Basic Machine Learning with R

MACHINE LEARNING with kaggle™

WORKSHOP

19/02/2018

by Data Science ชิลชิล & Data Rockie

INTRO TO KAGGLE

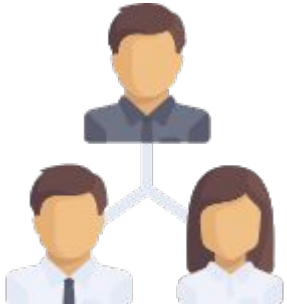
kaggle™

The popular platform for
data science competitions

Data Science Competitions

Prize Money

Time Limit



Teamwork

Real-world Machine Learning Problems

Real-world Machine Learning Problems

Predict House Price

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Real-world Machine Learning Problems

Revenue Forecasting

https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

Real-world Machine Learning Problems

Predict HIV Progression

https://www.kaggle.com/c/hivprogression

Real-world Machine Learning Problems

Classify Satellite Images

https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection

Acquired by Google in 2017

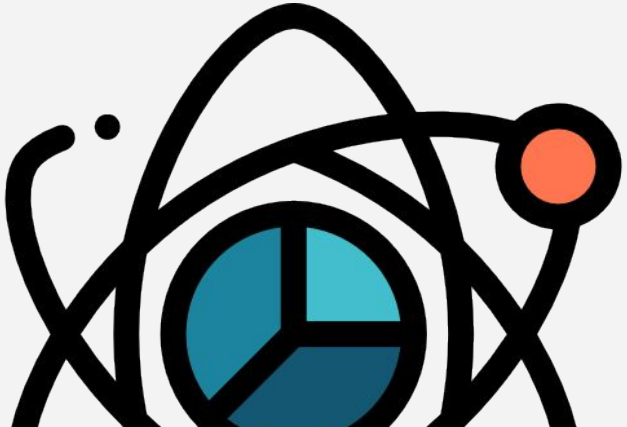
Is Kaggle a good way to become data scientist?

YES

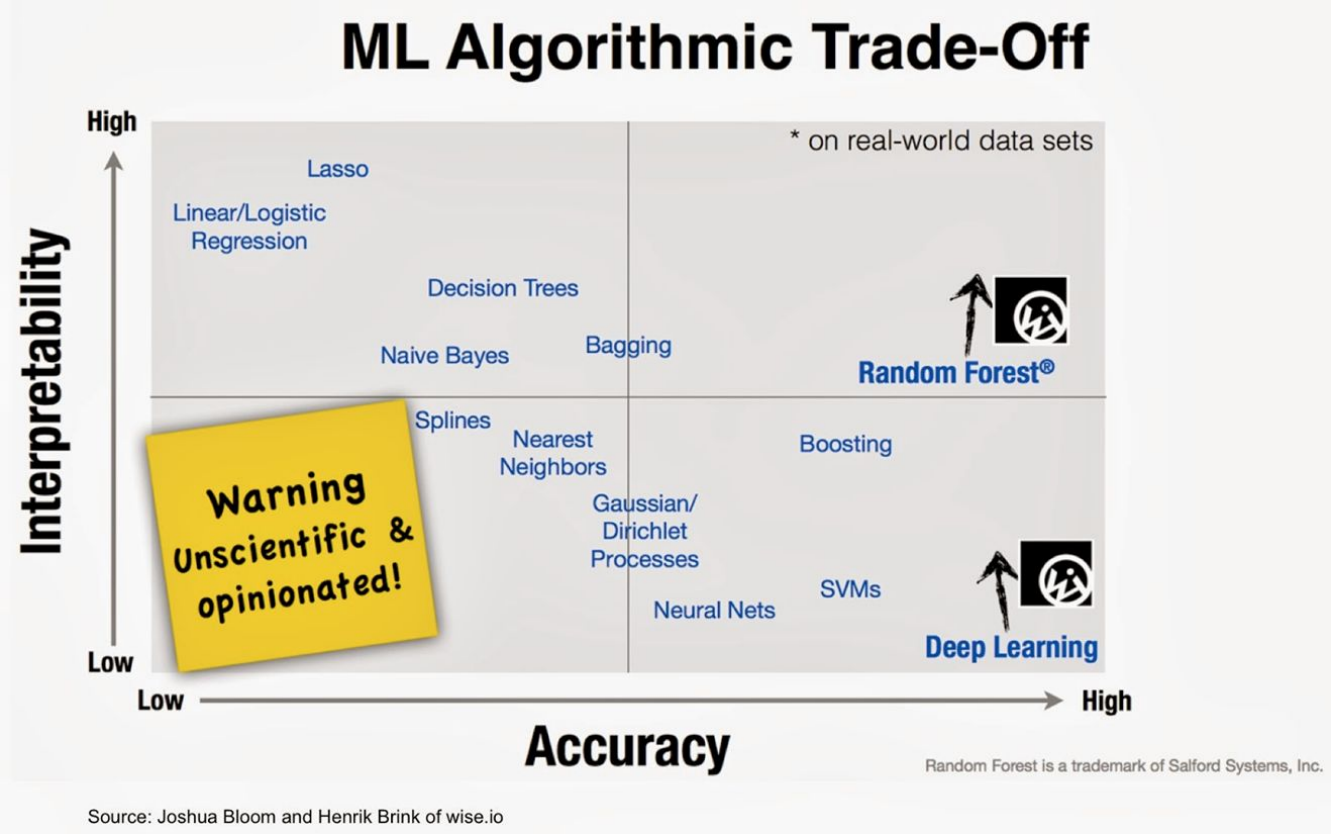
- Real-world problem, data
- Learn new library in R or Python
- Learn from many people & winners sharing their explorations / solutions.

