Machine Learning for the Real World: Safety & Trustworthiness

Rich Caruana
Computer Science Department
Cornell University

Summary:

If Your Hair Doesn't Turn Gray You're Not Paying Enough Attention!

- Three case studies:
  - Pneumonia risk prediction (to scare you)
  - C-Section rate evaluation (to make you think)
  - KDD Cup 2004 Competition (for fun)

Predicting Dire Outcomes of Patients with Community Acquired Pneumonia

Gregory Cooper, Vijoy Abraham, Constantin Aliferis, John Aronis, Bruce Buchanan, Rich Caruana, Michael Fine, Janine Janosky, Gary Livingston, Tom Mitchell, Stefano Monti, Peter Spirtes

Pneumonia

- Millions of cases each year in U.S. alone
- $$$ (hundreds of millions of dollars per year in U.S.)
- 6-8% of patients die from pneumonia
- But many patients can/should be treated as outpatients

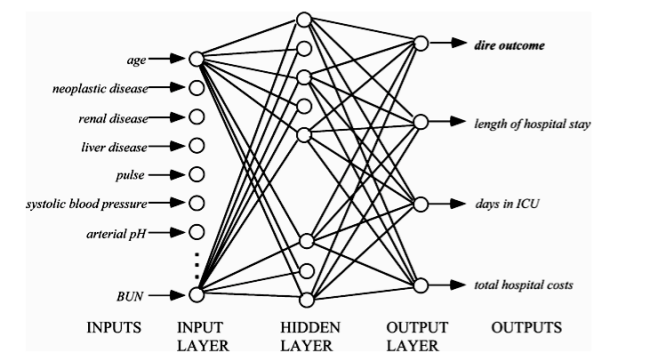
- Learn model to predict high risk patients
  - High risk => in hospital
  - Low risk => chicken soup and antibiotics

- Large multi-institution study (CMU, Pitt, many hospitals)
  - Each group uses their favorite learning method(s)

Learning Methods

- Logistic regression
- Naïve bayes
- Bayes nets (fan models, others)
- Rule-based methods
- Finite mixture models
- Artificial neural nets
- Feature selection
- Missing value imputation
- Multitask learning

Multitask Neural Nets



Multitask Neural Nets Perform Best

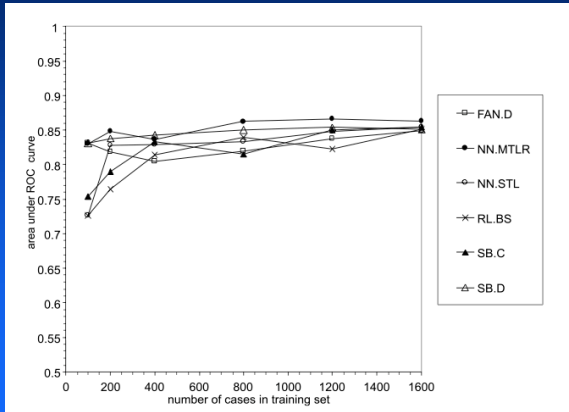


Fig. 5. The ROC curve areas for the models that perform the best when using 1601 training cases.

What model did we deploy?

Learning Methods

- Logistic regression
- Naïve bayes
- Bayes nets (fan models, others)
- Rule-based methods
- Finite mixture models
- Artificial neural nets
- Feature selection
- Missing value imputation
- Multitask learning

Learning Methods

- **Logistic regression**
- Naïve bayes
- Bayes nets (fan models, others)
- Rule-based methods
- Finite mixture models
- Artificial neural nets
- Feature selection
- Missing value imputation
- Multitask learning

Why did we deploy the worst model!?

Once upon a time…

- At a random meeting…
- Grad student doing rule learning said…
- We learned a funny rule last night…
- Learned the rule:

Asthma => Lower Risk (from pneumonia)

True Pattern in Data!!!

- Some patients admitted to hospital
- Some patients not admitted
- Asthmatics always admitted, treated aggressively
- Good news: aggressive treatment works for them!
- Lowers risk of dying compared to general population
- So asthmatics have less probability of death than others
- This rule is statistically true in data
- If we use this rule to decide to admit, may kill asthmatics!
- Did neural net learn this rule? You betcha!
- How do we excise this rule?
- What other bad true things did neural net learn?

In Summary

- Neural Net model probably would save more patients than it would kill
- But it probably would start killing at least one class of patients (asthmatics) that were safe before ML
- Bad medicine
- In this domain, intelligibility (or explanation) was critical before fielding model

What has the model you are about to deploy learned?

