

# The Allure of a Kaggle Competition

Diane Rucker

Data^3 Conference

March 7, 2015

# What is a Kaggle competition?

Kaggle

Cagle

Haggle

Waggle

Gaggle

Kaggle

Spell-check does not  
like the word  
“Kaggle”

**Kaggle is the world's largest  
community of data scientists.**

*They compete with each other to solve complex data science problems, and the top competitors are invited to work on the most interesting and sensitive business problems from some of the world's biggest companies through Masters competitions.*

# More about Kaggle competitions...

- ▶ Anyone, at any level, can enter (and do well)
- ▶ Test your analytics skills
- ▶ Keep it simple (to start)
- ▶ Draw on a community of experts
- ▶ Share what you learn

[www.kaggle.com](https://www.kaggle.com)








Create an account to check out the competitions

# Kaggle competitions de-coded

Competitions you  
can enter now

Competitions  
for Prizes  
(\$\$\$)

Competitions  
for  
Knowledge  
(Just for Fun)

Active Competitions		
	 <b>March Machine Learning Mania 2015</b> Predict the 2015 NCAA Basketball Tournament	25 days 135 teams \$15,000
	 <b>National Data Science Bowl</b> Predict ocean health, one plankton at a time	27 days 702 teams \$175,000
	 <b>Driver Telematics Analysis</b> Use telematic data to identify a driver signature	27 days 1091 teams \$30,000
	 <b>BCI Challenge @ NER 2015</b> A spell on you if you cannot detect errors!	7.4 days 263 teams \$1,000
	 <b>Microsoft Malware Classification Challenge (B...</b> Classify malware into families based on file content and characteristics	59 days 100 teams \$16,000

Name and  
purpose of  
competition

Time left in  
competition

Number of teams  
entered

Prize amount

# 15.071x, the Analytics Edge: the MIT MOOC

You don't need to be a rocket scientist. Or even a data scientist.

Learn this stuff during the first seven weeks.

Enter a Kaggle competition for fun (and a grade).

Try your skills against the rest of the teams.

## Prerequisites:

- ▶ Basic mathematical knowledge (at a high school level).

Intro to Analytics

An Intro to using R

Clustering

Seeing the Big Picture

Linear Regression

Logistic Regression

Trees

Visualize Data



kaggle

*"We will provide you with a dataset and problem, and ask you to use all of the methods you have learned to build the best possible predictive model. (From the edX site for 15.071x)"*

# The Analytics Edge Kaggle competition: Show of Hands

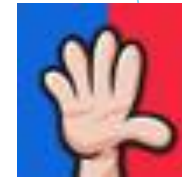
Asked & Answered

**Opinions are powerful little buggers, sometimes even more so than facts.**

Show of Hands is the fun and powerful way to peer into the brilliant and warped minds of people around the country. Answer questions daily and instantly see results sliced and diced by age, gender, income, party, geography.

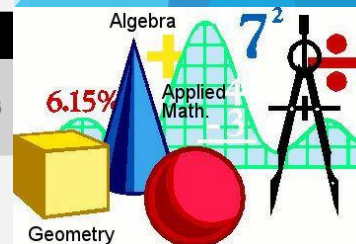
<https://www.showofhands.com/>

(or download the app)



# What types of questions might I see?

- Questions have varied topics
- Some questions have Yes/No answers
- Some questions have an Either/Or response format
- Some questions relate to personal life or activities
- Some questions relate to political or philosophical opinions



# Kaggling - A Basic Approach

## ▶ The Problem:

- ▶ **Can you predict whether people are happy?**

## ▶ Data:

- ▶ Answers to various poll questions on the app Show of Hands

## ▶ Solution:

- ▶ A model that helps you learn what predicts happiness by using informal polling questions.

## ▶ Ranking: AUC (Area Under the Curve)

- ▶ Assumes a random positive observation and negative observation
- ▶ AUC gives the proportion of the time you guess which is which **correctly**.
- ▶ Less affected by sample balance than accuracy.
- ▶ A perfect model will score an AUC of 1; Random guessing will score an AUC of around 0.5.

