Learning from the Best -- Kaggle Best Practices

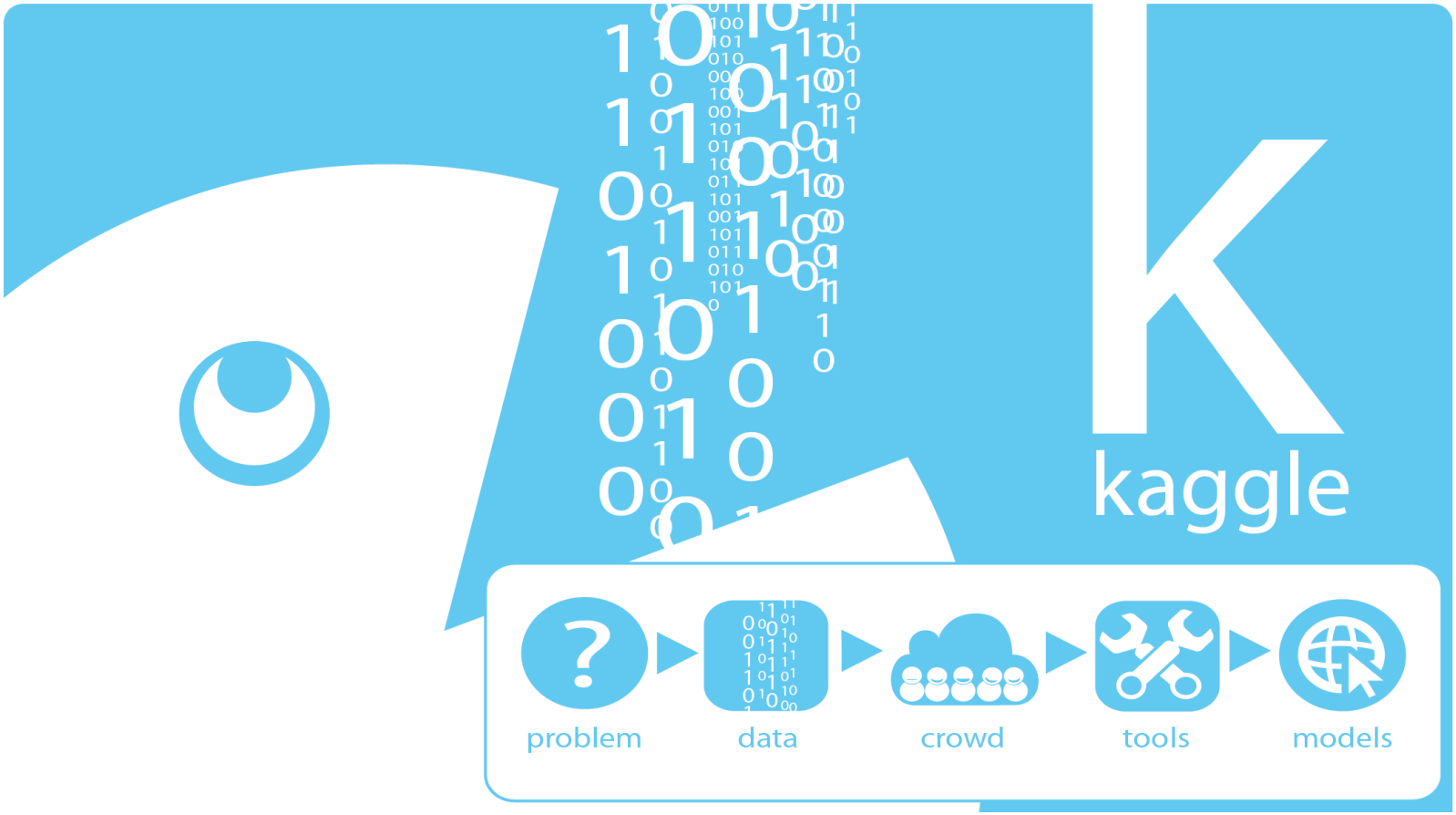


Image: http://www.36dsj.com/wp-content