Predicting urinary tract infections in the emergency department with machine learning

Authored by: [Is it credible]

R. Andrew Taylor*, Christopher L. Moore, Kei-Hoi Cheung, Cynthia Brandt Department of Emergency Medicine, Yale University School of Medicine, New Haven CT, United States of America.

This study was approved by the institutional review board (Yale Human Research Protection Program) and waived the requirement for informed consent.

Published: [Is the research relevant to today]

March 7, 2018.

The Problem: [Is a problem well defined]

Usually, urinary tract infections are misdiagnosed (occurs 30%-50% of the time in the US) because the golden standard metric for antibiotic determination takes at least 48 hours (urinary culture analysis).

Problem Importance: [Is it of high significance]

Previous research shows insufficient diagnosis which results in serious consequences such as delayed treatment and antibiotic resistance.

Objective: Apply machine learning to selected features of the diverse dataset of NHAMCS (adults only) to predict whether or not UTI is present. [purpose matches our needs]

Features used: [verify feature extraction process]

demographic information, vitals, laboratory results, past medical history, chief complaint, and structured historical and physical exam findings. [a total of 211 features! Seem quite relevant]

Methodology: [ways to try. Does it inspire our upcoming experiments?]

1. 80,387 adults who visit the ER were considered in this dataset.
2. Features to be fed to the model were 2 options; a) all 211 aforementioned features and b) 10 extracted features.
3. 6 different machine learning models were tried: random forest (tree ensembling, that is), extreme gradient boosting (XGBoost), adaptive boosting, support vector machine (SVM), elastic net, neural network, and logistic regression.
4. The predictions were compared. (what metric?)

=>This resembles our general dataset exploration process to an extent. It also matches our experiment design (the 3 models are explored).

Results: [how robust is the methodology]

For both the full features set and reduced features set, the XGBoost performed the best, outperforming all previous literature. For our experiments, from best to worst: Random forest, SVM, neural network.

Where Many Lessons are Learned… The Process Recapped.

Feature selection:

1. The symptoms and measurements of UTI were researched.
2. The dataset was searched for these symptoms.
3. Only the records with at least one of these symptoms available were collected.
4. Demographics, complaints, and past medical history were added.
5. The numbers, 211 and 10, were chosen through expert literature review to address user acceptance concerns (i.e. using an online calculator to predict UTI).

=> medications were NOT included to remove the bias we are trying to avoid: the misdiagnosis!

Preprocessing:

1. K-means clustering (data point goes to cluster with closest mean | 5 clusters) was performed on all continuous variables. Why clustering? To better represent the information. Why k=5? To represent the standard scale (critically low, low, normal, high, critically high) + inflection point: increasing k did not really affect the variance.
2. Errant text data were improved through regex searches.
3. Missing values were included and treated as categorical variables "not recorded". In the clinical context, it provides more info about the patient and improves performance.

=> clustering helps extract needed info from a feature. Missing values can actually SHOW MORE information!

n.b. This process was built upon that of the paper "Prediction of In‑hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data–Driven, Machine Learning Approach.**" View it here: https://onlinelibrary.wiley.com/doi/full/10.1111/acem.12876*

Model Selection and Training:

1. **The selected models:** random forest (tree ensembling, that is), extreme gradient boosting (XGBoost), adaptive boosting, support vector machine (SVM), elastic net, neural network, and logistic regression. Why? Because...
    - Their resilience to overfitting.
    - Their ability to model non-linear relations. Yesterday, we demonstrated that correlation analysis gave low scores for linearity. This is an important point!
    - Having so many experiments and limited time, they're easy and fast to implement.
    - Logistic regression is commonly used as the baseline in the medical field.
2. **The partitioning:** 80/20. Done at random.
3. **Hyperparameter tuning:** 10-fold cross-validation and grid searches were used.

Metrics:

- **Primary metric:** AUC of ROC (receiver operating characteristics). Why? Because they are good choices if the target is binary (UTI or no UTI).
- **Other metrics**: sensitivity, specificity, positive & negative likelihood ratios.
- **To compare the 211 and 10 feature cases:** confusion matrix used.

Comparisons:

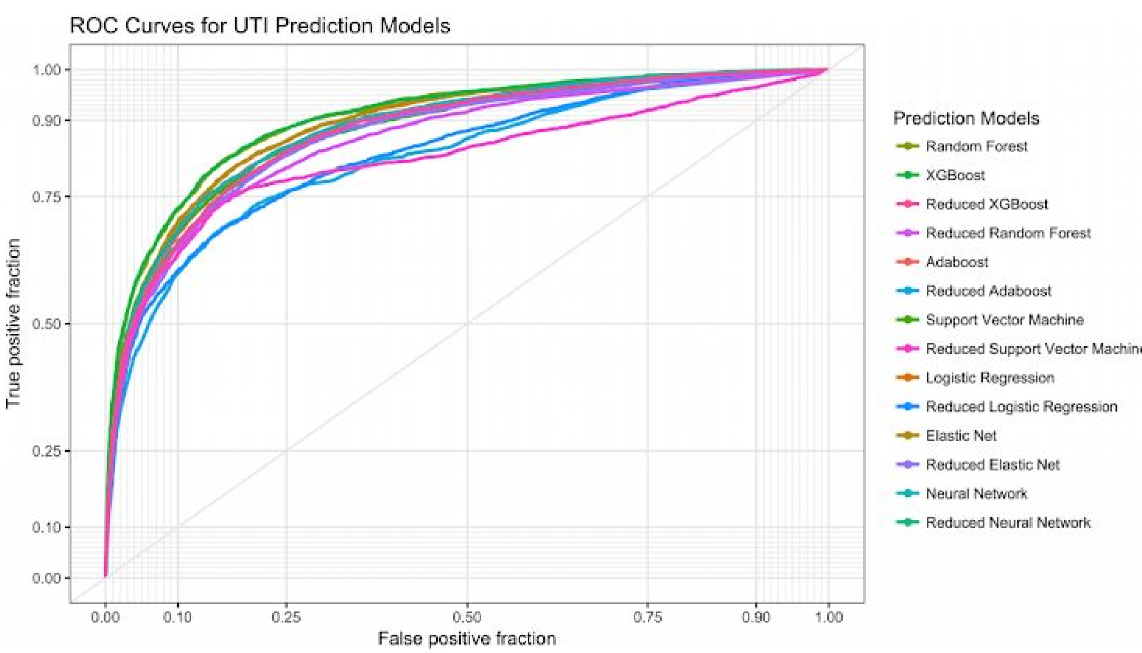

Fig 2. Receiver operating characteristic (ROC) curves for different machine learning models.

Table 4. Test characteristics of UTI prediction models on validation data*.

Models	AUC (95%CI)	Sensitivity (95% CI)	Specificity (95% CI)	+LR (95% CI)	-LR (95% CI)	Accuracy (95% CI)	P-value
XGBoost	**.904(.898-.910)**	61.7(60.0–63.3)	94.9 (94.5–95.3)	12.0(11.1–13.0)	.404(.387-.421)	87.5 (87.0–88.0)	NA
Random Forest	.902(.896-.908)	57.3(55.6–58.9)	96.0 (95.6–96.3)	14.3(13.0–15.6)	.445(.428-.462)	87.4 (86.9–87.9)	0.58
Adaboost	.880(.874-.887)	62.2(60.6–63.8)	92.3(91.8–92.7)	8.06(7.54–8.61)	.409(.392-.427)	85.6(85.1–86.2)	< .001
Support Vector Machine	.878(.871-.884)	49.6(47.9–51.2)	96.8(96.4–97.1)	15.3(13.8–16.9)	.521(.504-.538)	86.3(85.7–86.8)	< .001
ElasticNet	.892(.885-.898)	56.8(55.2–58.4)	94.9(94.5–95.2)	11.1(10.2–12.0)	.455(.438-.473)	86.4(85.9–87.0)	< .001
Logistic Regression	.891 (.884-.897)	57.5(55.8–59.1)	94.7(94.3–95.1)	10.9(10.0–11.8)	.449(.432-.466)	86.4(85.9–87.0)	< .001
Neural Network	.884 (.878-.890)	54.6(52.9–56.2)	95.3(95.0–95.7)	11.7(10.8–12.8)	.476(.460-.494)	86.3(85.8–86.8)	<001
Reduced XGBoost	**.877(.871-.884)**	54.7(53.0–56.3)	94.7(94.3–95.1)	10.4(9.6–11.3)	.479(.462-.496)	85.9(85.3–86.4)	< .001
Reduced Random Forest	.861(.853-.868)	54.8(53.1–56.4)	94.3(93.9–94.7)	9.66(8.94–10.4)	.479(.462-.497)	85.5(85.0–86.1)	< .001
Reduced Adaboost	.826(.817-.834)	61.9(60.3–63.5)	88.8(88.2–89.3)	5.50(5.21–5.81)	.429(.412-.448)	82.8(82.2–83.3)	< .001
Reduced Support Vector Machine	.822(.813-.832)	49.4(47.8–51.1)	95.8(95.4–96.1)	11.7(10.7–12.9)	.528(.511-.546)	85.5(84.9–86.0)	< .001
Reduced Elastic Net	.870(.863-.877)	52.4(50.7–54.1)	95.2(94.8–95.5)	10.9(9.99–11.8)	.500(.482-.571)	85.7(85.1–86.2)	< .001
ReducedLogistic Regression	.870(.863-.877)	53.3(51.6–54.9)	94.8(94.4–95.2)	10.3(9.52–11.2)	.492(.476-.510)	85.6(85.0–86.2)	< .001
Reduced Neural Network	.873(.867-.881)	54.0(52.3–55.6)	95.0(94.6–95.4)	10.9(10.0–11.8)	.485(.468-.502)	85.9(85.4–86.5)	< .001

* Test Characteristics determined at optimal AUC threshold

Full models were developed on 212 variables, while the reduced models were developed on 10 variables.

P-values obtained by AUC comparison to best performing model

Noteworthy limitations in previous literature: [mistakes to avoid]

Paper 1

- Small datasets
- Few features used (e.g. urine dipstick or urinalysis results)

=> *make sure to try several sets of features with varying sizes.*

Paper 2

- Female-only dataset
- Most patients were generally healthy
- high prevalence of UTI
- Based on points 2 and 3, very limited generalizability.

=> *if sampling is performed, pay attention to biases per attribute, and don't fall for Simpson's Paradox!*

References

Main paper: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194085

Supplementary paper: *https://onlinelibrary.wiley.com/doi/full/10.1111/acem.12876*