# Ensemble Learning and the Heritage Health Prize

Jonathan Stroud, Igii Enverga, Tiffany Silverstein,
Brian Song, and Taylor Rogers

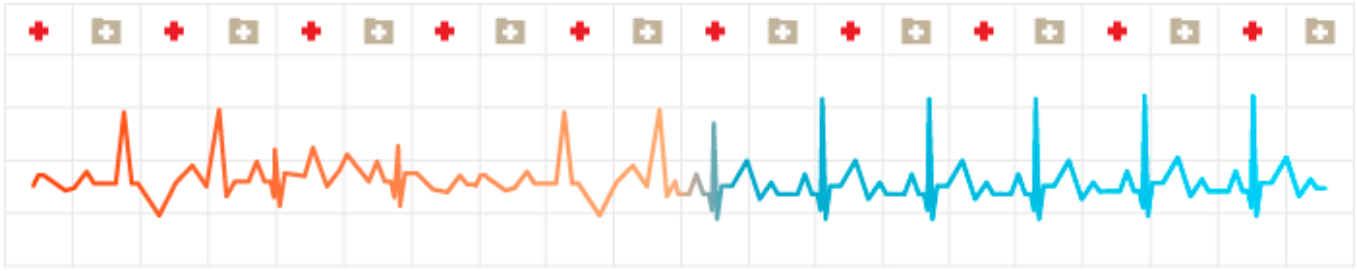
iCAMP 2012
University of California, Irvine
Advisors: Dr. Max Welling, Dr. Alexander Ihler, Sungjin Ahn, and Qiang Liu

December 4, 2013

# Outline

- Heritage Health Prize
  - Data
  - Evaluation
- Our Approach
  - Motivation
  - Models
  - Blending
- Results

# The Heritage Health Prize



**Improve Healthcare, Win $3,000,000.**

- Goal: Identify patients who will be admitted to a hospital within the next year, using historical claims data.[1]
- 1,659 teams

# Purpose

- Reduce cost of unnecessary hospital admissions per year
- Identify at-risk patients earlier

# Evaluation

Root Mean Squared Logarithmic Error (RMSLE)

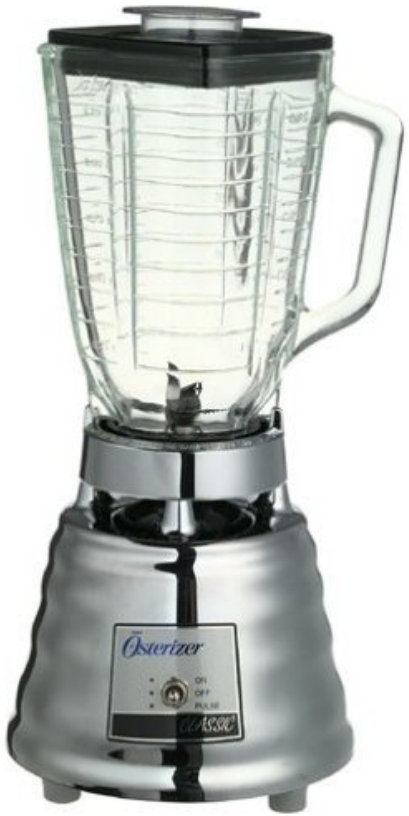
$$\varepsilon = \sqrt{\frac{1}{n}\sum_{i}^{n}[log(p_i + 1) - log(a_i + 1)]^2}$$

Threshold: $\varepsilon \leq .4$

# Our Approach

- Motivation
- Individual Models
- Optimized Ensemble

# Blending




Blend several predictors to create a more accurate predictor
Motivated by solution to the Netflix Prize [3]

# Prediction Models

- Preprocessing: Feature Selection
- K-Nearest Neighbors
- Random Forests
- Gradient Boosting Machines
- Logistic Regression
- Support Vector Regression
- Neural Networks