/uploads/2016/06/1510.jpg

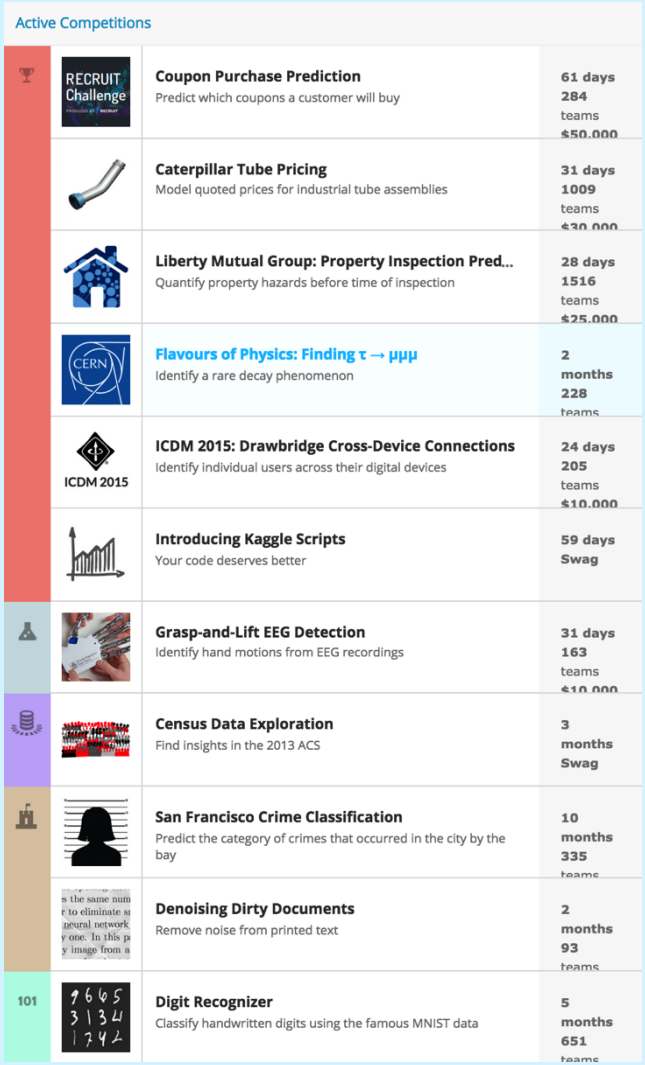
Cork Big Data Analytics, 2015-08-15
Johannes Ahlmann

