neighbors
Flexible Discriminant Analysis
Weighted Subspace Random Forest
Naive Bayes
Rule-Based Classifier
Observed Event Proportion
Bin Midpoint
0.1
0.3
0.5
0.7
0.9
Outbreak Source
Eggs
MeatPoultry
Produce
AnimalContact