# Feature Selection

| | AA | AB | AC | AD | AE | AF | AG | AH | AI | AJ | AK | AL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | PayDelay_max | PayDelay_min | PayDelay_ave | PayDelay_stdev | LOS_max | LOS_min | LOS_ave | LOS_stdev | LOS_TOT_UNKNOWN | LOS_TOT_SUPRESSED | LOS_TOT_KNOWN | dsfs_max |
| 2 | 45 | 24 | 30 | 10.0124922 | 0 | 0 | 0 | -1 | 4 | 0 | 0 | 4 |
| 3 | 55 | 0 | 20 | 10.15757901 | 0 | 0 | 0 | -1 | 32 | 0 | 0 | 12 |
| 4 | 85 | 0 | 30 | 23.76856 | 0 | 0 | 0 | -1 | 9 | 0 | 0 | 9 |
| 5 | 30 | 15 | 21 | 5.727128425 | 0 | 0 | 0 | -1 | 6 | 0 | 0 | 1 |
| 6 | 162 | 162 | 162 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 7 | 52 | 0 | 29 | 16.60247287 | 0 | 0 | 0 | -1 | 20 | 0 | 0 | 10 |
| 8 | 15 | 12 | 13 | 1.643167673 | 0 | 0 | 0 | -1 | 5 | 0 | 0 | 1 |
| 9 | 73 | 0 | 22 | 14.38821091 | 1 | 1 | 1 | -1 | 43 | 0 | 1 | 12 |
| 10 | 28 | 0 | 12 | 10.41766662 | 0 | 0 | 0 | -1 | 9 | 0 | 0 | 9 |
| 11 | 162 | 0 | 55 | 64.13267498 | 0 | 0 | 0 | -1 | 9 | 0 | 0 | 10 |
| 12 | 66 | 0 | 33 | 23.22675641 | 0 | 0 | 0 | -1 | 22 | 1 | 0 | 12 |
| 13 | 86 | 23 | 56 | 29.28082421 | 0 | 0 | 0 | -1 | 6 | 0 | 0 | 2 |
| 14 | 43 | 15 | 33 | 16.16580754 | 0 | 0 | 0 | -1 | 3 | 0 | 0 | 5 |
| 15 | 162 | 25 | 58 | 50.83989029 | 2 | 1 | 1 | 0.707106781 | 7 | 0 | 2 | 6 |
| 16 | 43 | 0 | 21 | 30.40559159 | 0 | 0 | 0 | -1 | 2 | 0 | 0 | 10 |
| 17 | 73 | 0 | 22 | 27.72105956 | 0 | 0 | 0 | -1 | 15 | 0 | 0 | 5 |
| 18 | 50 | 14 | 30 | 12.45563556 | 0 | 0 | 0 | -1 | 8 | 0 | 0 | 1 |
| 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 20 | 132 | 0 | 29 | 29.21995032 | 0 | 0 | 0 | -1 | 44 | 0 | 0 | 12 |
| 21 | 101 | 0 | 47 | 24.55139796 | 0 | 0 | 0 | -1 | 44 | 0 | 0 | 12 |
| 22 | 108 | 0 | 43 | 26.90792744 | 1 | 1 | 1 | -1 | 30 | 0 | 1 | 11 |
| 23 | 37 | 21 | 27 | 7.762087348 | 0 | 0 | 0 | -1 | 4 | 0 | 0 | 3 |
| 24 | 71 | 0 | 24 | 9.948280632 | 0 | 0 | 0 | -1 | 44 | 0 | 0 | 11 |
| 25 | 27 | 0 | 14 | 13.52774926 | 0 | 0 | 0 | -1 | 3 | 0 | 0 | 4 |
| 26 | 59 | 10 | 33 | 13.89484317 | 0 | 0 | 0 | -1 | 15 | 0 | 0 | 10 |
| 27 | 19 | 0 | 9 | 13.43502884 | 0 | 0 | 0 | -1 | 2 | 0 | 0 | 1 |
| 28 | 37 | 0 | 22 | 14.62537231 | 0 | 0 | 0 | -1 | 12 | 0 | 0 | 8 |
| 29 | 58 | 0 | 32 | 17.03917056 | 0 | 0 | 0 | -1 | 13 | 0 | 0 | 9 |
| 30 | 42 | 0 | 23 | 11.65280562 | 0 | 0 | 0 | -1 | 12 | 0 | 0 | 10 |
| 31 | 60 | 0 | 46 | 14.82982415 | 1 | 1 | 1 | -1 | 19 | 0 | 1 | 7 |
| 32 | 50 | 14 | 24 | 7.553933066 | 1 | 1 | 1 | -1 | 20 | 0 | 1 | 11 |
| 33 | 44 | 0 | 22 | 18.1042436 | 0 | 0 | 0 | -1 | 11 | 0 | 0 | 7 |
| 34 | 58 | 19 | 32 | 16.77200048 | 0 | 0 | 0 | -1 | 5 | 0 | 0 | 6 |
| 35 | 95 | 0 | 47 | 23.60530274 | 4 | 4 | 4 | -1 | 27 | 0 | 1 | 9 |
| 36 | 162 | 15 | 64 | 60.52520136 | 0 | 0 | 0 | -1 | 5 | 0 | 0 | 4 |
| 37 | 77 | 0 | 27 | 18.60506049 | 2 | 1 | 1 | 0.577350269 | 27 | 0 | 3 | 6 |
| 38 | 24 | 0 | 10 | 11.36331818 | 0 | 0 | 0 | -1 | 8 | 0 | 0 | 8 |
| 39 | 21 | 21 | 21 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 40 | 38 | 38 | 38 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 41 | 162 | 0 | 47 | 62.65185189 | 0 | 0 | 0 | -1 | 11 | 0 | 0 | 10 |
| 42 | 75 | 0 | 39 | 21.46771151 | 0 | 0 | 0 | -1 | 14 | 0 | 0 | 9 |
| 43 | 64 | 17 | 26 | 12.24837652 | 0 | 0 | 0 | -1 | 12 | 0 | 0 | 7 |
| 44 | 90 | 0 | 37 | 22.99721276 | 0 | 0 | 0 | -1 | 27 | 0 | 0 | 9 |
| 45 | 52 | 0 | 24 | 11.95879112 | 0 | 0 | 0 | -1 | 44 | 0 | 0 | 11 |

