Insurance Programs
and Analytic Services

Milliman

The Kaggle Challenge

Dmitriy Guller, ACAS
Actuarial Associate Sr., Modeling Division, ISO

Mark Goldburd, FCAS, MAAA
Consulting Actuary, Milliman

Outline

- Description of the Competition
- Model Building
- Lessons Learned

What is Kaggle?

- World's largest community of data scientists (220,000+ members)
- Crowdsourcing of predictive modeling problems
  - Many predictive modelers competing with each other may come up with a better model than domain experts
- Host of competitions to solve complex data science problems
  - Cover many different fields
  - Few insurance-related challenges
  - For people new to data science, beginner challenges to get them started
- Competition sponsors post a problem and related datasets
- Players submit predictions and are ranked by some objective function
- Top finishers often get a prize

Liberty Mutual Competition

- Predict expected fire losses for insurance policies
  - Significant portion of total property losses
  - Low frequency and high severity
- Objective function: maximize weighted Gini on the test dataset
- Ultimately 634 teams participated
  - Competition open to Liberty Mutual employees for training purposes

Model Building

Insurance Programs
and Analytic Services

Milliman

Description of Data

Training Set

- **452,061** policy records
- **1,188** have claims (0.26%)

Test set

- **450,728** policy records
- **???** claims

random 50% used to rank competitors on the **public leaderboard**

other 50% used to determine **final standings** (not revealed until the end of the competition)

- **Target variable**: transformed ratio of losses to amount of insurance
- Total of 300 anonymized potential predictor variables:

Basic Insurance Variables (17)
- 9 categorical
- 8 continuous

var1, var2, var3, ..., var17

Crime Variables (10)

crime1, crime2, crime3, ..., crime10

Geodemographic Variables (37)

geodem1, geodem2, geodem3, ..., geodem37

Weather Variables (236)

weather1, weather2, weather3, ..., weather236

Generalized Linear Models

- Standard statistical method
  - Commonly used in the industry for class plan analysis
- Many model runs
  - Find significant variables
  - Transformations and bucketing
- Best GLM model:
  - Pure premium, using Tweedie (p = 1.5) distribution
  - 14 variables, 22 degrees of freedom
  - Mostly basic insurance variables
- Public leaderboard Gini of 0.40023 – **44th** place

Generalized Linear Mixed Models

- **Challenge**: ‘Basic’ insurance variables included categorical variables with many levels – many of which had little credibility
- **Solution**: Use GLMM -- extension of GLM
- Introduces ‘random effects’ in addition to ‘fixed effects’
    - Fixed effects are fully credible variables
    - Random effects are variables to which credibility is applied
- Integrated GLM and credibility framework
- Public leaderboard Gini of 0.41387 – **23rd** place

Ensembling

- Combining information from multiple models to come up with better estimates
- For most of the competition, we worked on our own models separately
  - Formed team with two weeks to go
- Different approaches taken
  - Frequency/severity modeling using GLMM – best model had Gini of 0.40518
  - Pure premium modeling using GLMM – best model had Gini of 0.41387
- What if we just took predictions from both models and averaged them?
- Public leaderboard Gini of 0.41904 – **16th** place

Elastic Net

Regular GLM

Minimizes Total Deviance

Elastic Net GLM

Minimizes Total Deviance subject to a penalty term for size of coefficient estimates

Example using OLS:

$$SSE_{EN} = \underbrace{\sum_{i}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2}_{\text{SSE}} + \lambda \underbrace{\left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2 \right)}_{\text{Penalty}}$$