NO

- Skip data collection & cleaning
- New library may not be practical



Model accuracy & interpretability Tradeoff



Interpretability
Easy to understand

VS

Accuracy
Easy to get high prediction score

“As long as complex models are properly validated, it may be **improper to use a model that is built for interpretation rather than predictive performance.**”

- Kuhn and Johnson “Applied Predictive Modeling”

Accuracy > Interpretability

Get started with Kaggle

What Kaggle problem looks like?

Example: Predict who will like ice cream?

Name	Age	Weight (KG)	Like Ice Cream?
Chatri	24	60	Yes
Tong	30	50	No
Somsri	42	48	No
Thanet	18	72	We have to predict the answers
Petch	35	48	
Pongsak	26	62	

Data Science Process

"Turn **data** into **insights**"

Collect Data > Maintain >
Explore Data > Data Cleaning > Build Model >
Visualize > Present

Kaggle Process

kaggle

"Turn **data** into **prediction**"

~~Collect Data > Maintain >~~

Explore Data > Data Cleaning > Build Model >

~~Visualize > Present~~

“No matter how advanced is your Machine Learning algorithm, **the results will be bad if the input data is bad**.”

- *Kaiser Fung “Numbersense: how to use Big Data to your advantage”*

WORKSHOP OVERVIEW

What will we do today?

Kaggle competitions for beginners

- No time limit
- Public leaderboard reset every 2 months

2 Workshops

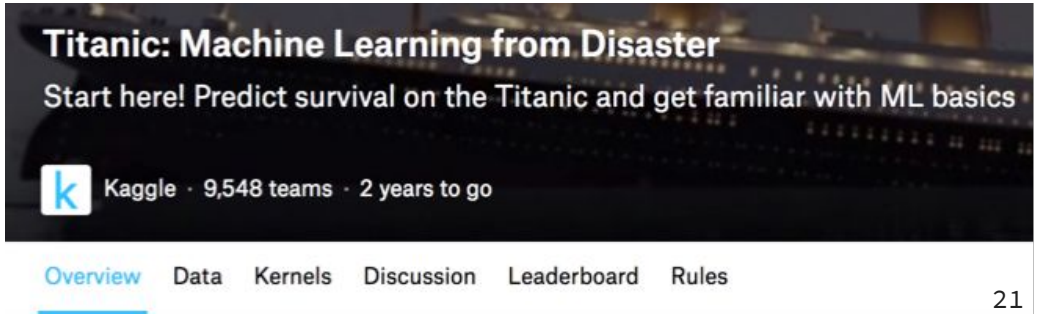
1. **Classification Problem:** Who is likely to survive on Titanic?
2. **Regression Problem:** predict the house price from its quality