Disclaimer

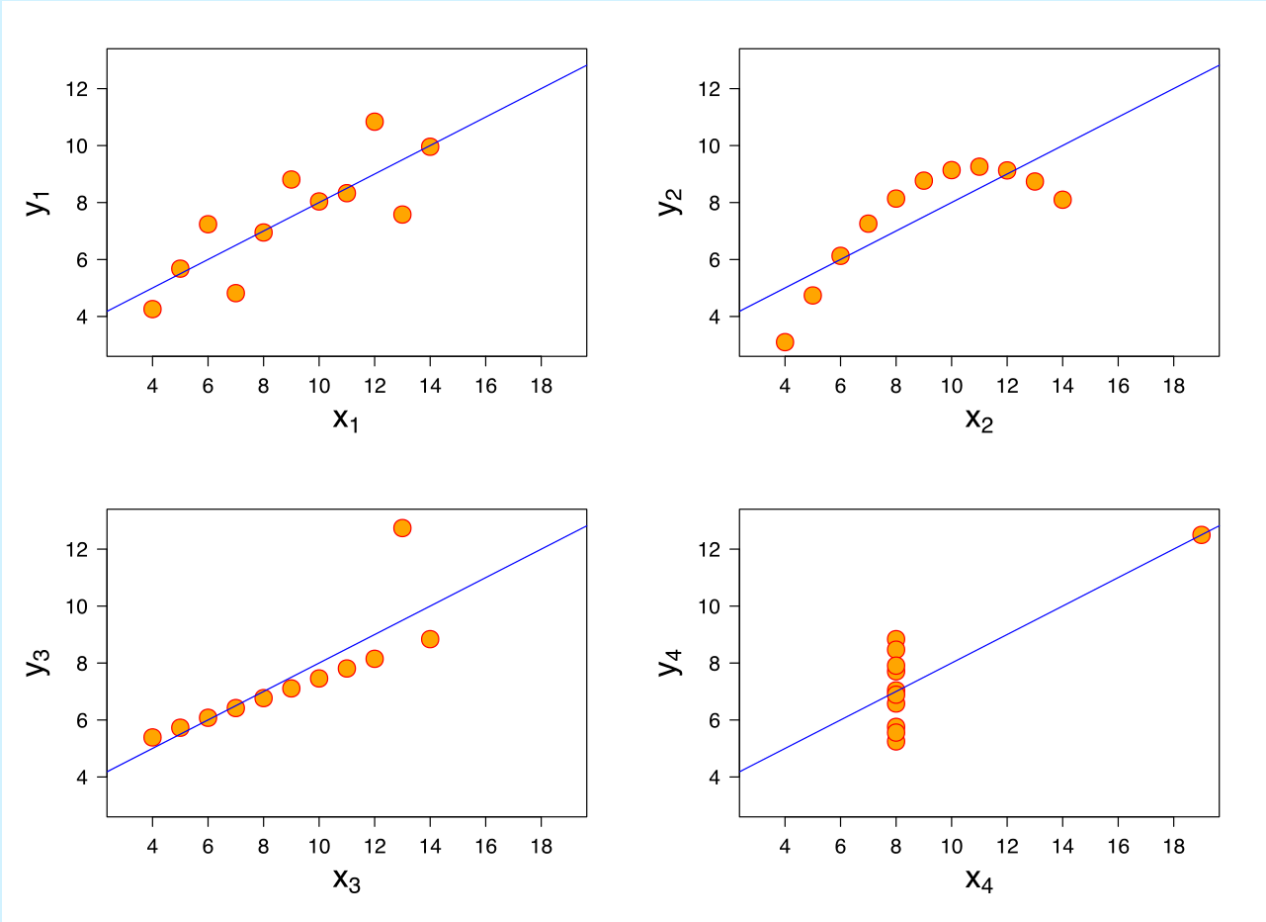
- Insights from Kaggle winners or near-winners
- Data Analytics is a huge field, this can only be a small, Kaggle-specific view
- Much of the structure is “borrowed” from David Kofoed Wind’s blog post and Thesis on Kaggle Best Practices

Kaggle

- Platform for predictive modelling and analytics competitions
- Open to anyone
- “Crowdsourcing”
- Automated scoring
- Leaderboard
- Public, private and competitions with awards



Don’t rely on simple Metrics



- mean(x) = 9
- mean(y) = 7.50
- variance(x) = 11
- variance(y) = 4.1
- correlation = 0.816
- lm = 3.00 + 0.500x

We need to remind ourselves; over and over again

It is so easy to become complacent!

