Heritage Health Provider Competition: How we made it to top 10

Presented by

Ad Tayal

Supervisors:

Thomas F. Coleman

Yuying Li

Cheriton School of Computer Science

UNIVERSITY OF WATERLOO

uwaterloo.ca

May 2013

Outline

- Competition
- Challenges
- Algorithms
- Computation
- Conclusion

Competition to Improve Healthcare

- 71+ Mil individuals are admitted to hospitals each year
- $30 to $40 Billion spent on **unnecessary** hospitalizations every year

UNIVERSITY OF WATERLOO

Competition to Improve Healthcare

- Can we identify those most at risk and ensure they get treatment they need?
- Heritage Provider Network (HPN) sponsored a global incentivized competition… with a grand prize of $3 Mil
- Predict hospitalization in the next year using historical claims data

UNIVERSITY OF WATERLOO

In the News…

- Competition attracted 1,660 teams including data scientists, biostatisticians, physicians, engineers, and many leading industry players from all around the world

THE WALL STREET JOURNAL. May the Best Algorithm Win…

With $3 Million Prize, Health Insurer Raises Stakes on the Data-Crunching Circuit

The New York Times

Change the World, and Win Fabulous Prizes

The Economist

Incentive prizes

Healthy competition

Apr 10th 2011, 17:00 by G.F. | SEATTLE

Forbes Kaggle's Predictive Data Contest Aims To Fix Health Care

COMPUTERWORLD Health care network offers $3M for illness-prediction algorithm

The Washington Post

Like public policy? Want to win some money?

POPSCI THE FUTURE NOW Company Offers $3 Million for an Algorithm That Predicts Whether You'll Get Sick

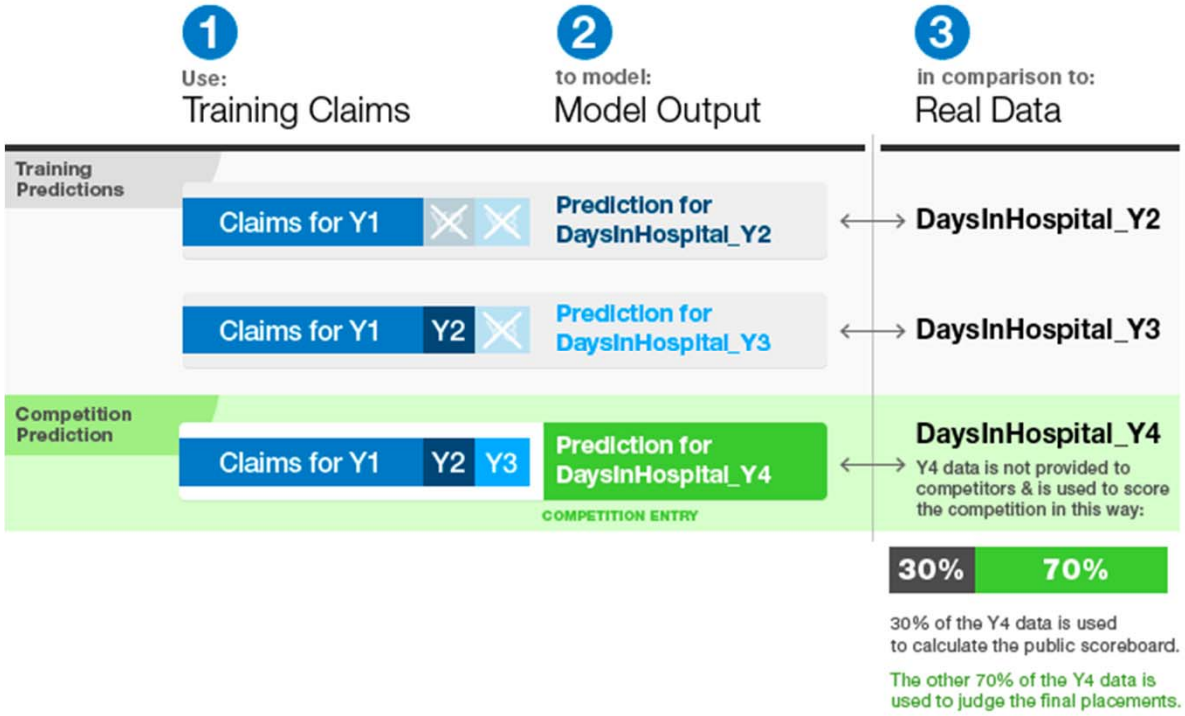
BUSINESS INSIDER Heritage Provider Network Wants To Fix Healthcare With $3 Million And Tech