Tuning parameter ‘lambda’ controls size of penalty

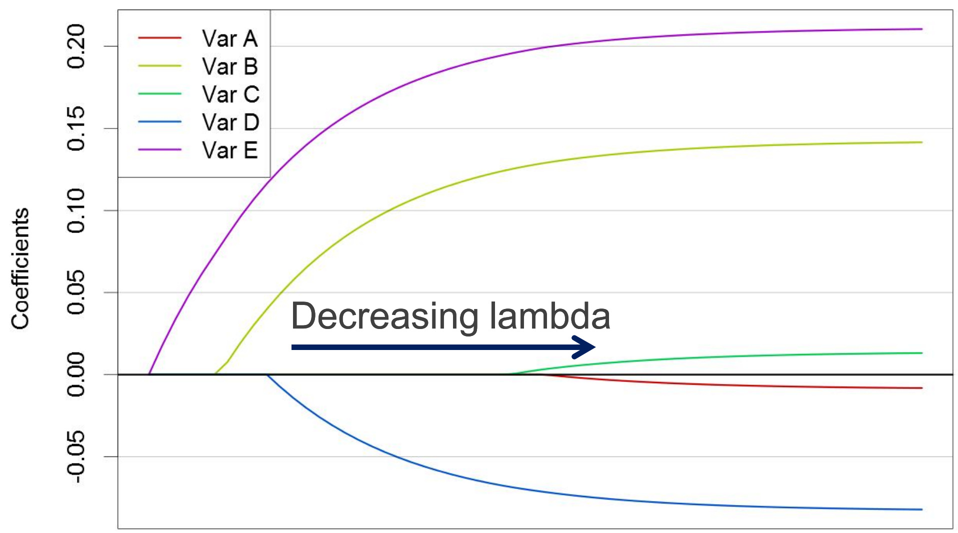
Elastic Net

Regular GLM

Minimizes Total Deviance

Elastic Net GLM

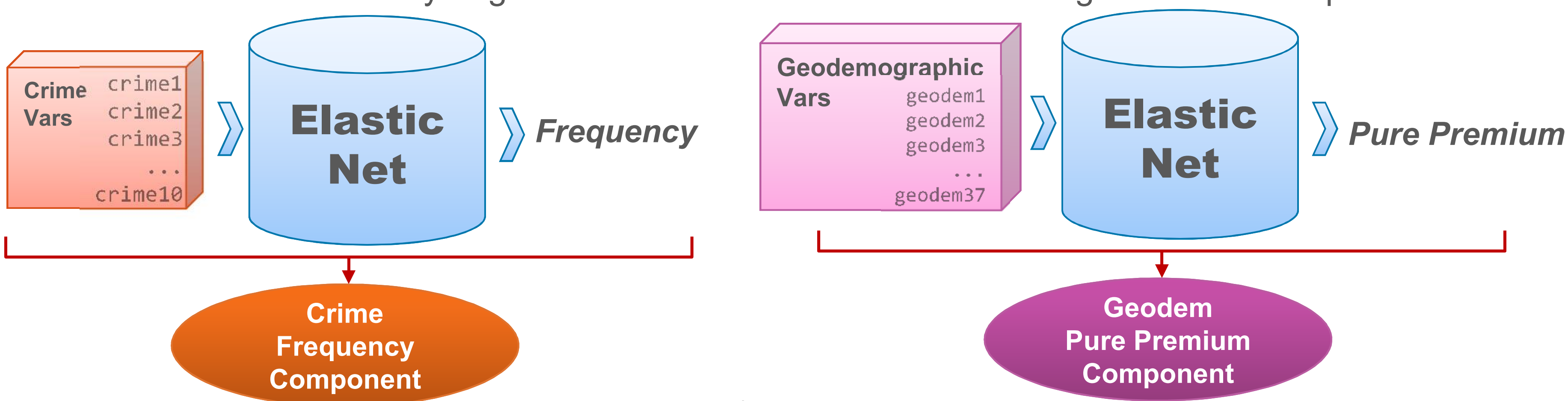
Minimizes Total Deviance subject to a penalty term for size of coefficient estimates



- Compared to GLM, elastic net has worse fit on training data and better fit on holdout data
- Just like GLMM, elastic net can be thought of as GLM with credibility

