

# kaggle™

is a competition platform  
for (aspiring) data  
scientists

# MKaa

Merja Kajava



<https://emerginginsightsonline.com/>



Verified account

KAGGLER



Highest  
549th

Current  
1301st  
/360,755

4,867.3 points

Joined 2 years ago

Ranking method changed 13 May 2015 (%)



Profile

Results

Scripts

Forum



63rd/1233



59th/532



523rd/2236



534th/1785



477th/1568



212th/634



688th/1687



308th/718

10

Competitions

## Bio

Education:

MSc in Information Processing Science

## Scripts

MKaa has no [scripts](#) yet

Why participate in Kaggle  
The data  
Competition steps  
Tips

Why participate in  
Kaggle competition?

**1**

Learn from the best.

Forums

Scripts

Solutions from prize winners

2

Work with cool datasets.



imagination at work

Flights in GE Flights Quest



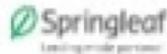
Driver telematic analysis



Amazon employee access rights

# +1 You can also win money

## Active Competitions

	 Springleaf Leading growth partners	<b>Springleaf Marketing Response</b> Determine whether to send a direct mail piece to a customer	59 days 383 teams 131 scripts \$100,000
	<b>RECRUIT Challenge</b> RECRUIT CHALLENGE	<b>Coupon Purchase Prediction</b> Predict which coupons a customer will buy	40 days 530 teams 348 scripts \$50,000
	 Caterpillar Tube Pricing	<b>Caterpillar Tube Pricing</b> Model quoted prices for industrial tube assemblies	10 days 1261 teams 848 scripts \$30,000
	 Liberty Mutual Group: Property Inspection Pred...	<b>Liberty Mutual Group: Property Inspection Pred...</b> Quantify property hazards before time of inspection	7.3 days 2220 teams 1995 scripts \$25,000

What kinds of  
competitions Kaggle has?

**kaggle**

Host

Competitions

Scripts

Jobs

Community ▾

**Public**

Welcome to Kaggle's data science competitions.

New to Data Science?

[Tutorials on the Titanic competition »](#)

Want to learn from other's code?

[Kaggle's top rated scripts »](#)**Download**

Choose a competition &amp; download the training data.

**Build**

Build a model using whatever methods and tools you prefer.

**In-class****Academic Machine Learning Competitions****Private**[Competition Details](#) » [Get the Data](#) » [Make a submission](#)**This competition is private-entry.** You can view but not participate.

# What languages can you use?



python™ **julia**



**WEKA**  
The University  
of Waikato

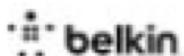


Gnu Octave  
(no Matlab)

Any open-source language  
(sometimes also sponsor's  
proprietary languages)

# What is the data like?

# Data comes from companies and non-profit organizations



Data sizes vary

 Zip ~1 MB

 Zip ~6 GB

# Data comes in all shapes

- Customer data
- Log files
- Timeseries
- HTML pages
- Images
- Documents

# How does Kaggle competition work?

# Competition flow

## Springleaf Marketing Response

●

A horizontal timeline bar consisting of a thin blue line with three circular markers. The first marker is at the left end, labeled "Fri 14 Aug 2015". The second marker is in the middle, labeled "Merger and 1st Submission Deadline". The third marker is at the right end, labeled "Mon 19 Oct 2015 (55 days to go)".

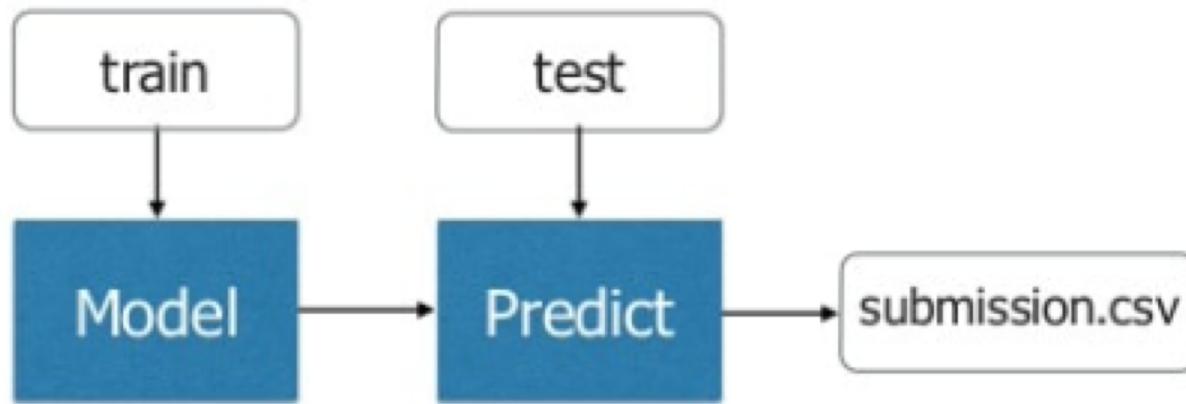
Fri 14 Aug 2015

Merger and 1st Submission Deadline

Mon 19 Oct 2015 (55 days to go)

Duration typically 4 to 8 weeks  
Max. 5 entries per day

# Build prediction model



Calculate CV to  
cross-validate

# Evaluate submission

Typical evaluations

Area under the ROC curve

Normalized Gini coefficient

RMSLE

...