Workshop 1: Titanic

Kaggle Page: https://www.kaggle.com/c/titanic

Kaggle Project Tabs

1. Overview (description, goal, evaluation metric, submission format)
2. Data (Test, Train, Data dictionary)
3. Kernels (Learn from others)
4. Discussion (Webboard)
5. Leaderboard (Public & Private)
6. Rules (Timeline)



Many Features & 1 Target Variable

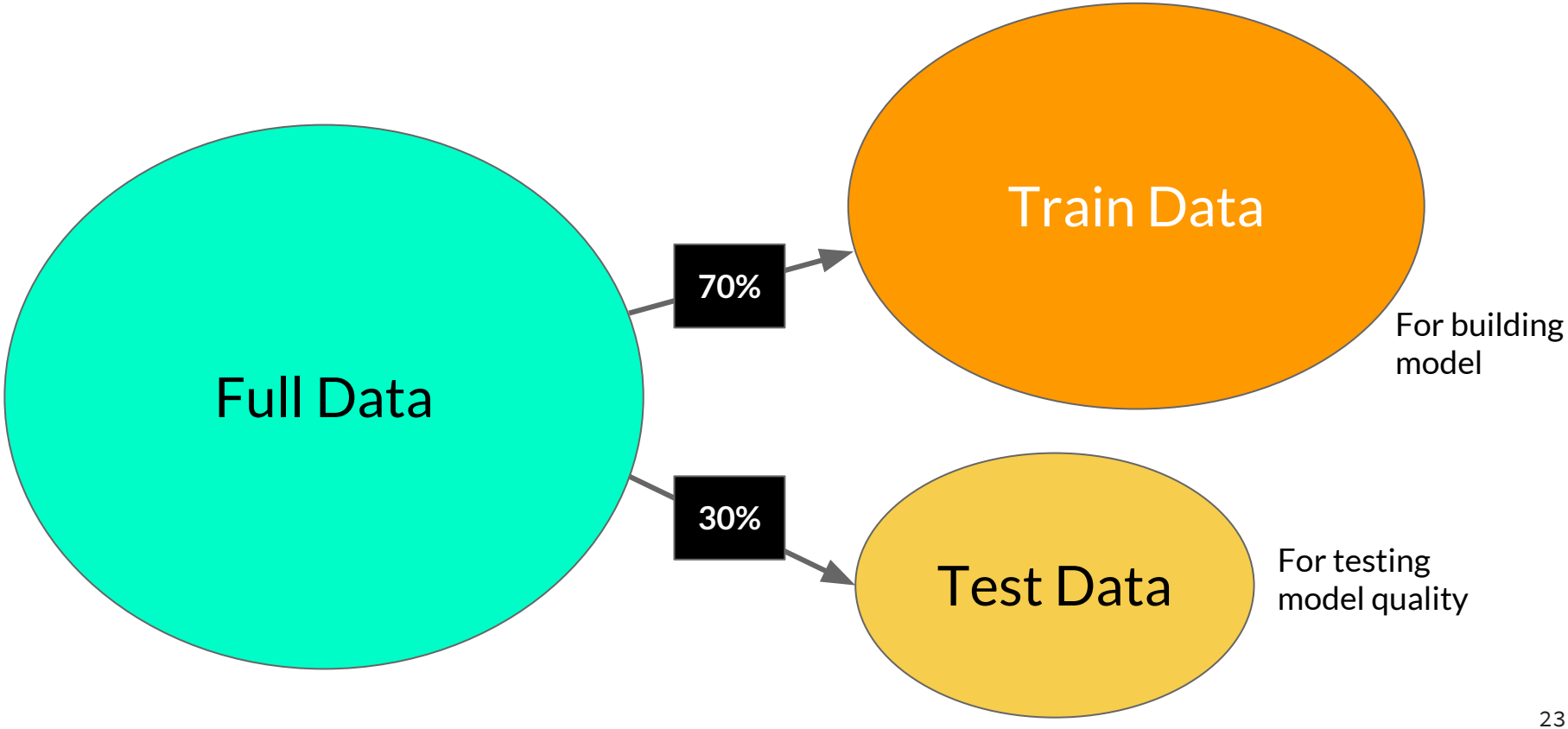
Example: Predict who will like ice cream?