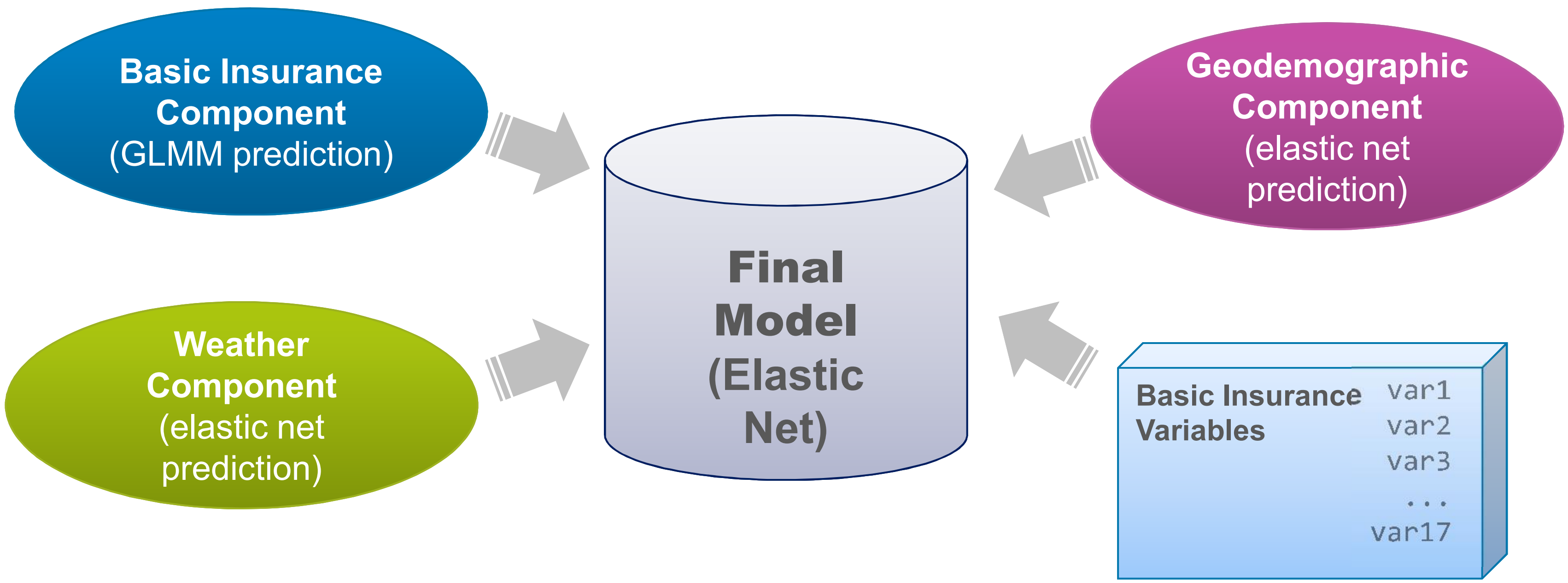
'Components' from elastic net

- **Challenge**: data included large number of crime, weather and geodemographic variables
  - Little informational value
  - Many were **highly correlated**
- **Solution**: create 'components' using elastic net models
  - Crime, geodemographic, and weather variables were all used in isolation to predict the target variable
  - Elastic nets were used to create pure premium, frequency, and severity components
  - Combined a very large number of variables into several single variable components



Final Model

- All the components were combined with another elastic net
- Not a true multivariate approach
  - There is risk of double-counting effects that were captured by components
  - All basic insurance variables were put into the combined model again to compensate



Public Leaderboard

Completed • $25,000 • 634 teams

Liberty Mutual Group - Fire Peril Loss Cost

Tue 8 Jul 2014 – Tue 2 Sep 2014 (2 months ago)

Dashboard

Public Leaderboard - Liberty Mutual Group - Fire Peril Loss Cost

This leaderboard is calculated on approximately 50% of the test data.
The final results will be based on the other 50%, so the final standings may be different.

See someone using multiple accounts?
Let us know.

#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	**Mark & Dmitriy** *	0.44776	160	Tue, 02 Sep 2014 00:16:08
2	↑5	DataRobot *	0.43285	154	Tue, 02 Sep 2014 17:18:29 (-3.1h)
3	—	Xinxin *	0.43131	152	Tue, 02 Sep 2014 14:19:23
4	—	Team Larry	0.43093	201	Sat, 30 Aug 2014 11:28:15 (-3.4d)
5	↑1	REX@CatchingFire	0.43012	174	Tue, 02 Sep 2014 23:55:27 (-3h)
6	↓4	Gauss, Anshul, and Gaurav	0.42995	149	Tue, 02 Sep 2014 18:04:02 (-3.4d)
7	↓2	Patrick Chan	0.42959	219	Tue, 02 Sep 2014 20:34:57 (-1.7h)
8	↑3	LM_Prometheus&Arcadian&CrazyHorse	0.42725	198	Tue, 02 Sep 2014 18:28:15 (-17.3h)
9	—	LM_Statistically_Speaking	0.42470	174	Tue, 02 Sep 2014 18:14:07 (-40.1h)
10	↓2	Ivanhoe	0.42434	206	Tue, 02 Sep 2014 23:28:55 (-2.8d)
11	↑27	gmilosev	0.42314	123	Tue, 02 Sep 2014 21:27:44 (-6.2h)
12	↓2	Leustagos and Titericz	0.42236	210	Tue, 02 Sep 2014 22:18:15 (-21d)