THE DAILY CALLER

Is a $3 million health-care prize for innovation the answer to Obama's call?

UNIVERSITY OF WATERLOO

Claims Data and Evaluation



$$\varepsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\ln(p_i + 1) - \ln(a_i + 1)]^2}$$

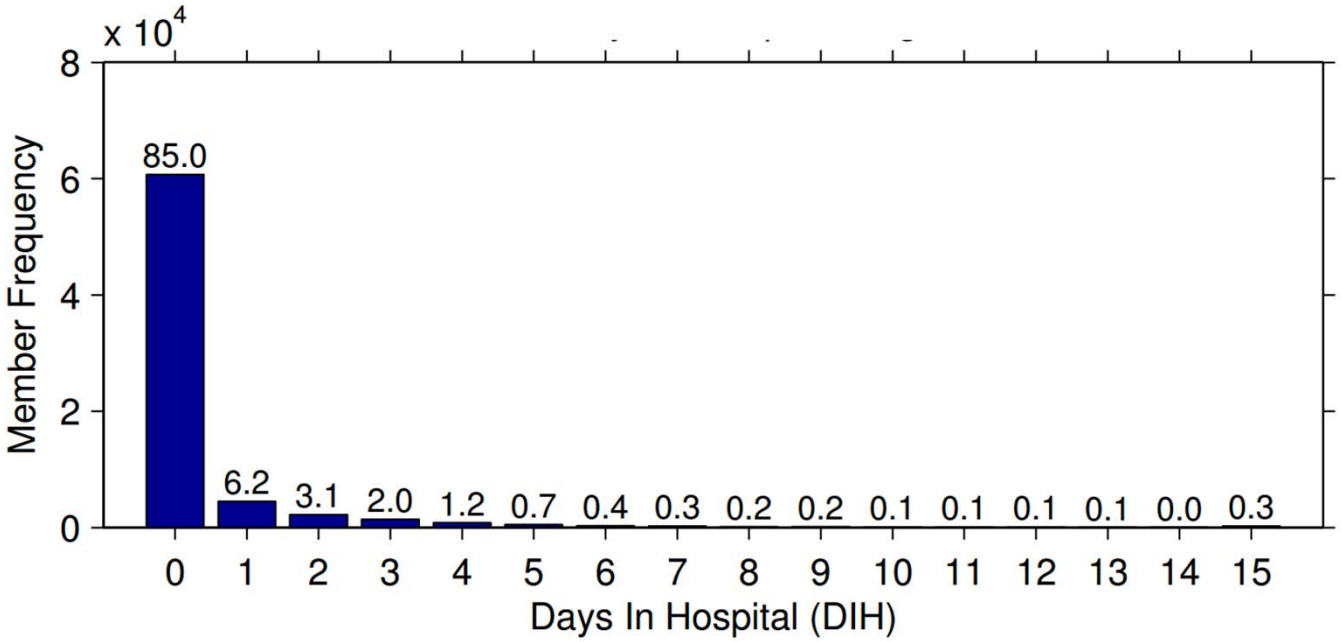
UNIVERSITY OF WATERLOO

Challenges

- Raw data provided
  - Participants required to construct their own features
- Anonymized data
  - “Data can either be useful or perfectly anonymous but never both”, Ohm (2009)
- Noisy and missing data
  - Inconsistent processing
- Mixed attributes
  - Numerical, ordinal, categorical, identifiers