Image: https://upload.wikimedia.org/wikipedia/commons/thumb/e/ec/Anscombe%27s_quartet_3.svg

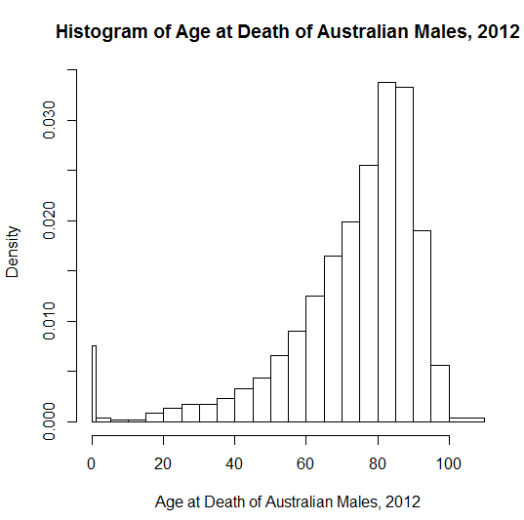
Get to know the Data

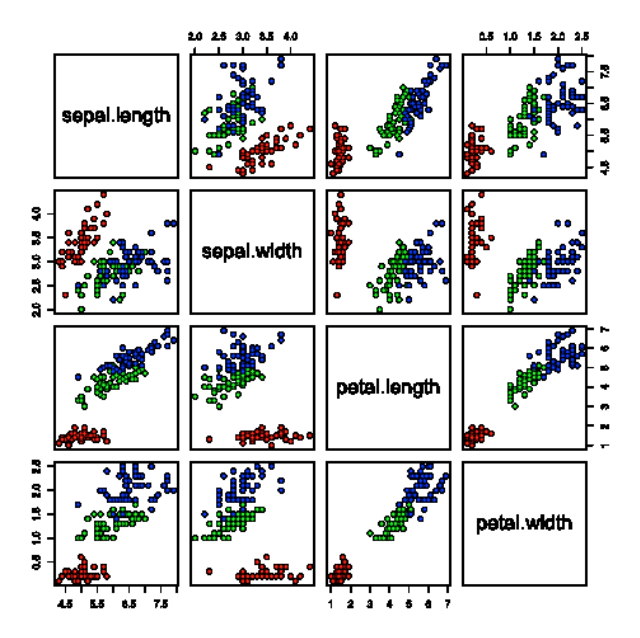
Visualize the Data

- how can we visualize 5 dimensions? 2000?
- simple metrics are not enough
- start feature-by-feature, or pair-wise

Understand the Shape and Patterns of the Data

- what do the attributes mean? how are they related?
- skew
- scale
- factors (“man”, “woman”)
- ordinals (“good”, “better”, “best”)
- missing data, data inconsistencies
- shape
- “step-functions”
- “outliers”?
- structural differences between train and test set





“Feature Engineering is the most important Part”

- Most kagglers use same few algorithms (logistic regression, random forest, gbm)
- Subject matter expertise often not a huge factor
- Err on the side of too many features.
  Thousands of features usually not a problem
- Examples
  - pairwise: a-b, a/b, a*b, 1/a, log(a), |a|
  - date => weekday, day of month, time
  - GPS locations => velocity, acceleration, angular acceleration, segment into stops, segment into accelerating and braking phases, mean/median/stddev/centiles/min/max, etc.
  - text => ngrams, tokenize, stemming, stopwords

