Predicting death in patients with pneumonia under the age of 21

Kristen McGarry, Ryan Barbaro & Nigel Michki

December 12, 2016

Prediction

Predict death amongst patients with pneumonia under the age of 21

Importance: WHO 156 million cases of pneumonia in children with 20 million requiring hospitalization

In the US mortality rate is low < 1 per 1000 cases, but still more than 1,000 children die from pneumonia each year

To improve survival we need to focus efforts on patients at high risk of mortality

HCUP KID Dataset

- Healthcare Cost and Utilization Project (HCUP) Kids' Inpatient Database (KID)
- Collection of administrative data concerning inpatient hospitalizations for patients under the age of 21
- Largest publically-available all-payer pediatric inpatient care database in the US

Table 1

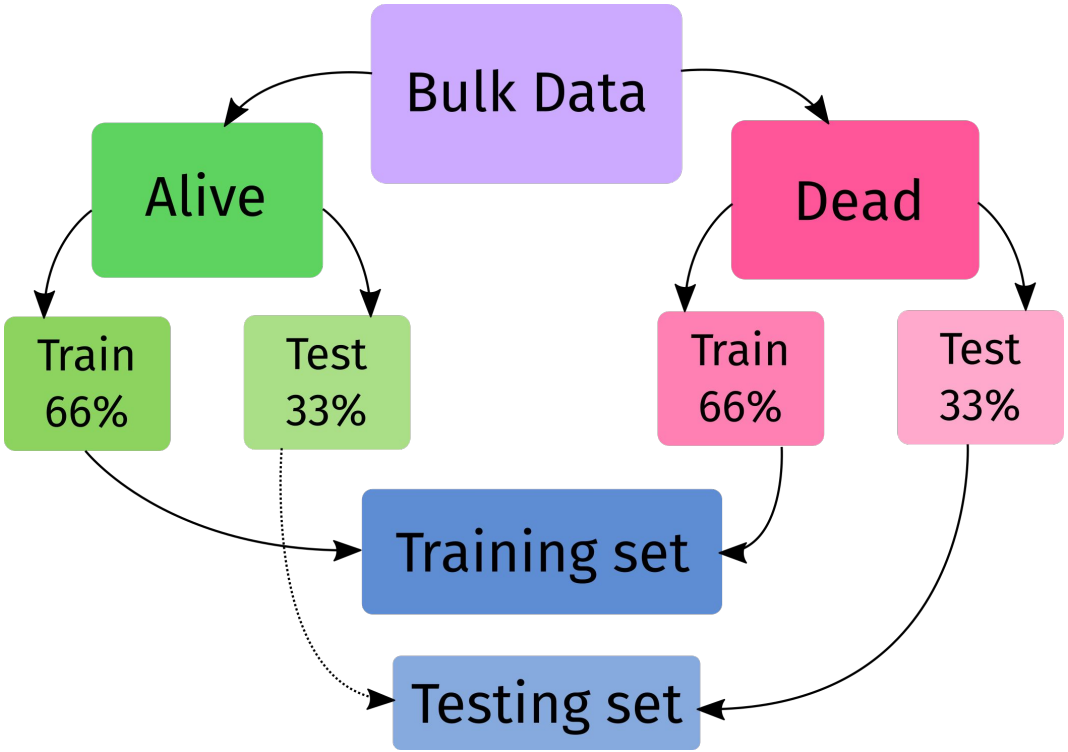
	TEST	TRAIN	p
n	3895	5619	
n_Organ_Failures (mean (sd))	1.59 (0.86)	1.58 (0.86)	0.615
n_Comorbidity (mean (sd))	0.81 (0.93)	0.81 (0.92)	0.897
n_diagnoses (mean (sd))	14.07 (6.54)	14.09 (6.52)	0.892
Operating_Room_Procedure (mean (sd))	0.25 (0.43)	0.27 (0.44)	0.086
Transferred_In (mean (sd))	0.30 (0.46)	0.29 (0.45)	0.113
Admitted_Weekend (mean (sd))	0.26 (0.44)	0.27 (0.44)	0.467
Month_of_Admission (mean (sd))	6.32 (3.74)	6.24 (3.71)	0.343
Childrens_Hospital (mean (sd))	0.32 (0.47)	0.31 (0.46)	0.390
n_Hostpital_Beds (mean (sd))	2.64 (0.61)	2.67 (0.59)	0.040
Mean_Income_of_Zipcode (mean (sd))	2.21 (1.14)	2.22 (1.15)	0.690
Pay_Type (mean (sd))	1.50 (0.65)	1.52 (0.65)	0.177
DIED (mean (sd))	0.07 (0.26)	0.07 (0.26)	0.578
Test_or_Train = TRAIN (%)	0 (0.0)	5619 (100.0)	<0.001

