Motivation and pre-requisites

Jeffrey Leek

May 18, 2016

About this course

- This course covers the basic ideas behind machine learning/prediction
- Study design - training vs. test sets
- Conceptual issues - out of sample error, ROC curves
- Practical implementation - the caret package
- What this course depends on
- The Data Scientist's Toolbox
- R Programming
- What would be useful
- Exploratory analysis
- Reporting Data and Reproducible Research
- Regression models

Who predicts?

- Local governments -> pension payments
- Google -> whether you will click on an ad
- Amazon -> what movies you will watch
- Insurance companies -> what your risk of death is
- Johns Hopkins -> who will succeed in their programs

Why predict? Glory!

http://www.zimbio.com/photos/Chris+Volinsky

Why predict? Riches!

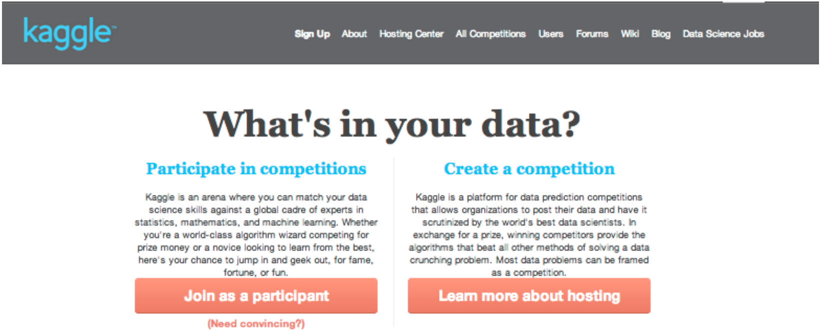
Improve Healthcare,
Win $3,000,000.

COMPETITION GOAL

Identify patients who will be admitted to a hospital within the next year, using historical claims data.

`http://www.heritagehealthprize.com/c/hhp`

Why predict? For sport!



What's in your data?

Participate in competitions

Kaggle is an arena where you can match your data science skills against a global cadre of experts in statistics, mathematics, and machine learning. Whether you're a world-class algorithm wizard competing for prize money or a novice looking to learn from the best, here's your chance to jump in and geek out, for fame, fortune, or fun.

Create a competition

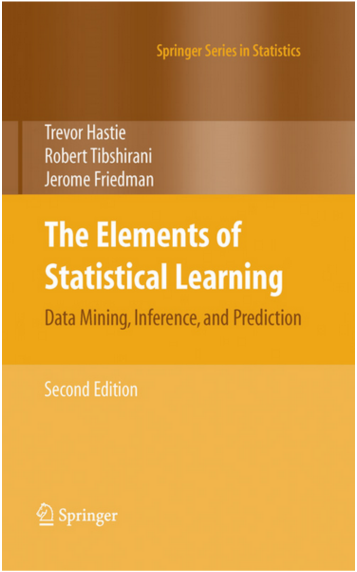
Kaggle is a platform for data prediction competitions that allows organizations to post their data and have it scrutinized by the world's best data scientists. In exchange for a prize, winning competitors provide the algorithms that beat all other methods of solving a data crunching problem. Most data problems can be framed as a competition.

`http://www.kaggle.com/`

Why predict? To save lives!

`http://www.oncotypedx.com/en-US/Home`

A useful (if a bit advanced) book



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

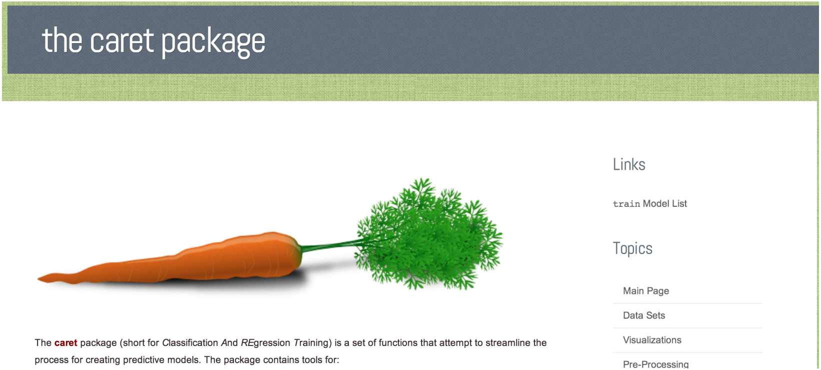
Data Mining, Inference, and Prediction

Second Edition

Springer

The elements of statistical learning

A useful package



the caret package

Links

`train` Model List

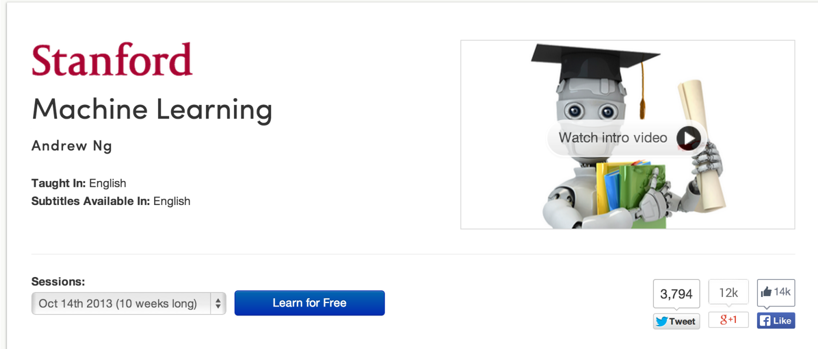
Topics

- Main Page
- Data Sets
- Visualizations
- Pre-Processing

The **caret** package (short for *C*lassification *A*nd *RE*gression *T*raining) is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

http://caret.r-forge.r-project.org/

Machine learning (more advanced material)



Stanford

Machine Learning

Andrew Ng

Taught In: English

Subtitles Available In: English

Sessions:

Oct 14th 2013 (10 weeks long)

https://www.coursera.org/course/ml

Even more resources

- List of machine learning resources on Quora
- List of machine learning resources from Science
- Advanced notes from MIT open courseware
- Advanced notes from CMU
- Kaggle - machine learning competitions