# Submit entry

Private leaderboard

submission.csv



Public leaderboard

~10-30% of test data

# Choose two entries for final

Practical choice

Best entry in public leaderboard

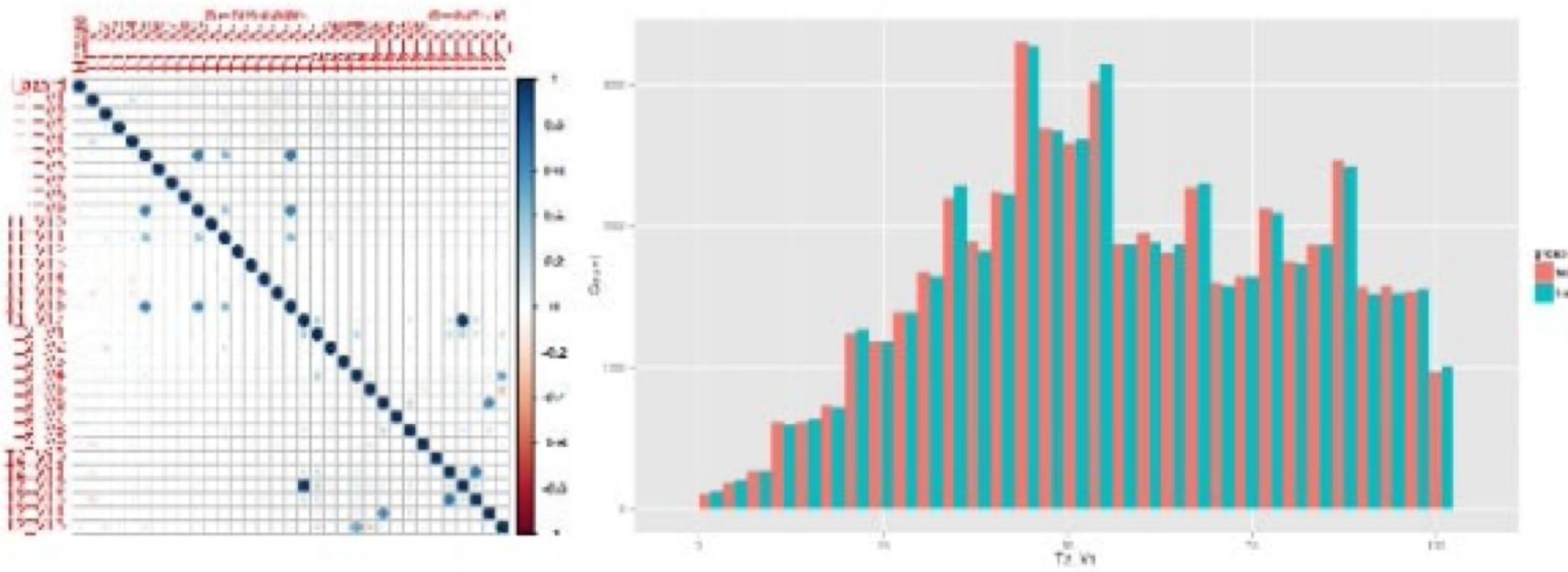
+

Best CV from local entries

# Tips

# Look at data. Visualize it.

## Liberty Mutual Group: Property Inspection Prediction



### Source

<https://www.kaggle.com/justfor/liberty-mutual-group-property-inspection-prediction/explore-data/notebook>

<https://www.kaggle.com/odiseo1982/liberty-mutual-group-property-inspection-prediction/compare-variables-between-train-and-test/files>