Private Leaderboard

Completed • $25,000 • 634 teams

Liberty Mutual Group - Fire Peril Loss Cost

Tue 8 Jul 2014 – Tue 2 Sep 2014 (2 months ago)

Dashboard

Private Leaderboard - Liberty Mutual Group - Fire Peril Loss Cost

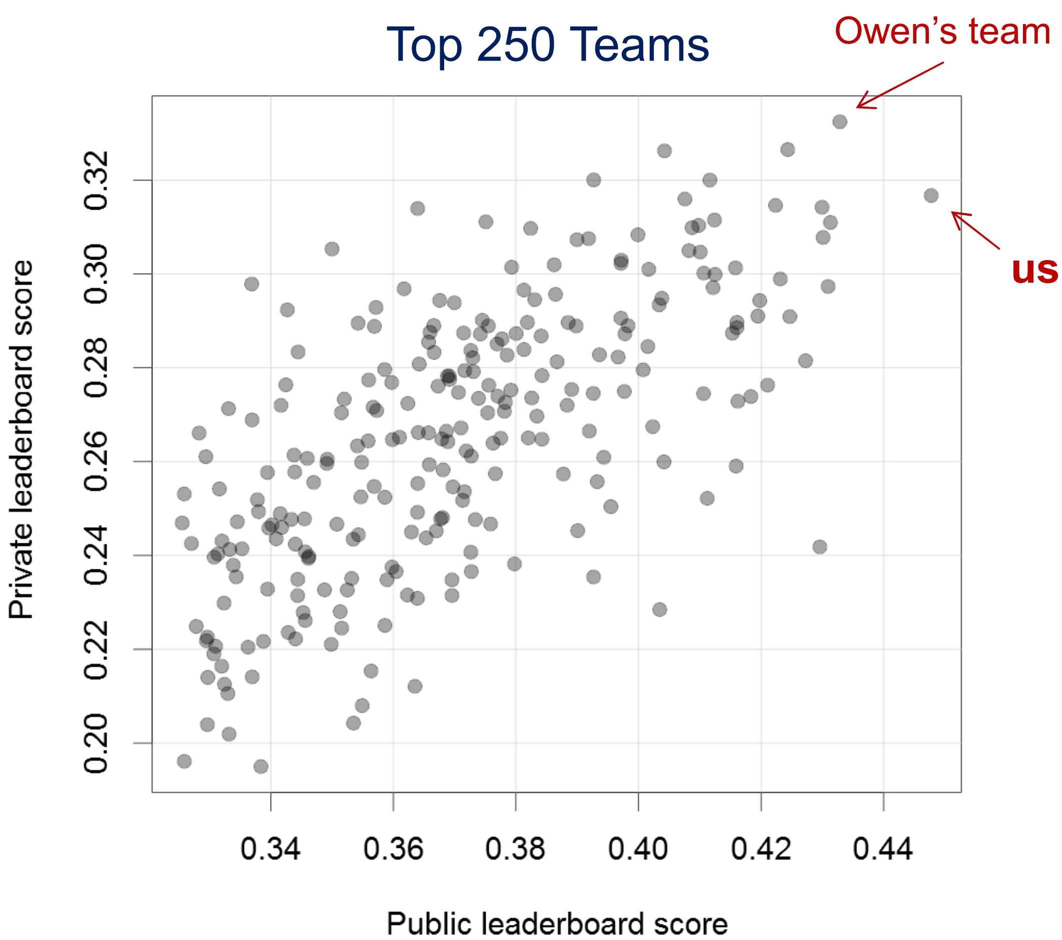
This competition has completed. This leaderboard reflects the final standings.

See someone using multiple accounts? Let us know.