Using Machine Learning to Model Standard Practice: Retrospective Analysis of Group C-Section Rate via Bagged Decision Trees

Rich Caruana — Cornell CS
Stefan Niculescu — CMU CS
Bharat Rao — Siemens Medical
Cynthia Simms — Magee Hospital

C-Section in the U.S.

- C-section rate in U.S. too high
  - Western Europe has lower rates, but good outcomes
  - C-section is major surgery => tough on mother
  - C-section is expensive => tough on everyone
- Why is U.S. rate so high?
  - Convenience (rate highest Fridays, before sporting events, …)
  - Litigation
  - Social and Demographic issues
- Current controls inadequate
  - Financial: pay-per-patient instead of pay-per-procedure
  - Physician reviews: monthly/quarterly evaluation of rates

Risk Adjustment

- Some practices specialize in high risk patients
- Some practices have low-risk demographics
- ***Must*** correct for patient population seen by each practice

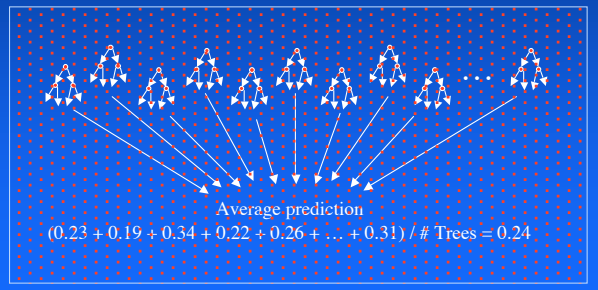
- Model *Standard Practice* and compare to it
  - Not trying to improve outcomes
  - Maintain quality of outcomes while reducing c-sec rate
  - Compare physician practices to other physician practices
  - Warn physicians if rate higher than other practitioners

Data

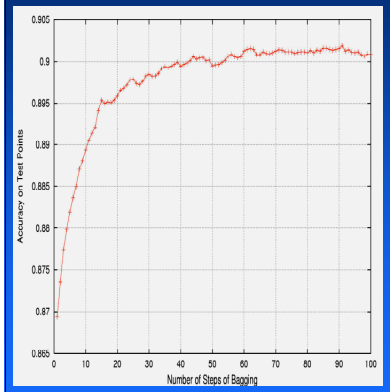
- 22,175 expectant mothers
- 17 physician practices
- 16.8% c-section rate on average
- C-section rate varies from 13% to 23% in practices

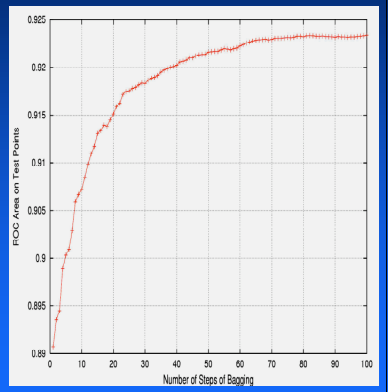
Bagged Decision Trees

- Draw 100 bootstrap samples of data
- Train trees on each sample -> 100 trees
- Average prediction of trees on out-of-bag samples



Bagging Improves Accuracy and AUC





Is Good Accuracy/AUC Good Enough?

- NO! NO! NO!

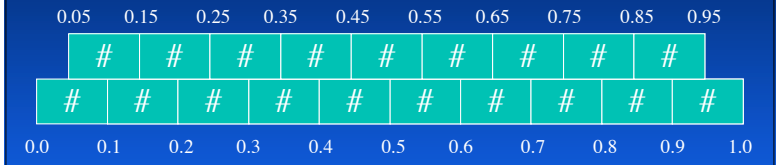
- Model must be just as accurate for low risk patients as for high risk patients as for moderate risk patients!
- Otherwise predictions can be biased on low or high risk practices.
- This means models must be well calibrated.
- We had never heard of calibration before.
- We discovered that some models make poorly calibrated probabilistic predictions, while others are well calibrated.