Missing Values

- Chose to omit observations that had any NAs in the variables of interest

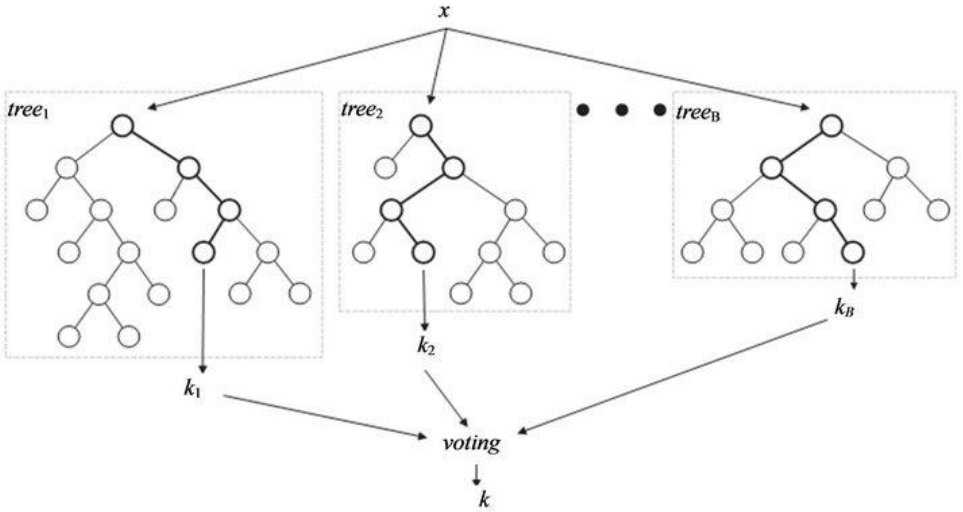
	Percent (number) of cases with deaths
Filtered data with NAs *included*	7.36% (703 out of 9514)
Filtered data with NAs *omitted*	7.32% (696 out of 9514)

Constructing test/train data



Models

1. K-Nearest Neighbor
2. **Random Forest**
    a. Generates multiple uncorrelated classification trees
        i. Training set for each tree = sampling with replacement
    b. Counts “votes” for the classifications
    c. *Advantages:* Avoids overfitting, fast with large datasets, inherent cross-validation



Performance Measures

- Accuracy
  - Gives quick idea of performance
  - Number of correct classifications over all classifications
- F1
  - Average recall (TP/(TP+**FN**)) and precision (TP/(TP+**FP**))
  - Average measure of false negatives and false positives

Results

Random forest

Sample fraction (alive)	**% Dead**	**Accuracy** [Train, CV, Test]	**F1** [Train, CV, Test]
1/1	7.3	99.1, 92.2, **91.6**	98.1, 5.1, **5.6**
1/2	13.6	99.7, 84.9, **85.2**	99.0, 15.9, **25.2**
1/4	24.0	99.8, 77.6, **77.1**	99.7, 44.3, **43.5**
1/10	44.1	99.7, 68.5, **68.7**	99.7, 62.4, **64.8**

Results

K-nearest neighbors

Sample fraction (alive)	**% Dead**	**Accuracy** [Train, CV, Test]	**F1** [Train, CV, Test]
1/1	7.3	94.8, 93.0, **91.6**	44.8, 9.9, **14.4**
1/2	13.6	92.0, 85.1, **84.9**	58.8, 21.9, **21.1**
1/4	24.0	88.7, 75.2, **77.6**	71.4, 34.5, **41.7**
1/10	44.1	87.8, 63.4, **66.6**	85.3, 55.5, **60.8**

Results

- Random forest outperforms k-nearest neighbors in all cases
  - Inherent cross-validation in the learning model due to multiple decision trees voting
  - k-space is large for this model (11 features of interest)
- Increasing F1 scores by reducing sample size drastically reduces accuracy
  - Implies: poor performance not due to overfitting or ‘squashing’ of dead patient records
  - Too many features? Features that determine death not selected? Too random an occurrence?

Q&A