- Used Market Makers method [2]
- Reduced each patient to vector of 139 features

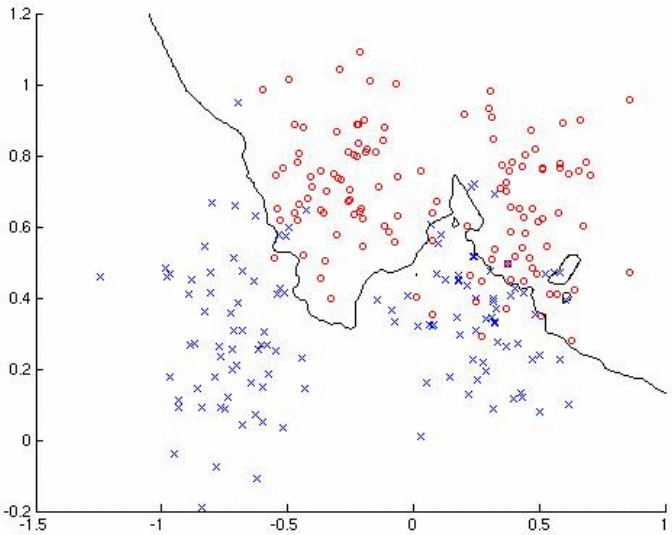
# Kitchen Sink

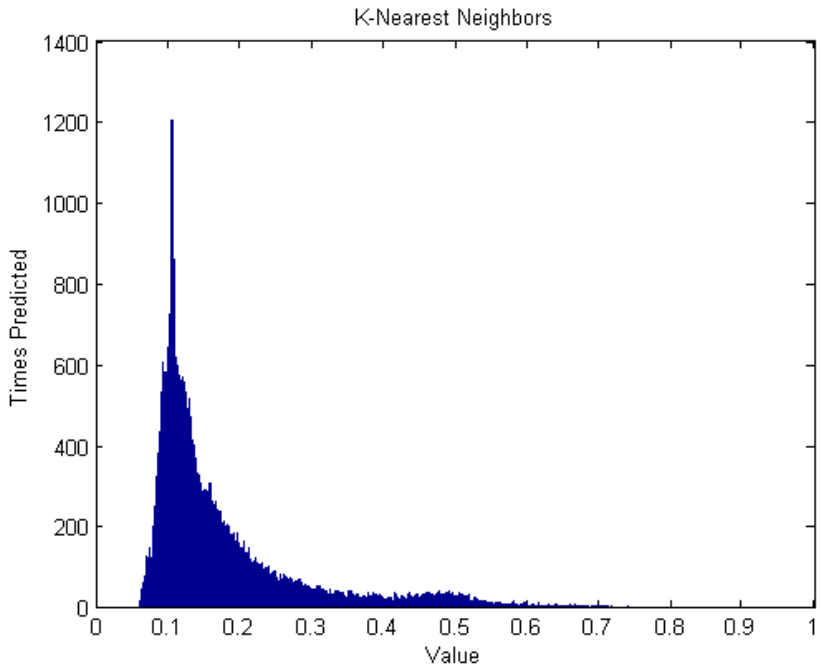
- Generate large number of features
- Need method for eliminating useless features
  - Compare distributions of feature in test and training sets
  - Basis pursuit

# Basis Pursuit

- Find best generated features (reduces error the most)
- Add to feature set
- Repeat until enough useful features are found