Challenges

- Highly unbalanced target distribution
- Hospitalization for a week or longer occurs about 1% of the time, but influences the score the most
- Need specialized algorithms…



UNIVERSITY OF WATERLOO

More challenges…

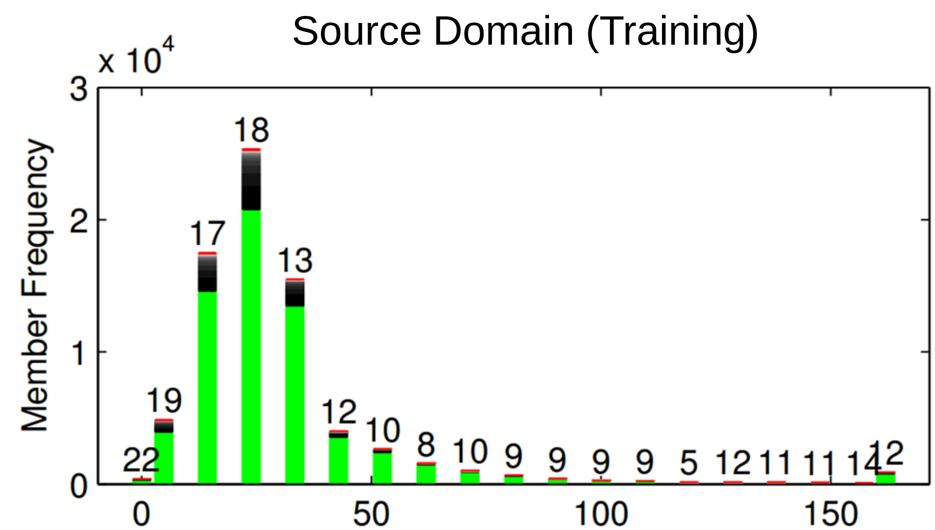
- Covariate Shift

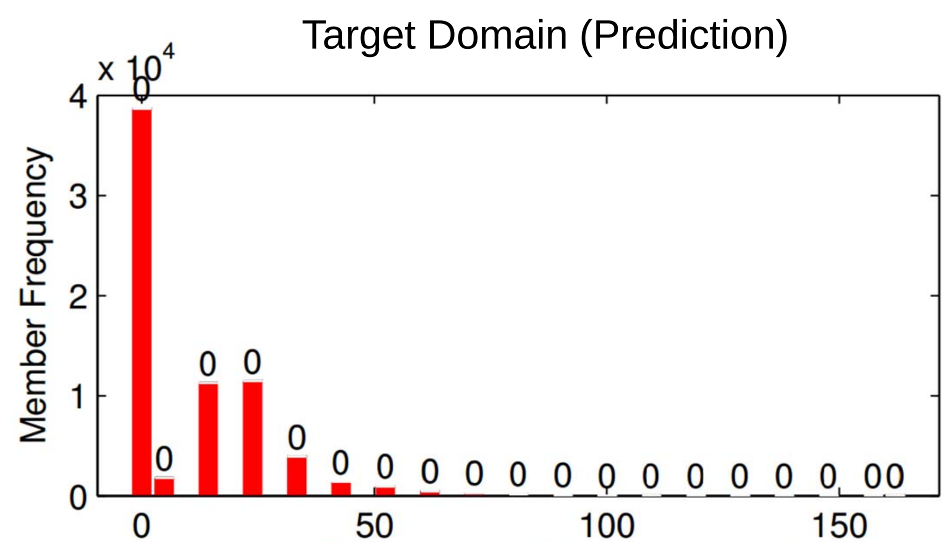
Conditional distributions are the same in the source and target domains.