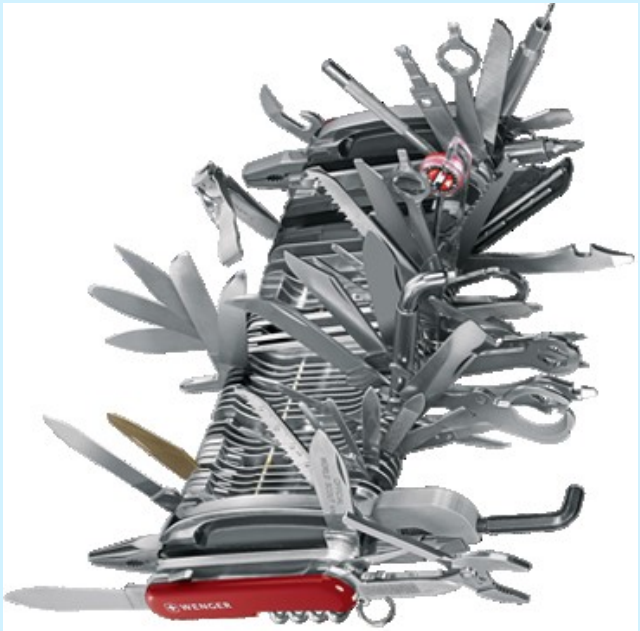


Image: https://content.etilize.com/images/300/300/1017951585.jpg

How the Kaggle Leaderboard works

- Public train and test data
- Secret holdout validation data
- Automated scoring
- Public leaderboard against test data
- Private leaderboard against validation data
- Final scoring is giving strong weight to validation data

#	Δ1w	Team Name * in the money	Score @	Entries	Last Submission UTC (Best – Last Submission)
1	—	Sajid Umair *	1.00000	2	Mon, 22 Jun 2015 11:46:16 (-0.1h)
2	—	ericychen	1.00000	7	Tue, 28 Jul 2015 04:14:57 (-32.1d)
3	—	PiPi 2	1.00000	2	Tue, 30 Jun 2015 21:19:14 (-0h)
4	—	Philosopher	1.00000	1	Fri, 03 Jul 2015 12:01:30
5	—	NP-hardly	1.00000	12	Wed, 15 Jul 2015 02:10:47 (-5.8d)
6	new	Yanzheng	1.00000	2	Wed, 29 Jul 2015 08:50:05
7	↓1	Boyuan Zhang	0.99522	13	Tue, 02 Jun 2015 15:58:13 (-0h)
8	↓1	nicolas gaude	0.99522	1	Sun, 28 Jun 2015 03:36:34
9	↓1	oussama absi	0.99043	1	Tue, 21 Jul 2015 15:08:49
10	↓1	mohit midha	0.93301	10	Tue, 30 Jun 2015 12:24:08 (-24d)
11	↓1	Raymond229	0.91388	3	Mon, 13 Jul 2015 02:42:12
12	↓1	Anjana Agrawal	0.90909	24	Fri, 19 Jun 2015 04:03:53
13	↑1802	Ankur singh chauhan	0.90909	5	Thu, 30 Jul 2015 19:29:00
14	↓2	edj	0.89474	7	Sun, 12 Jul 2015 08:08:09
15	↓2	Andy Dingler	0.88038	6	Fri, 24 Jul 2015 18:55:06
16	↓2	LovelyRaghav	0.86124	1	Sat, 06 Jun 2015 12:56:23
17	↓2	Haja Maideen	0.84689	1	Wed, 10 Jun 2015 13:59:26
18	↓2	王航	0.84211	1	Tue, 23 Jun 2015 08:37:00
19	↓2	zqs2008	0.84211	16	Wed, 24 Jun 2015 15:25:06
20	↓2	Stephen99	0.83732	26	Mon, 06 Jul 2015 11:11:23 (-35.6h)
21	↓2	joker125	0.83732	22	Mon, 06 Jul 2015 11:14:51 (-19.7h)
22	↓2	Eon Retief	0.83732	9	Sun, 05 Jul 2015 16:27:12
23	↓2	Sayantan Raha	0.83254	3	Wed, 03 Jun 2015 05:50:01
24	new	Alessio 2	0.83254	20	Fri, 31 Jul 2015 12:23:35 (-0.3h)
25	↓3	BrettKryway	0.82775	10	Mon, 06 Jul 2015 14:51:57 (-7.5d)
26	↓3	BOAZ.	0.82775	1	Sat, 04 Jul 2015 05:08:13
27	↓3	Chiraz BenAbdelkader	0.82775	25	Thu, 09 Jul 2015 12:50:49 (-4d)

“Overfitting to the leaderboard is a real issue”

- Kaggle lets you choose two final submissions
- Strong temptation to submit dozens or hundreds of solutions and to pick the ones that are performing “best”
- This leads to “manual overfitting”
- “The most brutal way to learn about overfitting?
  Watching yourself drop hundreds of places when a @kaggle final leaderboard is revealed”
  @benhammer