# K-Nearest Neighbors

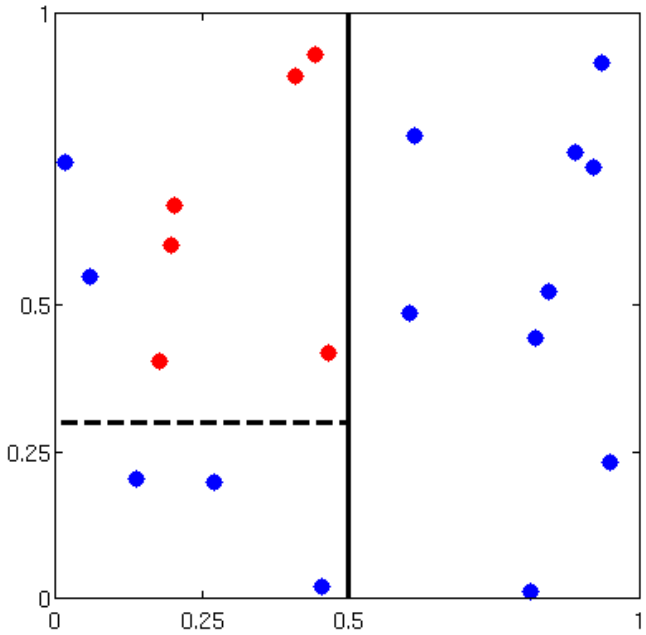


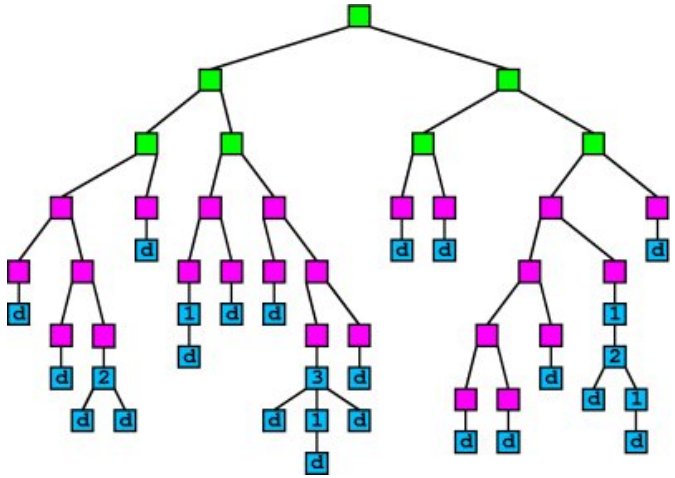




- Weighted average of closest neighbors
- Very slow

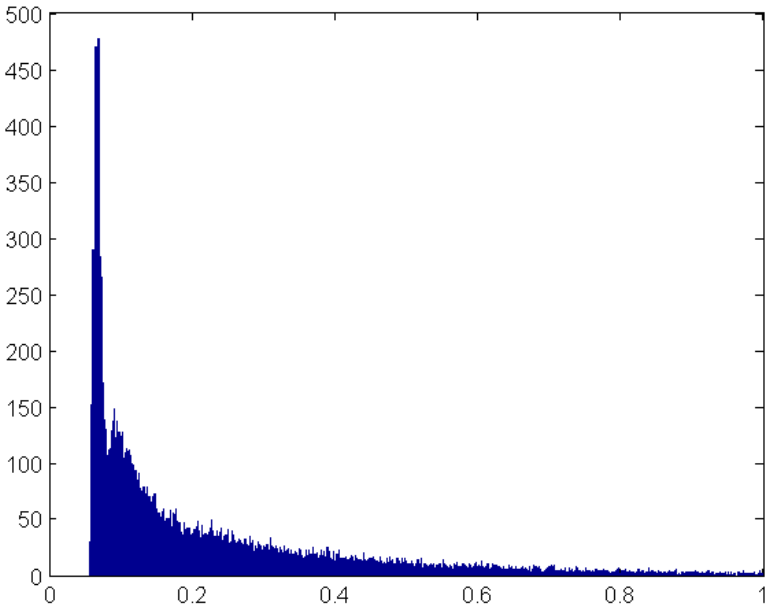
Neighbors: $k = 1000$
RMSLE: 0.475197
(996th place)