Calibration

- When you predict p = 0.2 for 1000 patients, about 200 of them better turn out positive.
- This must be true for all predicted p on [0,1].
- Model can be accurate but poorly calibrated
  - good threshold with uncalibrated probabilities
- Model can have good ROC but be poorly calibrated
  - ROC insensitive to scaling/stretching
  - only ordering has to be correct, not probabilities themselves

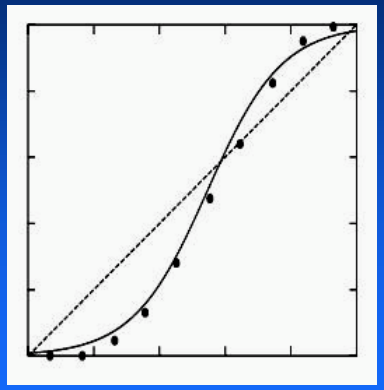
Measuring Calibration

- Bucket method

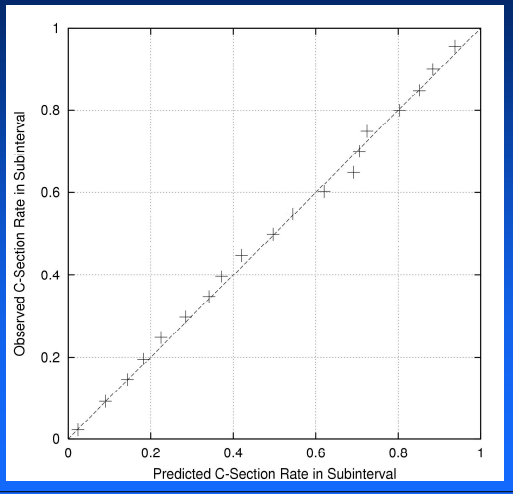


- In each bucket:
  - measure observed c-sec rate
  - predicted c-sec rate (average of probabilities)
  - if observed csec rate similar to predicted csec rate => good calibration in that bucket

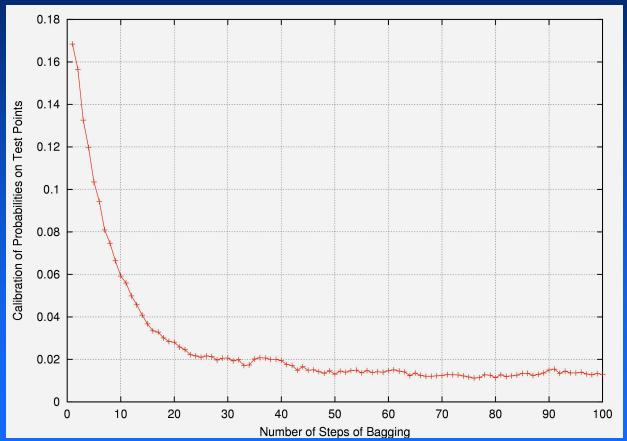
Not All Models Are Well Calibrated



Bagged Decision Tree Calibration Plot



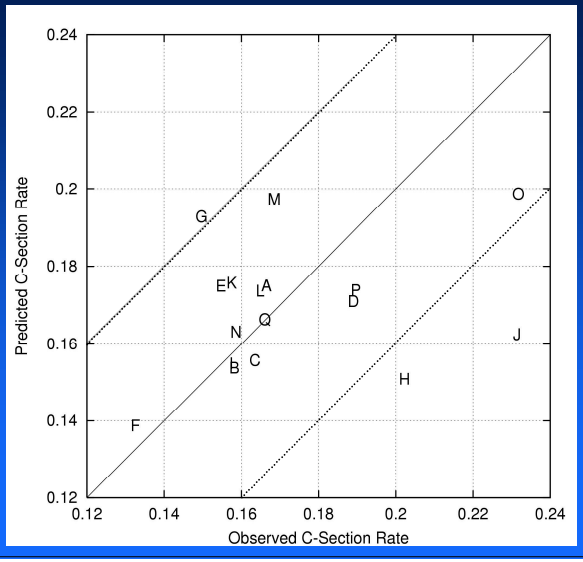
Bagging Improves Calibration an Order of Magnitude!



Aggregate Risk

$$Aggregate\_Risk(group) = \frac{\sum_{p \in group} prediction(p)}{\#\ patients \in group}$$

- Aggregate risk of c-section for a group is just the average probability of c-section for each patient in that group
- If a group's aggregate risk matches the observed c-sec rate, then the group is performing c-secs in accordance with standard practice (as modeled by learned model)



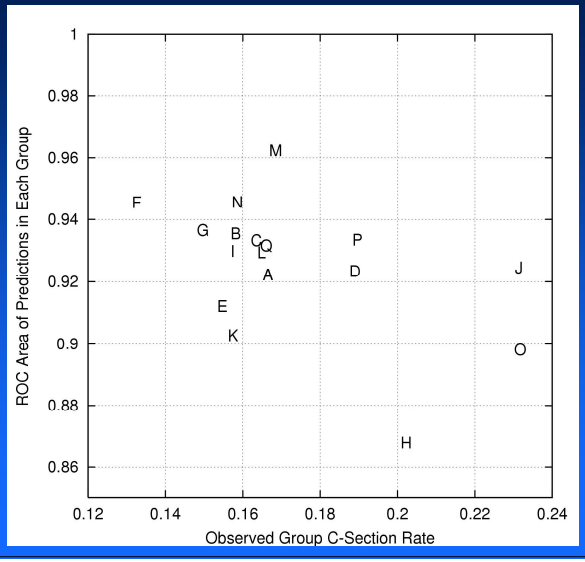
Are We Done Worrying Yet?