Other boundary conditions for the competition:

50% of data is available to use (“public data”)

Final results are based on total data set (100%)

Data is separated into “train” and “test” sets

Competition has a fixed time and fixed number of entries per day



# Digging In: A basic approach

## Explore

How familiar was I with the app that created the data? how did I look at the data? What was important? What was not?

## Create

How did I choose the models? How did I change them? What made the biggest difference? How did I generate the code?

## Do (Iterate)

how did I test my assumptions? What did I change based on the community feedback? How did my results match the public and private data?

# Exploring the Space

Explore

Create

Do  
(Iterate)

- ▶ All questions are not created equal (when it comes to predicting happiness)
- ▶ Some questions seem more predictive than others.
- ▶ Data had missing responses to both questions and biographical data; also included NA
- ▶ Some data had three response options (yes, no, maybe)
- ▶ Year of Birth (YOB) was more likely to be significant in a grouping (e.g. 20-29, 30-39)

## Approach:

- ▶ Process data to account for NA entries
- ▶ make responses numeric (1/ 0/ -1) for analysis
- ▶ Group YOB into age groups
- ▶ Separate Single / Married / Kids into different columns to determine significance

# Creating the Model

Explore

- ▶ Simple initial approach
- ▶ Assume all variables are significant until proven otherwise
- ▶ Started with basic logistic regression model
- ▶ Tested model with different assumptions based on exploring the space and the data

Create

## Approach:

- ▶ **Create logistic regression model and test**
- ▶ **Refine model based on significance of variables**
- ▶ **Test model with a selected list of “predictive questions”**
- ▶ **Refine or change approach based on AUC results on Public Data**

Do  
(Iterate)

Public Data - 50% of total dataset, available to test models. Results shown on Public Leaderboard

Final scores and rankings were based on the full dataset (Private Leaderboard)

# Iterate the Model

Explore

Create

Do  
(Iterate)

- ▶ Logistic regression model put me in the top 50 on the Public Leaderboard (!!!)
- ▶ Refining my model initially took me up to 19<sup>th</sup> place (out of 1685 teams)
- ▶ Tried various other approaches but decided to stick with logistic regression model

## Approach:

- ▶ Test probabilities of happiness based on current model
- ▶ Tweak model based on AUC results
- ▶ Rank questions in data by significance
- ▶ Submit 5 entries per day for Kaggle competition
- ▶ Best entry used for final scoring with the full dataset (Private Leaderboard)

# Why a logistic regression model in R?

- ▶ Great starting point for binary dependent variables
- ▶ Classic, effective, fairly simple
- ▶ Less complex model than other alternatives (CART, for example)
- ▶ Takes a short time to run (multiple iterations can be easily tested)
- ▶ Can quickly tell if the model will yield predictive results

One of the top final scores in this competition used a Random Forest model, which can classify lots of random data, run efficiently on large databases, and detect variable interactions.

Although a Random Forest model was a good choice, it would take longer to run. I chose not to use it based on the number and complexity of variables in the dataset.

# What I learned from “Kaggling”

- ▶ A good model starts with an understanding of the data (and what it really means)
- ▶ Keep the model simple
- ▶ Explore - Create - Do - (Repeat)
- ▶ All variables are not created equal.
- ▶ All data are not created equal.
  - ▶ Caveat 1: Public and Private Data are different
  - ▶ **Caveat 2: Do. Not. Over. Fit. The. Model.**
  - ▶ (Cross-validation!)

## Competition Results (Final)

Public Leaderboard  
AUC = 0.75177

33<sup>rd</sup> out of  
1685 teams

Private Leaderboard  
AUC = 0.76954

265<sup>th</sup> out of  
1685 teams

# Questions?

Diane Rucker  
Twitter: [@perky\\_r](#)