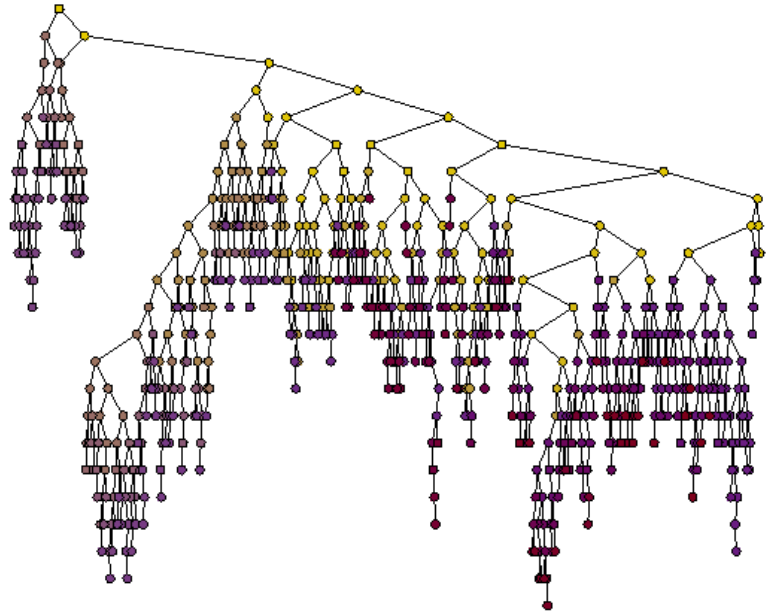
# Decision Trees



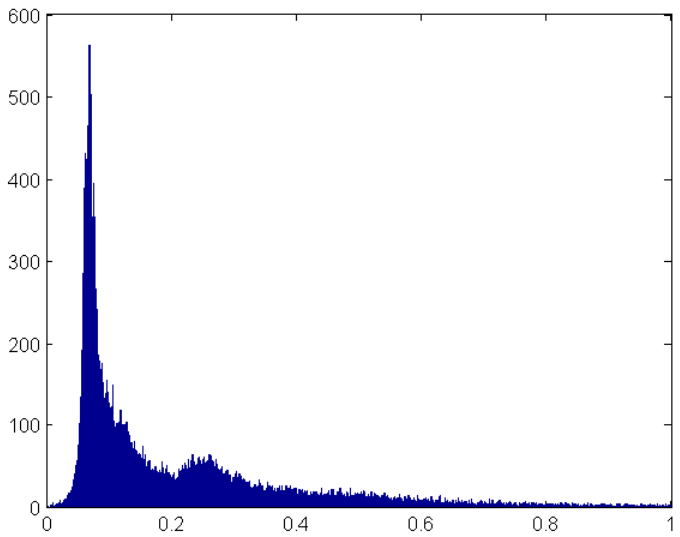
# Random Forests

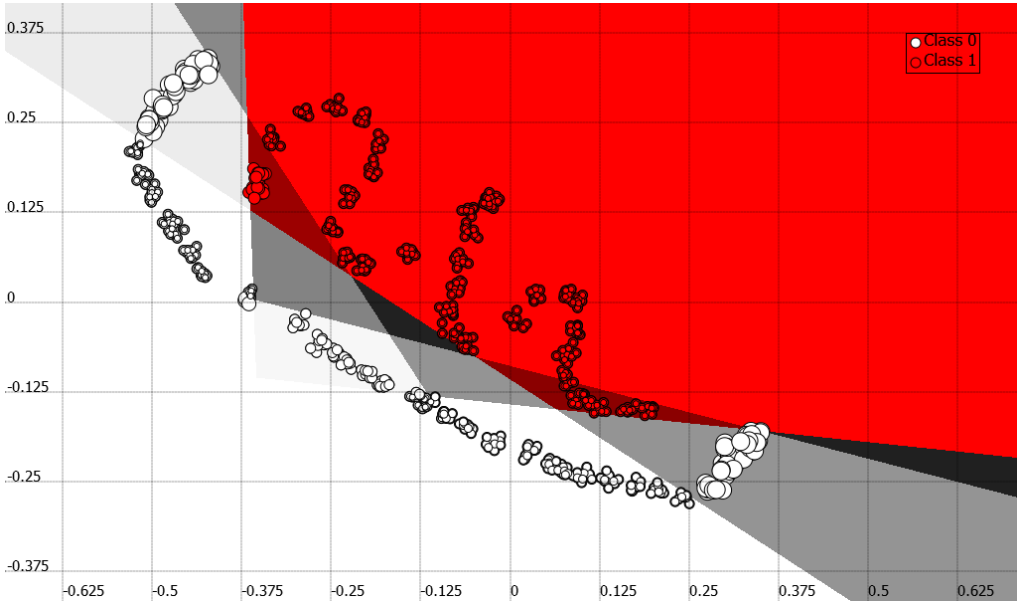


RMSLE: 0.464918
(469th place)

