INT 303 BIG DATA ANALYTICS

#Lab 6: How to get started in Kaggle

Jia WANG

Jia.wang02@xjtlu.edu.cn

Xi'an Jiaotong-Liverpool University
西交利物浦大學

	Lecture	Lab	Project
Week1	Introduction	N/A	
Week2	What is data and big data?	Preparation	
Week3	Big data Gramma	Basics	
Week4	Data collection and Data scraping	Pandas	Project 1(Start 10.01)
Week5	Data Virtualization	Data Scraping	
Week6	Infrastructure that supports Big Data processing	Data Virtualization	
Week7	How to tell a good story	Q&A	
Week8	Representing Data and Engineering Features	Q&A	(Due 11.05)
Week9	Dimensionality Reduction	Feature Engineering	
Week10	Big Data Analysis Models	Dimensionality Reduction	Project 2 (Start 11.15)
Week11	Bagging&Boosting	Models	
Week12	Boosting&Stacking	Bagging&Boosting	
Week13	Ethics and social issues.	Boosting&Stacking	
Week14	Review	Q&A	(Due 12.24)

kaggle™

is a competition platform for (aspiring) data scientists

MKaa

Verified account

Merja Kajava

https://emerginginsightsnow.com/

KAGGLER

Highest†	Current†
549th	1301st /360,755

4,867.3 points
Joined 2 years ago
†Ranking metrics changed 18 May 2013 [*]

Profile | Results | Scripts | Forum

TOP 10%	TOP 25%	TOP 25%						
AfSIS						amazon	criteo	10
63rd/1233	59th/532	523rd/2236	534th/1785	477th/1568	212th/634	688th/1687	308th/718	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.