#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	DataRobot *	0.33245	154	Tue, 02 Sep 2014 17:18:29 (-2.1d)
2	↓1	Ivanhoe *	0.32652	206	Tue, 02 Sep 2014 23:28:55 (-2.8d)
3	↑51	barisumog *	0.32626	40	Tue, 02 Sep 2014 08:19:53
4	—	datalab.se	0.32006	58	Sat, 02 Aug 2014 06:34:17 (-21h)
5	↑334	paulperry	0.32002	18	Tue, 02 Sep 2014 21:06:38 (-0.3h)
6	↓1	**Mark & Dmitriy**	**0.31673**	**160**	**Tue, 02 Sep 2014 00:16:08**
7	↑16	tryhard	0.31597	54	Tue, 02 Sep 2014 23:53:05 (-39h)
8	↓2	Leustagos and Titericz	0.31462	210	Tue, 02 Sep 2014 22:18:15 (-17.2d)
9	↑5	Gauss, Anshul, and Gaurav	0.31423	149	Tue, 02 Sep 2014 18:04:02 (-3.4d)
10	↑587	n_m	0.31396	8	Tue, 02 Sep 2014 16:20:34 (-1.1h)
11	↓4	backdoor	0.31149	47	Tue, 02 Sep 2014 09:29:15 (-20.9d)
12	↓4	Michael 2	0.31112	31	Tue, 02 Sep 2014 20:33:19 (-9.2d)

Public vs Private Leaderboard

Public Rank	Private Rank
1	6
2	1
3	13
4	34
5	18
6	9
7	210
8	82
9	48
10	2



Lessons Learned

Insurance Programs and Analytic Services

Milliman

Proper Cross-Validation is Crucial

- Used to measure predictive performance of model
- Often we don't have enough data to split into training and test sets

Data:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Procedure:

1. Randomly split into groups:

18 15 9 17 6	23 20 25 21 13	12 24 16 2 4	5 14 10 11 8	3 22 7 1 19

2. train on / test on

test on	train on	train on	train on	train on
18 15 9 17 6	23 20 25 21 13	12 24 16 2 4	5 14 10 11 8	3 22 7 1 19

3. Repeat for remaining groups:

18 15 9 17 6	23 20 25 21 13	12 24 16 2 4	5 14 10 11 8	3 22 7 1 19
18 15 9 17 6	23 20 25 21 13	12 24 16 2 4	5 14 10 11 8	3 22 7 1 19
18 15 9 17 6	23 20 25 21 13	12 24 16 2 4	5 14 10 11 8	3 22 7 1 19
18 15 9 17 6	23 20 25 21 13	12 24 16 2 4	5 14 10 11 8	3 22 7 1 19

Proper Cross-Validation is Crucial

- Plan out before model building
- Encompass all model building steps
- Preferable to train/test split
  - May do both – split first, then cross-validate inside the training dataset
- Lack of cross-validation can leave you flying blind
- Helps prevent “overfitting to public leaderboard” phenomenon

GLM: Simplistic But Not Simple

- One of the less powerful statistical learning methods
- Linear model
  - Most phenomena are non-linear
  - Interactions and transformations are an imperfect solution
- Requires a lot of manual fine-tuning
  - Makes proper cross-validation very difficult

Elastic Net

- Elastic net can do GLM better
  - GLM is a special case of elastic net
- Many non-obvious improvements over GLM
  - Shrinks coefficients towards grand mean, just like credibility procedure
  - When α is zero (ridge regression), there is direct connection to Buhlmann-Straub credibility method
  - When α > 0, variable selection is automatically performed

Try Many Approaches

- At the onset of a modeling project, it is difficult to know which approach or method will be optimal

- Some things that didn't work for us:
  - Principal Component Analysis
  - MARS Models
  - Tree-based learning methods (e.g., Random Forest)

Sometimes It's Not Either/Or

- When the choice is between using one method or the other, the optimal answer may be using one method on top of another
- Fitting a model to the residuals of the other model is called **boosting**
    - The second model can fill in where the first model systematically misses
- **Ensembling** is another way to combine multiple models
    - Instead of fitting one model on top of another, different model predictions are averaged in some way
- Our top model was always a straight average of pure premium and frequency/severity model

Competitions are Educational

Competition Platforms:

kaggle

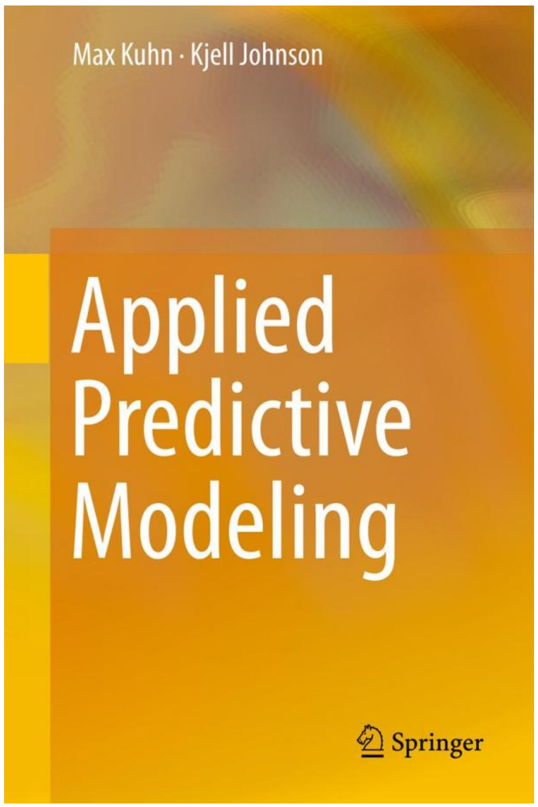
TUNEDIT

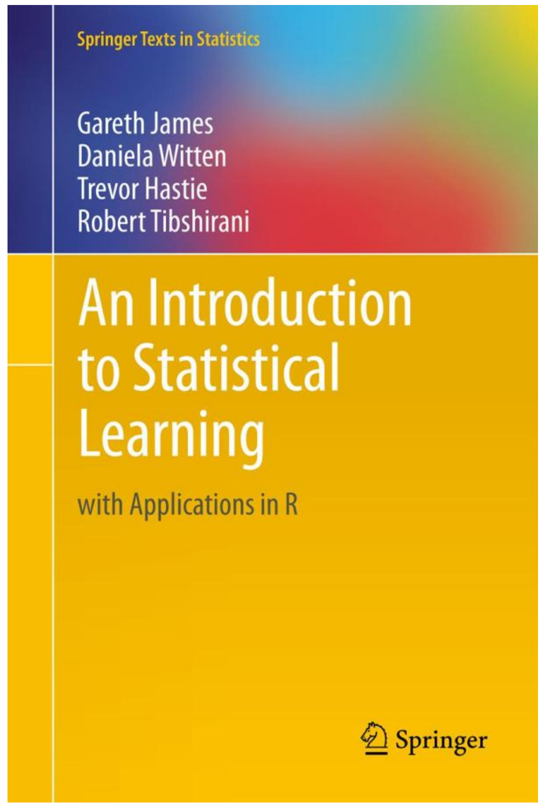
CrowdANALYTIX

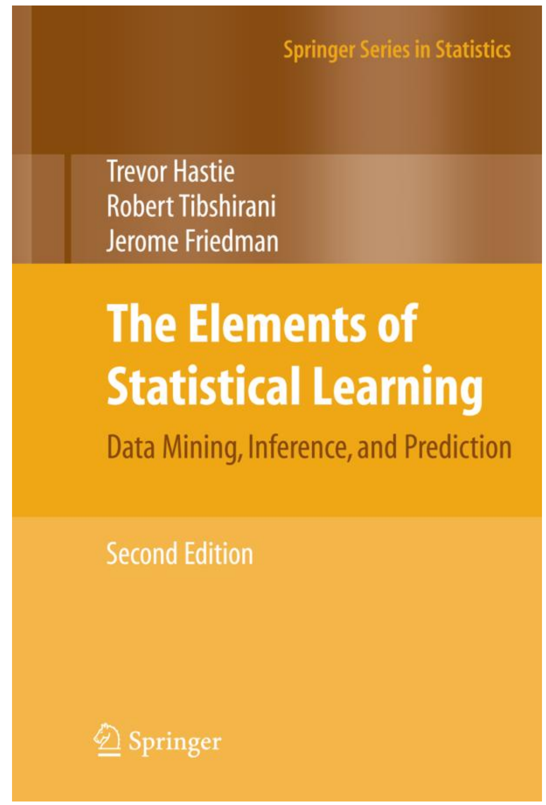
INNOCENTIVE®

- Clearly defined goals
- Competitive drive to come up with a better idea
- Instant feedback
- Learn from the best
- Bring back fresh ideas to work

Very Helpful Books







References

- GLM
  - Anderson, D. et. al. (2007) *A Practitioner's Guide to Generalized Linear Models.*
  - Frees, Edward W. et. al. (2014) *Predictive Modeling Applications in Actuarial Science.*
- GLMM
  - Klinker, F. (2010) "Generalized Linear Mixed Models for Ratemaking: A Means of Introducing Credibility into a Generalized Linear Model Setting," https://www.casact.org/pubs/forum/11wforumpt2/Klinker.pdf.
- R and R Studio
  - http://cran.r-project.org/
  - http://www.rstudio.com/