Are We Done Worrying Yet?

Not Yet. What else could be wrong?

Additional Check

- Could models have good calibration in tails, but still be inaccurate on some groups of patients?
- Suppose two kinds of patients, A and B, both with same **true** risk: p(A) = p(B) = 0.2
- Patients A and B differ, model might be more accurate or better calibrated on A than B, so a practice seeing patients A might over/under estimate relative to practice seeing B
- Sanity check: are models good on each physician group?



What Did We Deploy?

- Only "deployed" model on practices where we thought it was good enough
  - Probably good to deploy on practices F, J, and O
  - Need human expert to look at practice H
- Used multiple performance metrics, multiple ways of examining model performance
  - Deployed "inferior" model: bagged trees
- Tried to think of everything that could fool us
  - That's what made my hair turn gray!

Don't have to deploy something to turn your hair gray

Just run a competition!

KDD-Cup 2004

Chairs: Rich Caruana & Thorsten Joachims
Web Master++: Lars Backstrom

Cornell University

KDD-Cup Tasks

- Task1: Particle Physics
  - Accuracy
  - Cross-Entropy
  - ROC Area
  - SLAC Q-Score
- Task2: Protein Matching
  - Squared Error
  - Average Precision
  - Top 1
  - Rank of Last

Task 2: Protein Matching

- Data contributed by Ron Elber, Cornell University
- Finding homologous proteins (structural similarity)
- 74 real-valued features describing match between two proteins
- Data comes in blocks
- Unbalanced:
  typically < 10 homologs (+) per block of 1000
- Train: 153 Proteins (145,751 cases)
- Test: 150 Proteins (139,658 cases)

Pr 1	Pr 2	…	Pr N
–	+	…	–
–	–	…	–
+	–	…	+
–	+	…	–
…	…	…	…
–	–	…	–

and the winners are…

Task 2: Protein Winners

Rank	Top 1	Rank Last	Squared Error	Average Precision	Average Rank
1st	0.9200	45.62	0.0350	0.8412	4.125
2nd	0.9133	52.42	0.0354	0.8409	4.500
3rd	0.9067	52.45	0.0369	0.8380	5.500

Katharina Morik et al. (University of Dortmund):
HM Rank Last

David Vogel et al. (Aimed / University of Central Florida):
3rd place overall, HM Top1

Yan Fu et al. (Inst. of Comp. Tech., Chinese Academy of Sci.):
2nd place overall, HM Squared Error, HM Average Precision

Bernhard Pfahringer (University of Waikato):
1st place overall

Bootstrap Analysis of Results

- How much does selection of winner depend on test set? (139,658 points for 150 proteins)
- Bootstrap Algorithm:
  - Repeat many times (10,000 trials):
    - Take bootstrap sample of proteins from test set
    - Evaluate performance on bootstrap sample
    - Re-rank participants on sample
  - What is probability of each team winning/placing?

Protein Winners: Bootstrap Analysis

		Overall rank on bootstrap sample				
		1st	2nd	3rd	4th	5th
Overall Rank on official test set	1st	14%	29%	26%	16%	8%
	2nd	59%	24%	10%	5%	2%
	3rd	22%	28%	23%	14%	7%
	4th	4%	12%	22%	26%	17%
	5th	0%	2%	6%	12%	20%

• 10,000 bootstrap samples

Physics Winners: Bootstrap Analysis

		Overall rank on bootstrap sample				
		1st	2nd	3rd	4th	5th
Overall Rank on official test set	1st	100%	0	0	0	0
	2nd	0	100%	0	0	0
	3rd	0	0	94%	6%	0
	4th	0	0	6%	93%	1%
	5th	0	0	0	1%	76%

- 1000 bootstrap samples

Discussion

- You can't be too careful
- Good accuracy, good AUC usually not enough
- The right metric(s) are critical
- If best model is opaque, train simpler models, too
- Is model biased for some kinds of cases?
- Phased deployment in critical applications?
- Deploy multiple models and raise warning if they disagree too much?