Flights in GE Flights Quest

Driver telematic analysis

Amazon employee access rights

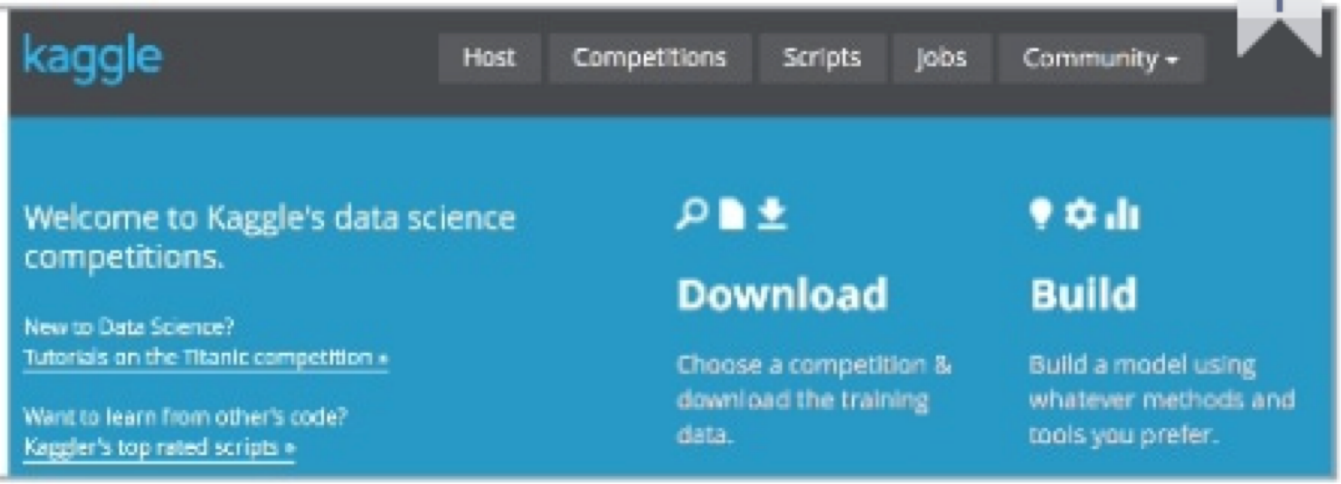
+1 You can also win money

Active Competitions

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

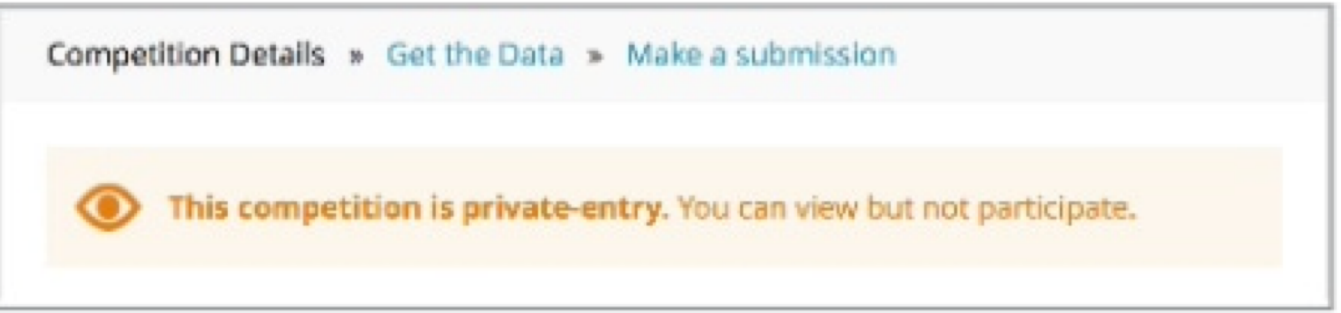
What kinds of competitions Kaggle has?

Public



In-class

Private



What languages can you use?

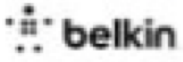
python™ julia

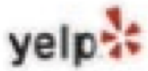
Gnu Octave
(no Matlab)

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

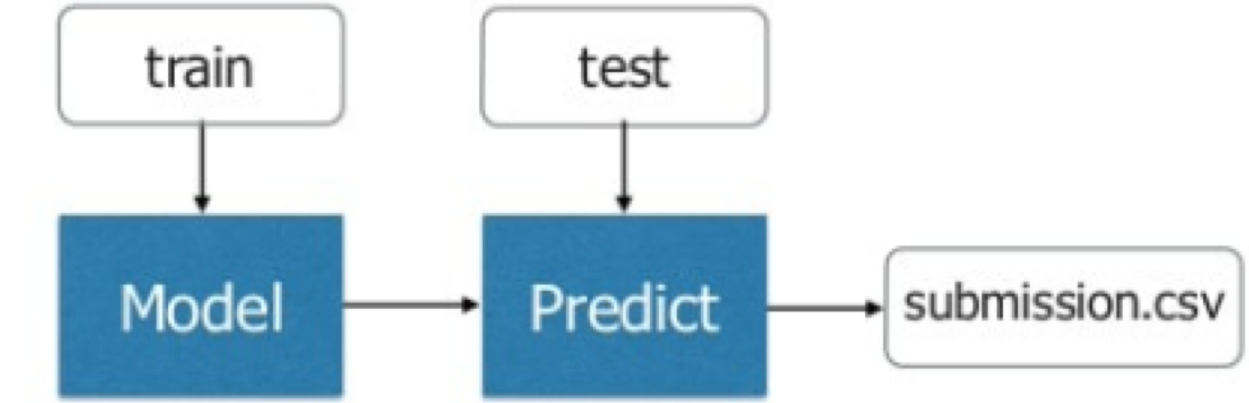
Customer data
Log files
Timeseries
HTML pages
Images
Documents

How does Kaggle competition work?