# Gradient Boosting Machines




Trees = 8000

Shrinkage = 0.002

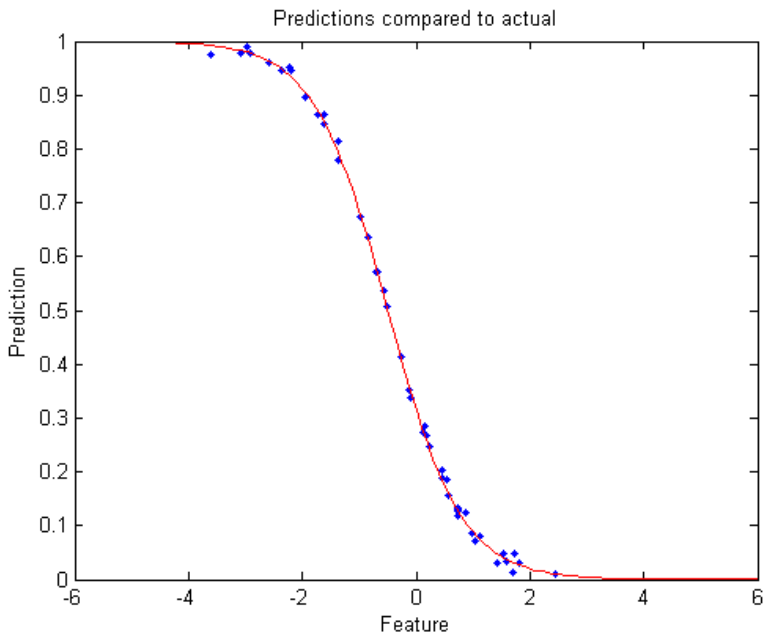
Depth = 7

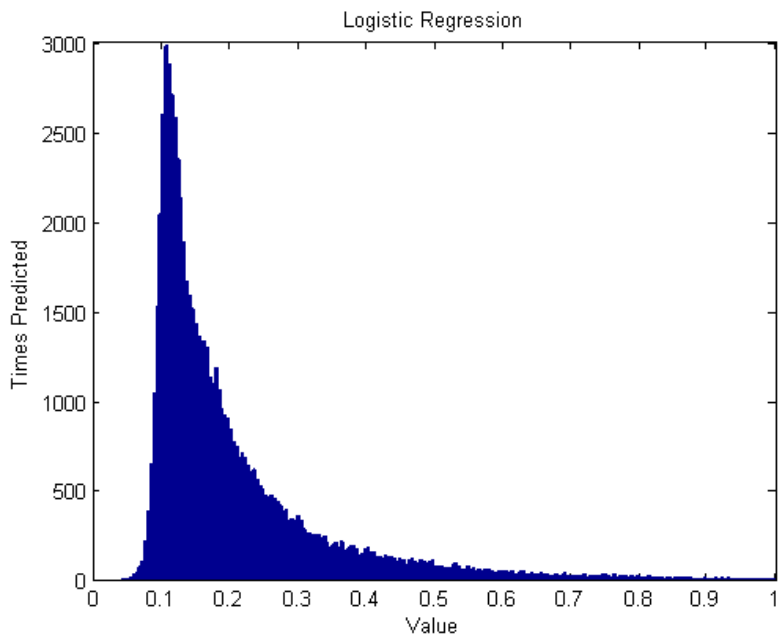
Minimum Observations = 100

RMSLE: 0.462998

(325th place)

# Logistic Regression

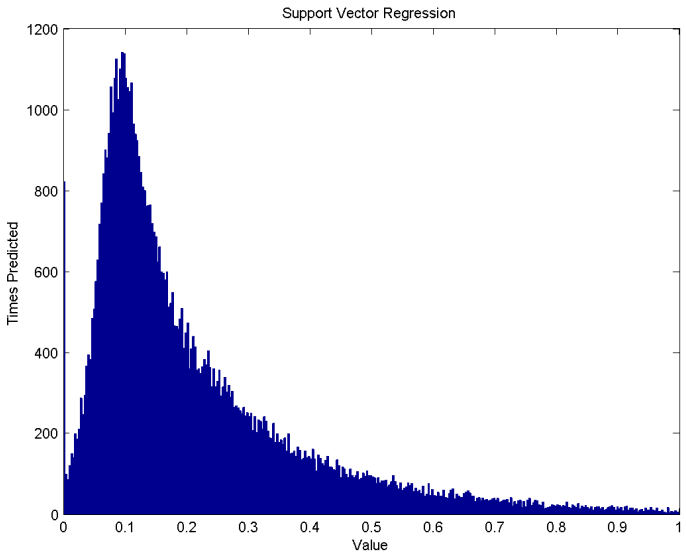


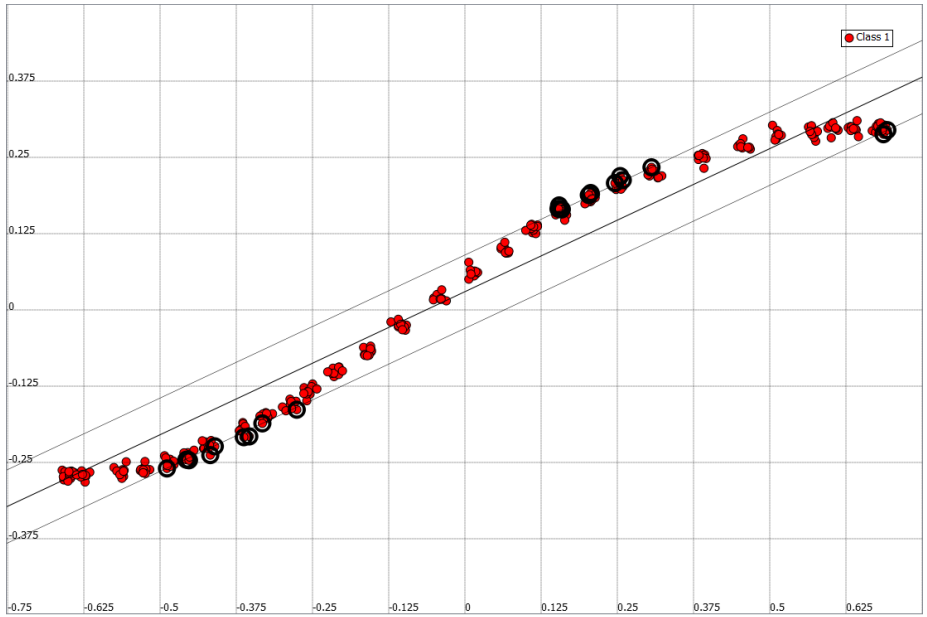




- Optimized with gradient descent

RMSLE: 0.466726
(580th place)