Name	Age	Weight (KG)	Like Ice Cream?
Chatri	24	60	Yes
Tong	30	50	No
Somsri	42	48	No
Thanet	18	72	We have to predict the answers
Petch	35	48	
Pongsak	26	62	

X: Features

Y: Target Variable

Data needs to be **splitted**



What is train & test data

Example: Predict who will like ice cream?

Name	Age	Weight (KG)	Like Ice Cream?
Chatri	24	60	Yes
Tong	30	50	No
Somsri	42	48	No
Thanet	18	72	We have to predict the answers
Petch	35	48	
Pongsak	26	62	

Train

Test

What is
Public Leaderboard & Private Leaderboard

Test Data

Name	Age	Weight (KG)	Like Ice Cream?
Thanet	18	72	We have to predict the answers
Petch	35	48	
Pongsak	26	62	
Kate	16	42	

Public

Private

WORKSHOP 1

Workshop 1: Titanic

Question: Who will survive from Titanic?

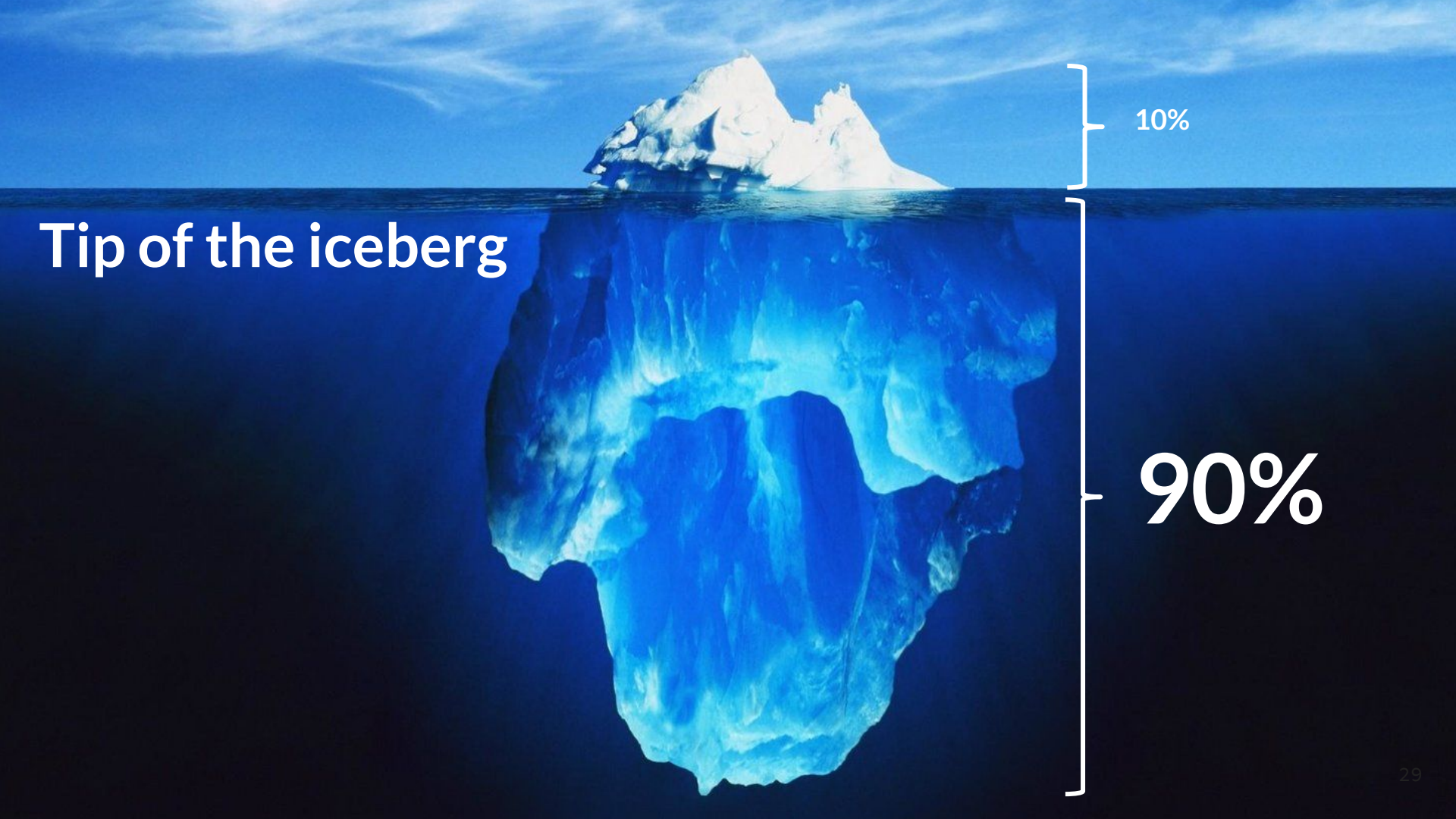
Step-by-step guide:

1. Load data
2. Explore data
3. Data cleaning
4. Model training

Titanic

The ship from England to United States in 1912

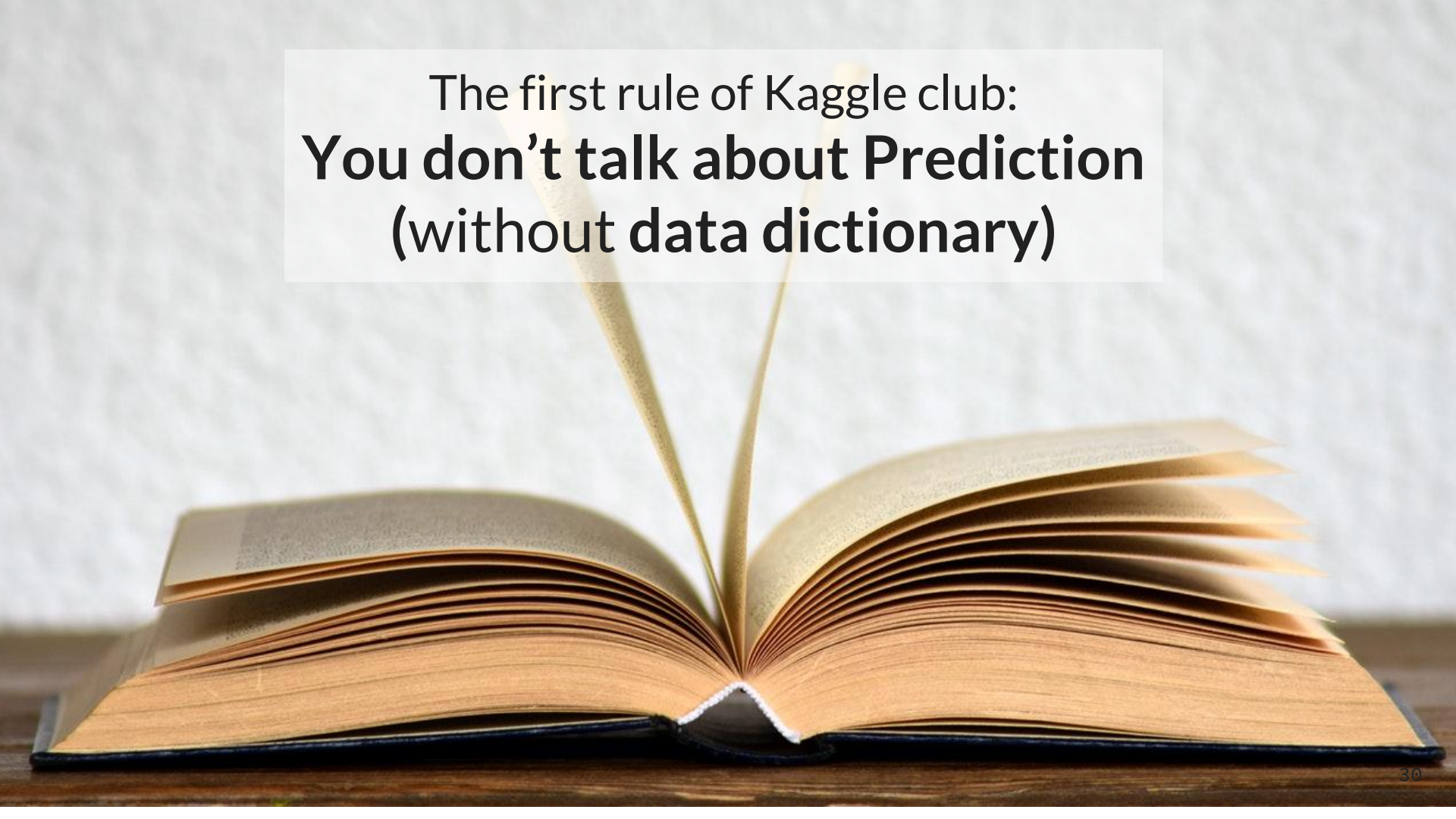
Tip of the iceberg



The first rule of Kaggle club:

You don’t talk about Prediction

(without **data dictionary)**



First step of ML:
Find the target variable

Titanic: Data Dictionary

Variable	Definition	Key
survival	(Target Variable) Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Let’s get to the code

Structure of "workshop1-titanic.R"

1. Load data
2. Explore data
   a. Summary
   b. Correlation
3. Clean data
   a. Impute missing values
   b. Convert data types
4. Train model using **decision tree** & Visualize
5. Train model with **random forest** & optimize using **random search**
6. Use model to predict the test data
7. Export file to submit to Kaggle

Titanic: Explore Data

Titanic Fact - https://www.encyclopedia-titanica.org/

Find answers from the data using R:

1. How many passengers?
2. How many people in each ticket class?
3. What is the average age of people on Titanic?
4. What is the correlation between variables?

Titanic: Cleaning Data

3 Types of Data Anomalies

1. Missing Values (where there should be data) - delete or impute
2. Wrong Data (misspelling, wrong column)
3. Incomplete Data (abbreviation, multiple measurement unit, incomplete address)