But marginal distributions may be different in the two domains.

$$P_s(Y|X = x) = P_t(Y|X = x)$$

$$P_s(X) \neq P_t(X)$$



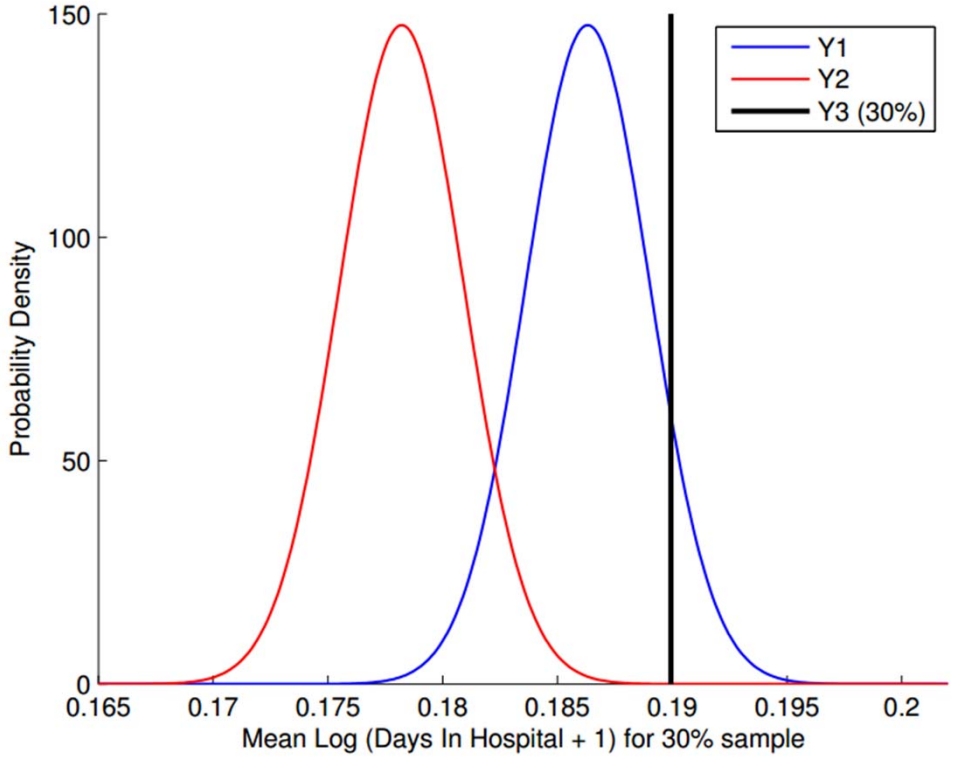


More challenges…

- Non-Stationarity

Conditional distributions are different in the source and target domains as well.

$$P_s(Y|X = x) \neq P_t(Y|X = x)$$



UNIVERSITY OF WATERLOO

Some standard algorithms…

- Linear regression (L1 and L2 regularization)
- Logistic regression (L1 and L2 regularization)
- Weighted SVM
- Decision Trees
- Bagging
- Boosting

And some of our new algorithms…

- Nonlinear SVM with Feature Selection
- SVR with Feature Selection
- Semi-supervised Support Vector Machines
- Regression with graduated L0 non-convexification for sparsity
- Mixture Gaussian ordinal logistic regression with unknown centers
- Rare class Lago variations with fixed kernel width
- RegSigNet

Blending

- Extensive cross-validation strategy used to evaluate different models
- Blending (or stacked generalization) used to combine models rather than selecting just one

Computation

× 5 datasets
× 4 weighting schemes
× 3 time horizons
× 5 problem types
× 10 folds
× ~10 algorithms
× ~50 parameters/algorithm
× ~10 min/algorithm/parameter