# Support Vector Regression



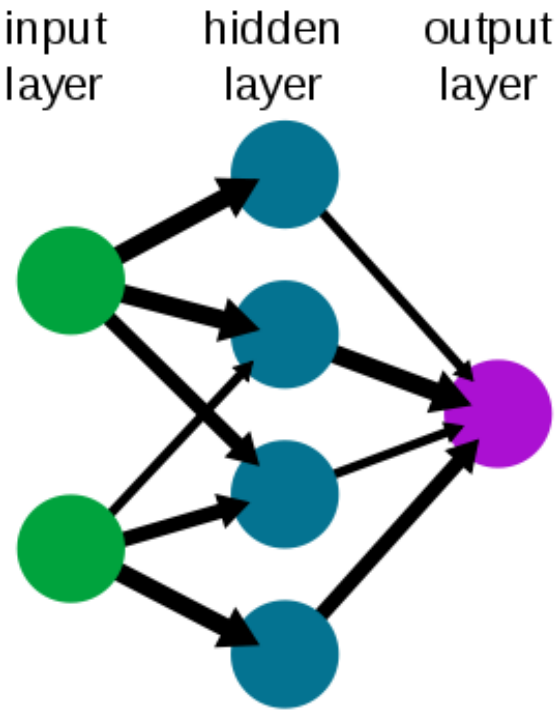



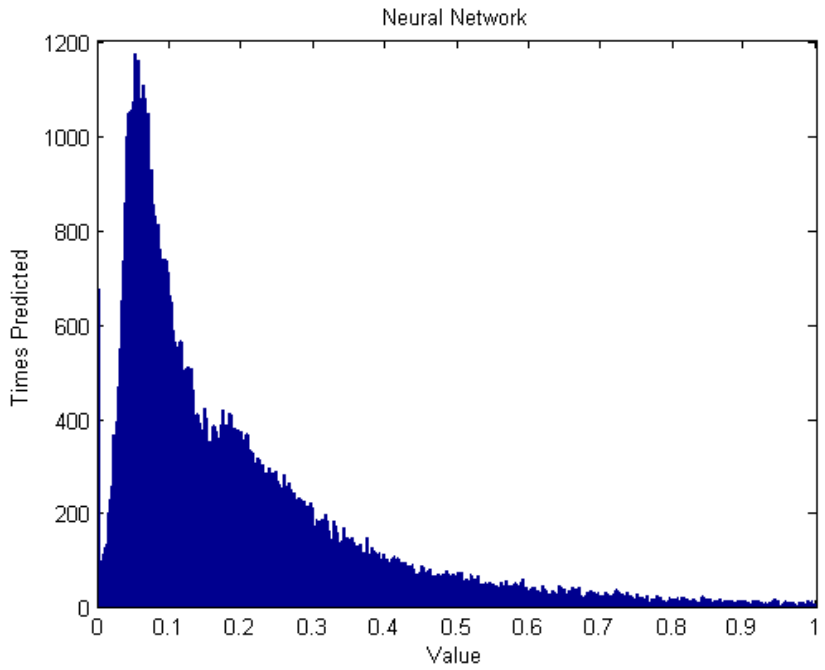

$\varepsilon = .02$

RMSLE: 0.467152

(629th place)

# Neural Networks




Number of hidden neurons = 7

Number of cycles = 3000

RMSLE: 0.465705

(511th place)

# Individual Predictors (Summary)

| | |
|---|---|
| K-Nearest Neighbors | 0.475197 (996th place) |
| Support Vector Regression | 0.467152 (629th place) |
| Logistic Regression | 0.466726 (580th place) |
| Neural Networks | 0.465705 (511th place) |
| Random Forests | 0.464918 (469th place) |
| Gradient Boosting Machines | 0.462998 (325th place) |