Competition flow

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

Build prediction model



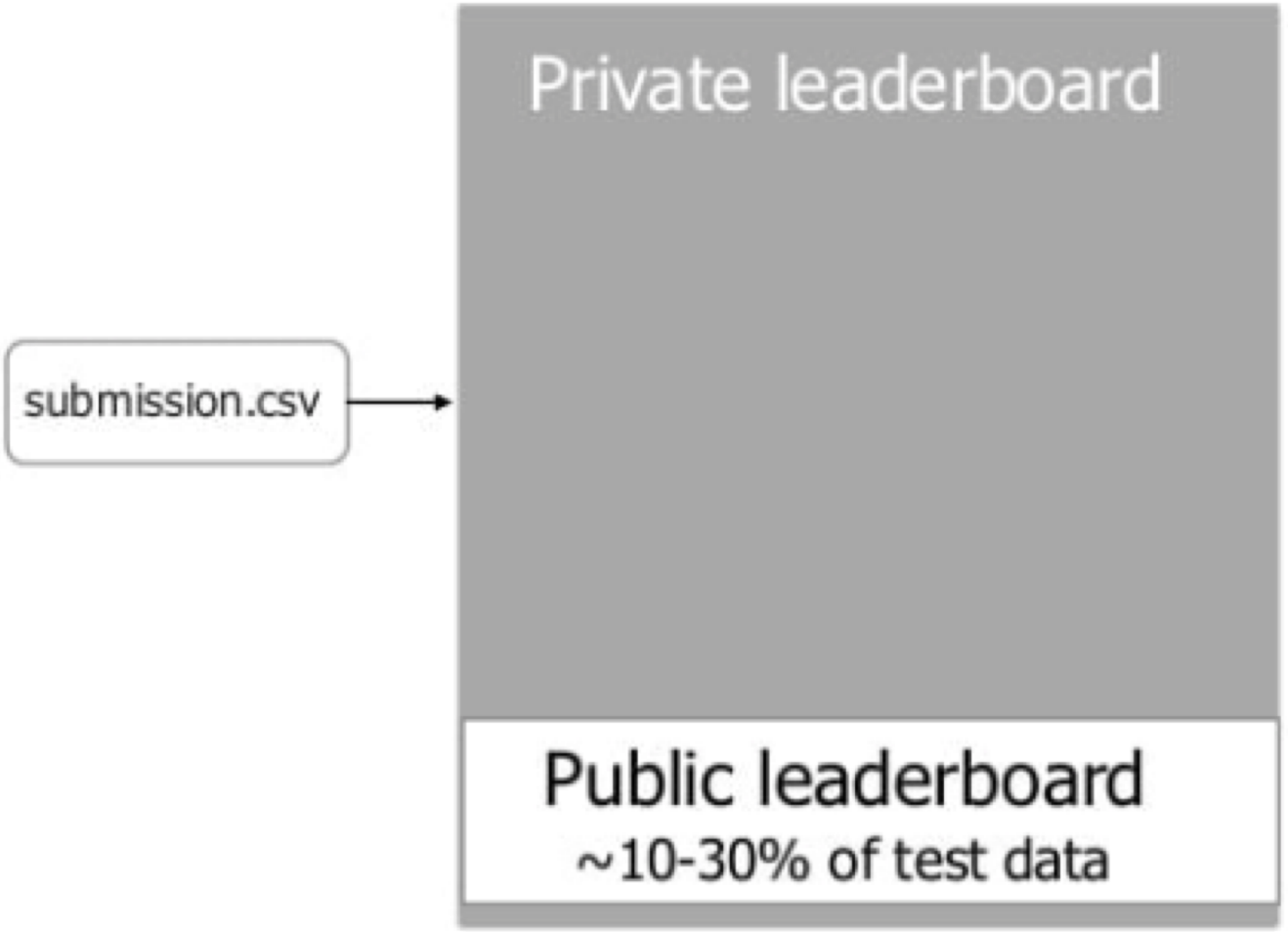
Calculate CV to cross-validate

Evaluate submission

Typical evaluations

Area under the ROC curve
Normalized Gini coefficient
RMSLE
...