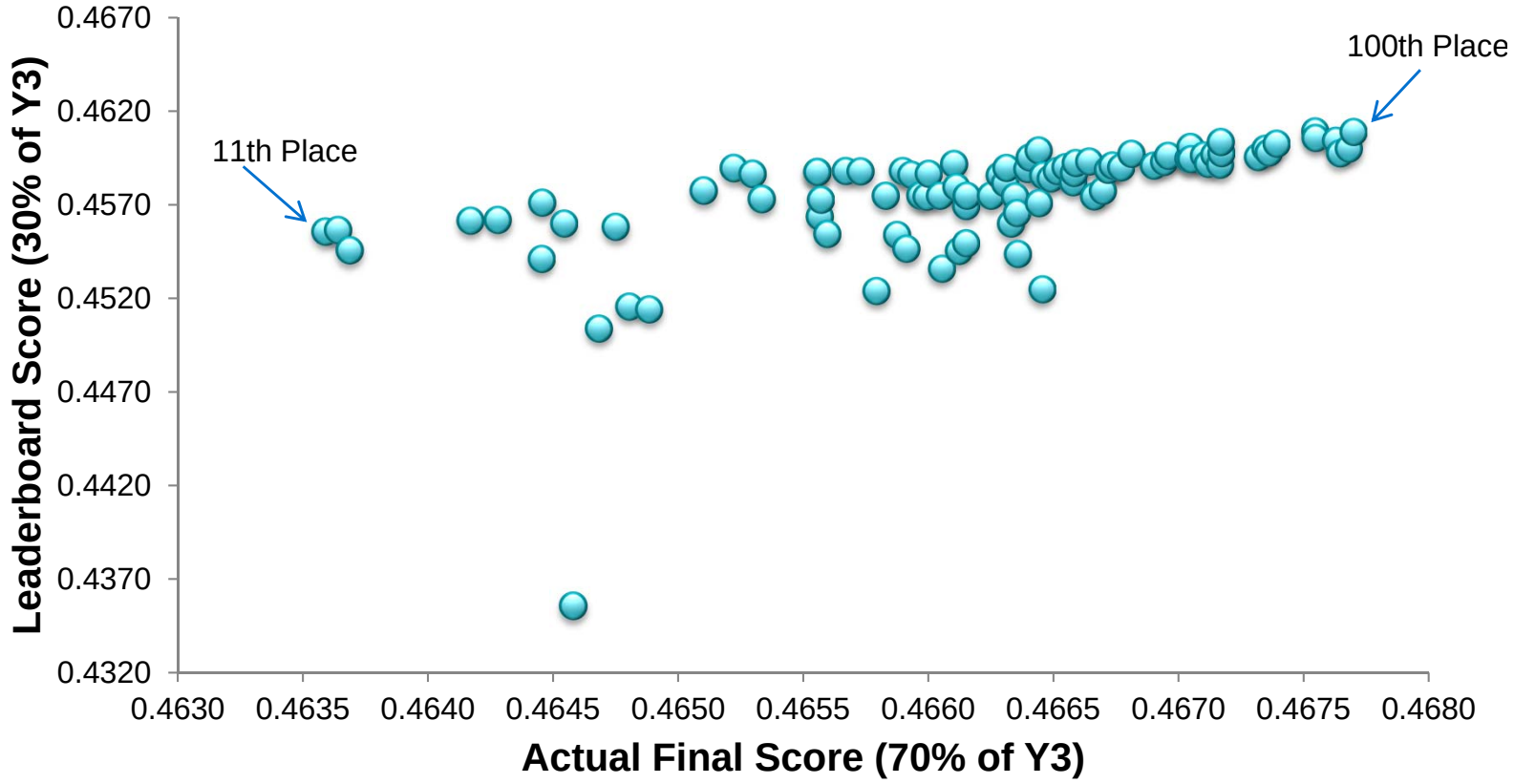
≈ 28 years computation time!

≈ 1 month computation time

Thanking:
Prof. Coleman: 192 cores (COPS)
Prof. Wan: 128 cores (Cabernet)
Prof. Li: 336 cores (M160)

Conclusion

- We ranked top 10 among more than 1600 teams (placement among top 10 will be publicly disclosed at DataPalooza Conference in June)



UNIVERSITY OF WATERLOO