# The Blending Equation

X as a combination of predictors

$$\tilde{X} = Xw$$

Minimize cost function

$$C = \frac{1}{n}\sum_{i=1}^{N}(Y_i - \tilde{X}_i)^2$$

# The Blending Equation

Optimizing predictors' weights

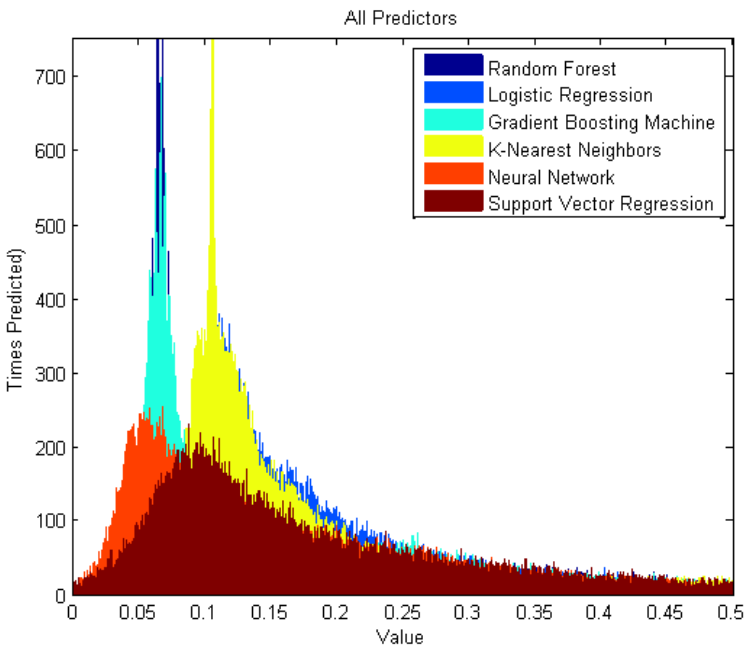
$$w_c = (Y^T X)(X^T X)^{-1}$$

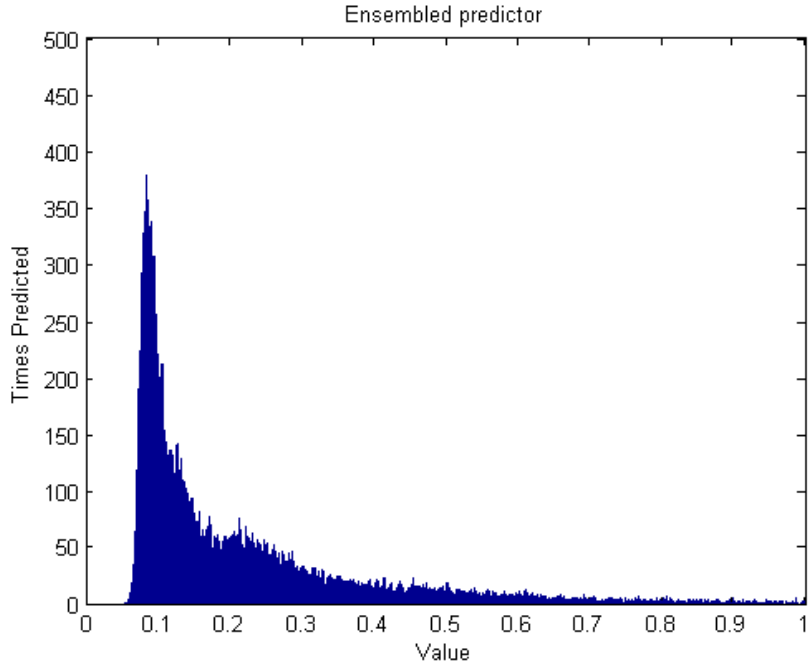
$$Y^T X = \sum_i X_{ic}^2 + \sum_{i=1}^{n} (X_i - Y_i)^2 - \sum_{i=1}^{n} Y_i^2$$

$$Y^T X = \sum_i X_{ic}^2 + n\varepsilon_0^2 - n\varepsilon_c^2$$

Y = Actual values (Unknown)
X = Our predictions (Known)
$\varepsilon$ = Feedback (Known)

# Blending Results




RMSLE: 0.461432

88th place on the final milestone leaderboard

# Results(Summary)

| | |
|---|---|
| K-Nearest Neighbors | 0.475197 (996th place) |
| Support Vector Regression | 0.467152 (629th place) |
| Logistic Regression | 0.466726 (580th place) |
| Neural Networks | 0.465705 (511th place) |
| Random Forests | 0.464918 (469th place) |
| Gradient Boosting Machines | 0.462998 (325th place) |
| **Blending** | **0.461432 (88th place)** |

# Observations and Problems

- Fewer repeated classifiers worked better
- Overfitting based on feedback
- Test and Training data were inconsistent

# Future Work

- Optimizing Blending Equation with Regularization Constant

$$w_c = (Y^T X)(X^T X + \lambda I)^{-1}$$

- More predictors
- Adjust for changes over time

# Questions

# References



Heritage provider network health prize, 2012.
http://www.heritagehealthprize.com/c/hhp.


David Vogel Phil Brierley and Randy Axelrod.
Market makers - milestone 1 description.
September 2011.


Andreas Töscher and Michael Jahrer.
The bigchaos solution to the netflix grand prize.
September 2009.