# Focus on feature engineering

Feature selection

Feature construction

Dates

Locations

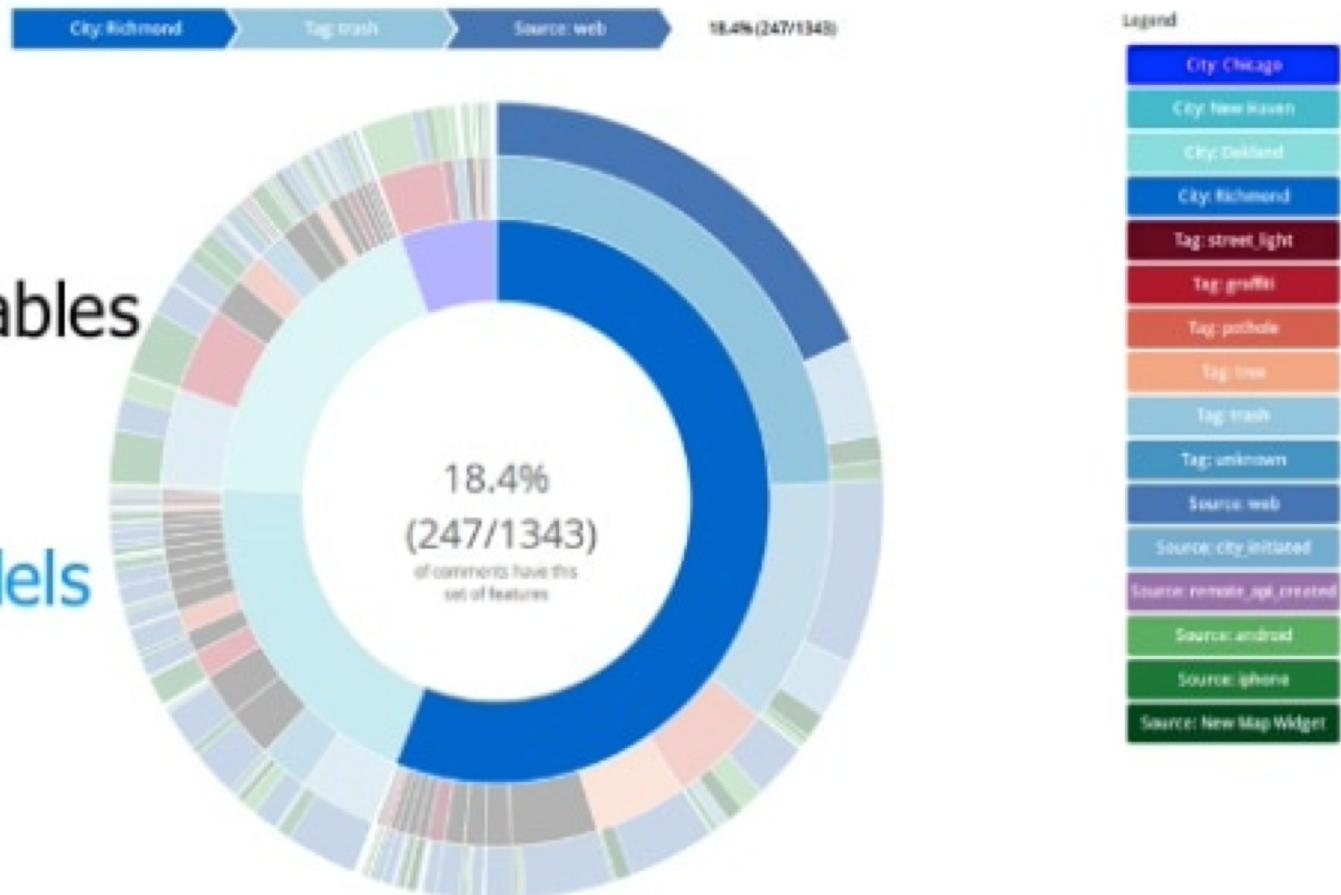
Categories

Segmentation

Statistics

# Build different models

3 target variables  
4 cities  
=  
**Build 12 models**



## Source

<https://www.kaggle.com/c/see-click-predict-fix/visualization/1390>

# Try different algorithms

Random forest

Vowpal Wabbit

GBM

Xgboost

# Build ensembles

Average of submissions

Weighted average of submissions

Ranked average of submissions

Stacked generalization

Blending

# Keep track of your submissions

## Submission id

Mon, 20 Oct 2014 17:49:05 Submission 008: Ensemble with 003 and 006 submissions. Feature P is converted to be no smaller than -0.41. <a href="#">Edit description</a>	008-2014102 0-submission .csv	0.40558	0.49566	<input type="checkbox"/>
Sun, 19 Oct 2014 18:09:14 Submission 007: Ensemble with 003 and 006 submissions. <a href="#">Edit description</a>	007-2014101 9-submission .csv	0.40594	0.49634	<input checked="" type="checkbox"/>
Sat, 04 Oct 2014 16:04:58 Submission 006: Benchmark with SVM in R with cost 10,000. Added space variables including Dept_h. <a href="#">Edit description</a>	006-2014100 4-submission .csv	0.43423	0.50596	<input type="checkbox"/>



# Next steps



# Start competing

Create Kaggle account

Choose competition

[Go for it!](#)

# Useful links

## Kaggle Blog

<http://blog.kaggle.com>

## Kaggle Competitions: Where to begin

<http://www.analyticsvidhya.com/blog/2015/06/start-journey-kaggle/>

## Kaggle Feature Engineering

<http://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/>

## Kaggle Ensembling Guide

<http://mlwave.com/kaggle-ensembling-guide/>