Submit entry



Choose two entries for final

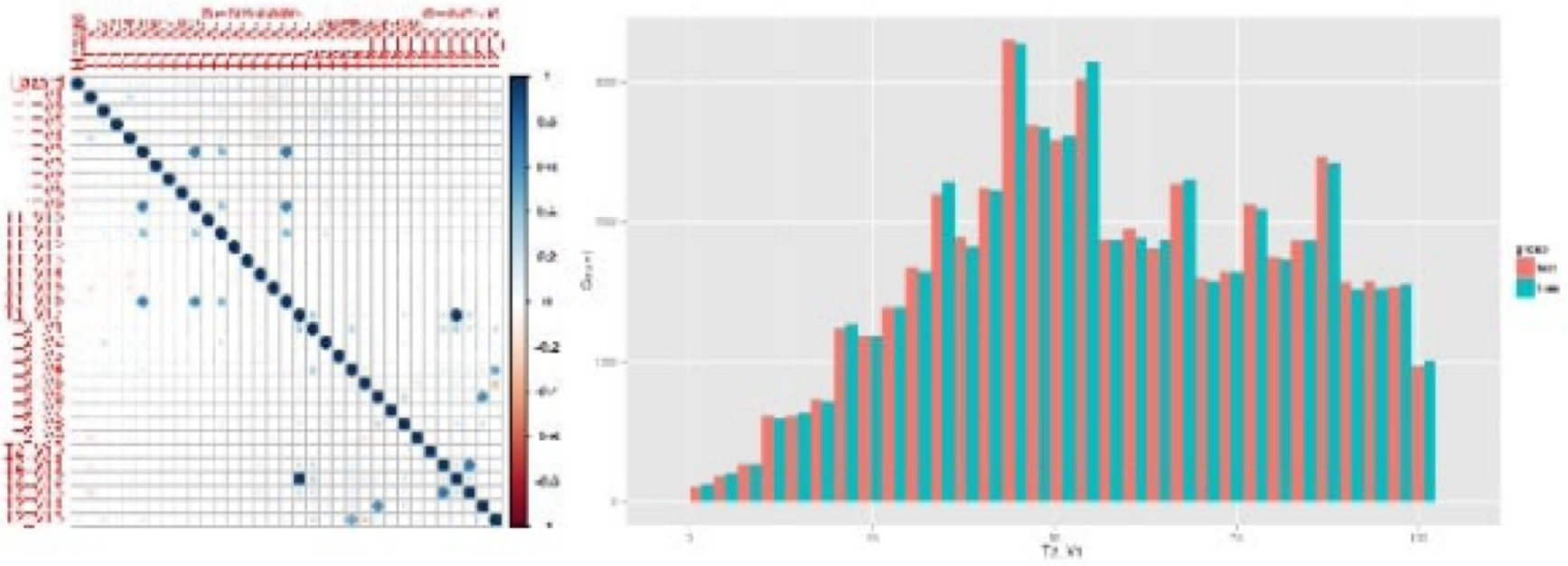
Practical choice

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

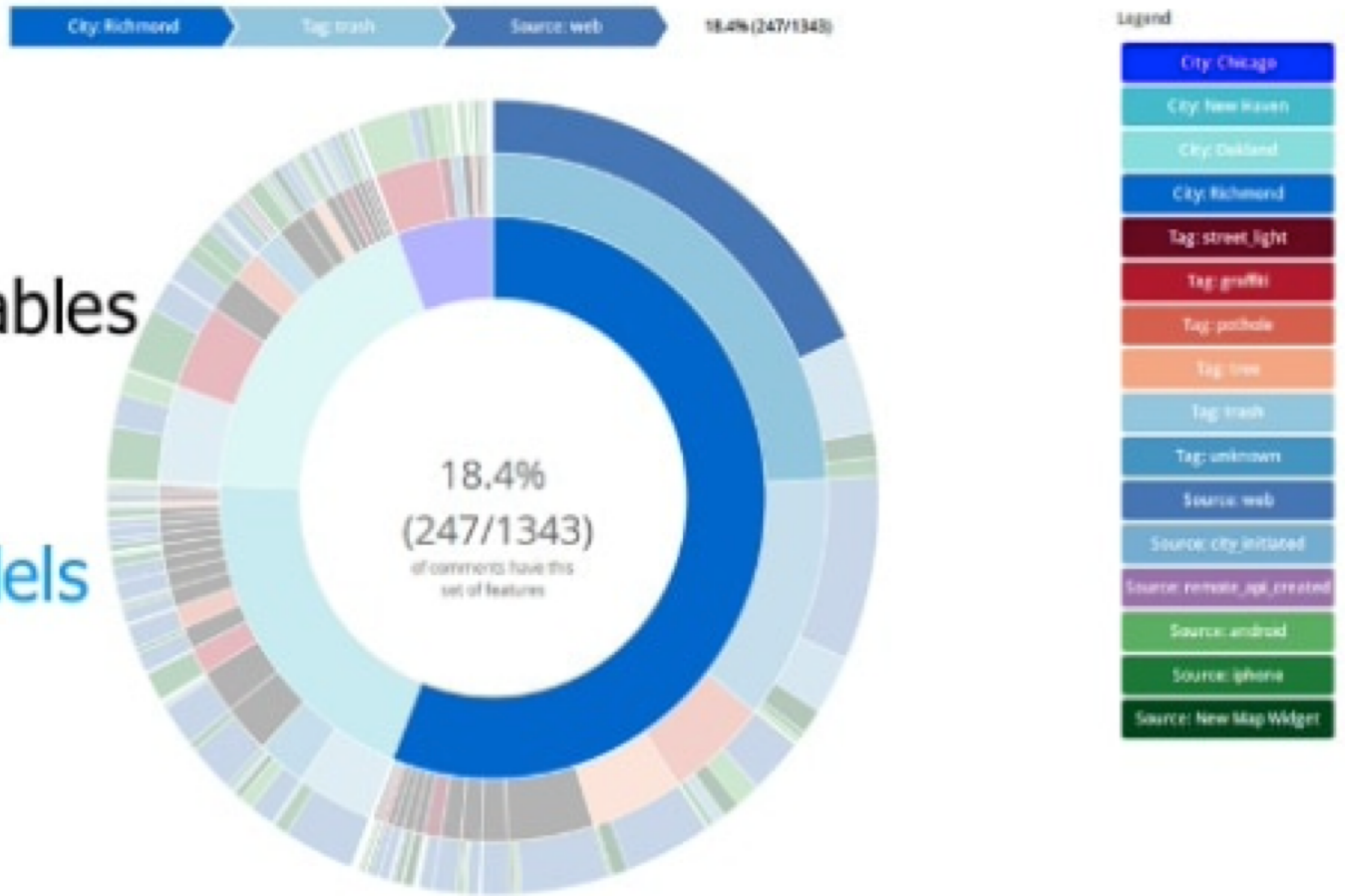
Feature selection

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

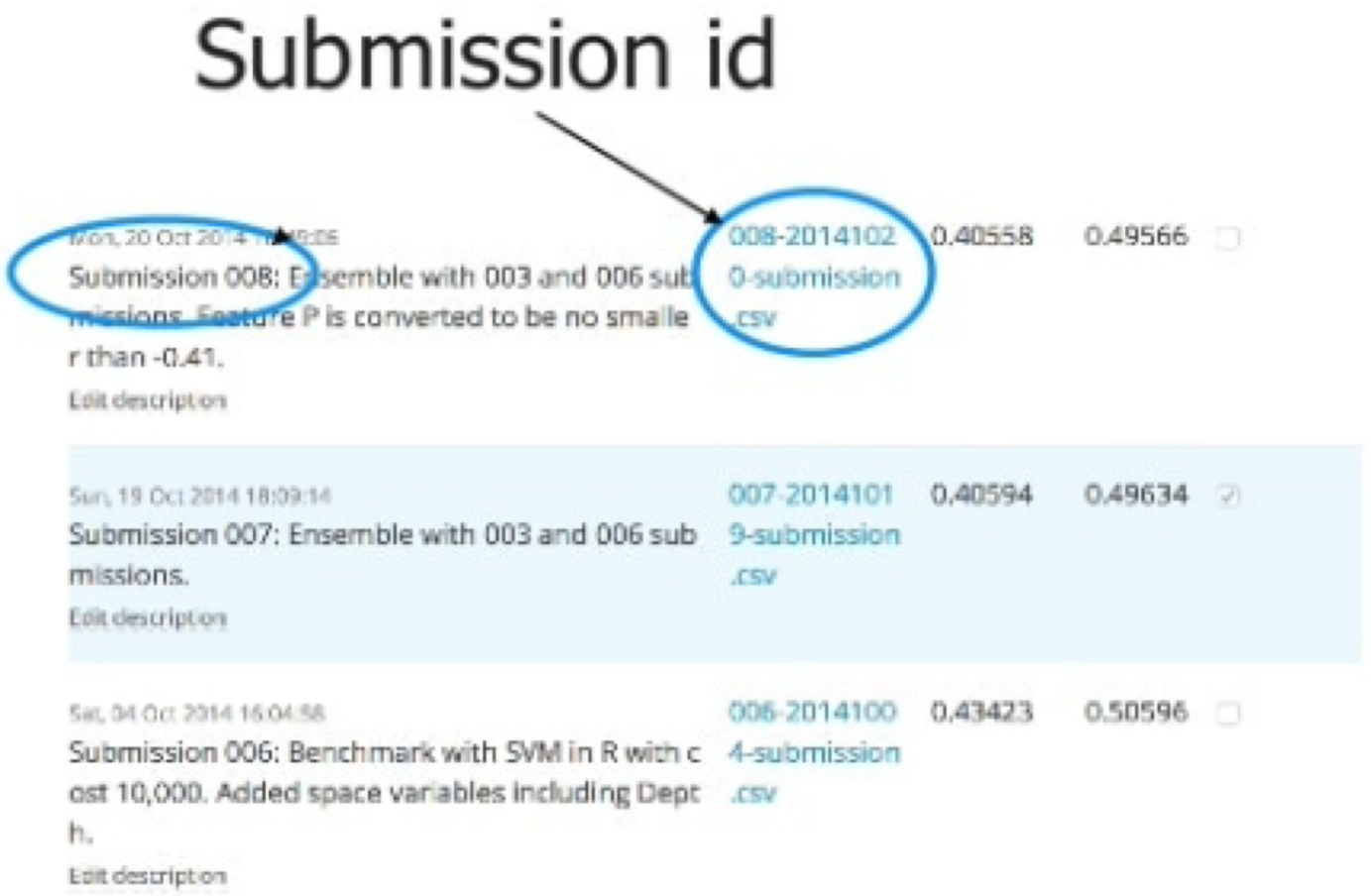
Random forest
Vowpal Wabbit
GBM
Xgboost

Build ensembles

Average of submissions
Weighted average of submissions
Ranked average of submissions

Stacked generalization
Blending

Keep track of your submissions



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/

THANK YOU

VISIT US

WWW.XJTLU.EDU.CN

FOLLOW US

@XJTLU

Xi'an Jiaotong-Liverpool University
西交利物浦大學