Image: https://republic.ru/images2/blog_photo_18/2013_06_10/listalka/rastyagivanie.jpg

“Overfitting to the leaderboard is a real issue”

- Need strong intrinsic measure of performance from train-set alone
  - k-fold cross-validation
  - bagging
- Possible to use public leaderboard in an intelligent way to glean information or in a weighted manner with CV score
- But resist the temptation to just pick the “best” two submissions

- Sidenote: the same “manual overfitting” issue applies to hyper-parameters as well, if we are not careful

Image: http://img.sparknotes.com/content/sparklife/sparktalk/tightfitting_Large.jpg

“Simple Models can get you very far”

- “I think beginners sometimes just start to “throw” algorithms at a problem without first getting to know the data.
  I also think that beginners sometimes also go too-complex-too-soon”
  – Steve Donoho
- Start with a simple baseline
- Usually “logistic regression” or “random forest” will get you very far. And even “random forest” is far from “simple”
- Complex algorithms often run much slower, reducing speed of learning iterations
- More model parameter means more risk of overfitting, and more arduous parameter-tweaking

Image: http://www.jeffbullas.com/wp-content/uploads/2013/07/The-Power-of-Simple-Writing.jpg

“Ensembling is a winning Strategy”

- “In 8 out of the last 10 competitions, model combination and ensembling was a key part of the final submission”
- Improves accuracy at the cost of explanatory value and performance
- Do it as a last step
- Works best if the models are less correlated and of reasonably high quality; ideally ensemble different algorithmic approaches
- Another opportunity for overfitting; what data to train/test them on?
- Needs to be use in a disciplined, well-founded manner, not just ad-hoc
- Methods:
  - naive weighting
  - bagging
  - AdaBoost
  - random forest already an ensemble

Image: https://pbs.twimg.com/profile_images/3536053177/89a7cf7df33fea05522399484b7b28f9_400x400.jpeg

“Predicting the right thing is important”

- What should I be predicting
  - correct derived variable
  - correct loss function
- Metric/loss function often given on Kaggle
  - AUC
  - Gini
  - MSE, MAE
- Understand what metric underlies your favorite algorithms
- But also more subtle understanding of the independent and dependent variables
- How to translate the outcome formulation into the correct derived variable; in the face of inconsistent and noisy data

Image: https://theosophywatch.files.wordpress.com/2012/09/seeing-the-future1.jpg?w=500

Miscellaneous

- First, build a reusable pipeline and put something on the leaderboard
- Understand the subtleties of different algorithms; prefer an algorithm you understand over a shiny new one
- Perform feature selection (i.e. random forest), and plug the features back into your "favorite" tool.
  (redundant variables, some collinearity)
- Imputation of missing data (i.e. using clustering)
- "Think more, try less"
- Choose the right tool for the right job
  (Excel, SQL, R, Spark, etc.)

Image: http://static.zoonar.com/img/www_repository5/f1/09/f1/8_42d970eeb4f447d441415716a2c7b439.jpg

Thank you

Resources

- Thesis – Competitive Machine Learning
  - expand from blog post: http://blog.kaggle.com/2014/08/01/learning-from-the-best/
- http://www.quora.com/What-do-top-Kaggle-competitors-focus-on
- http://www.slideshare.net/ksankar/data-wrangling-for-kaggle-data-science-competitions
- http://www.slideshare.net/ksankar/oscon-kaggle20?related=1
- http://www.slideshare.net/OwenZhang2/winning-data-science-competitions?related=1
- http://www.slideshare.net/SebastianRaschka/nextgen-talk-022015
- Kaggle Best Practices Youtube
- http://blog.kaggle.com/2014/08/01/learning-from-the-best/

- Many more Resources and Links: https://gist.github.com/codinguncut/c4359d9bc6f36549b625