Titanic: Evaluation Metric = Accuracy

Confusion matrix →

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

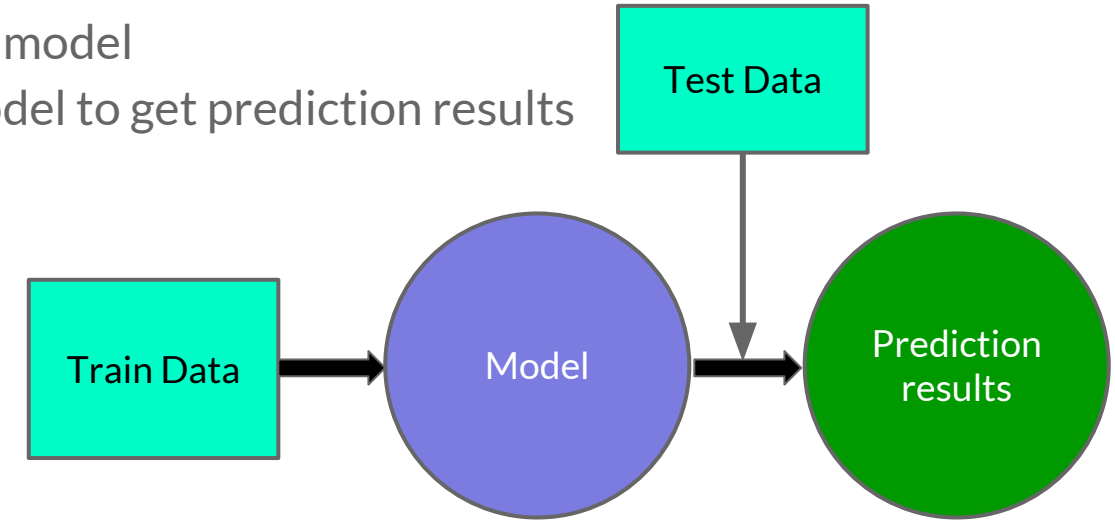
Correct answers

Correctly predict 0% -> Accuracy = 0
Correctly predict 50% -> Accuracy = 0.5
Correctly predict 100% -> Accuracy = 1

Titanic: Building Model

How to build a prediction model

1. Prepare features and target variables
2. Use train data to train the model
3. Feed test data into the model to get prediction results
4. Optimize the model
5. Submit to Kaggle



Titanic: Feature Engineering

Feature Engineering = Build a feature from the current attributes

e.g. **Turn numeric into categorical** --> Price (*$800, $5000, $8000, ...*) **to**
Price_less_than_5000, Price_more_than_5000

e.g. **Extract some parts** --> Date (*30-5-2017, ...*) **to**
Hour_of_day, Day_or_night

e.g. **One Hot Encoding** --> Color (*red, blue, green*) **to**
Is_red, Is_blue, Is_green

Titanic: Optimize the model with Grid Search

What is the easiest way to find the best parameter?`

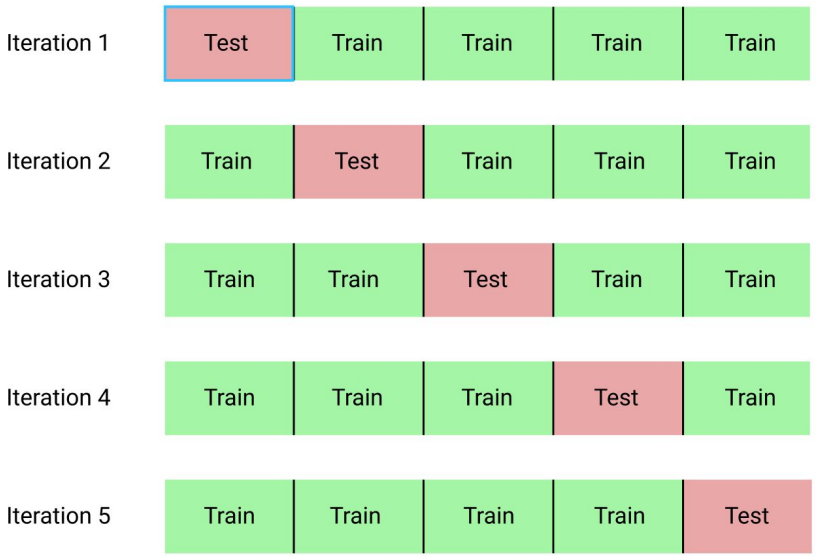

```
gbmGrid <-  expand.grid(interaction.depth = c(1, 5, 9),
                        n.trees = (1:30)*50,
                        shrinkage = 0.1,
                        n.minobsinnode = 20)

gbmFit2 <- train(Class ~ ., data = training,
                 method = "gbm",
                 trControl = fitControl,
                 verbose = FALSE,
                 ## Now specify the exact models
                 ## to evaluate:
                 tuneGrid = gbmGrid)
```

Titanic: Evaluate the model before submission

K-Cross Validation

Split data into *k* equal portions. Use 1 portion for test, and the rest for train for *k* times.

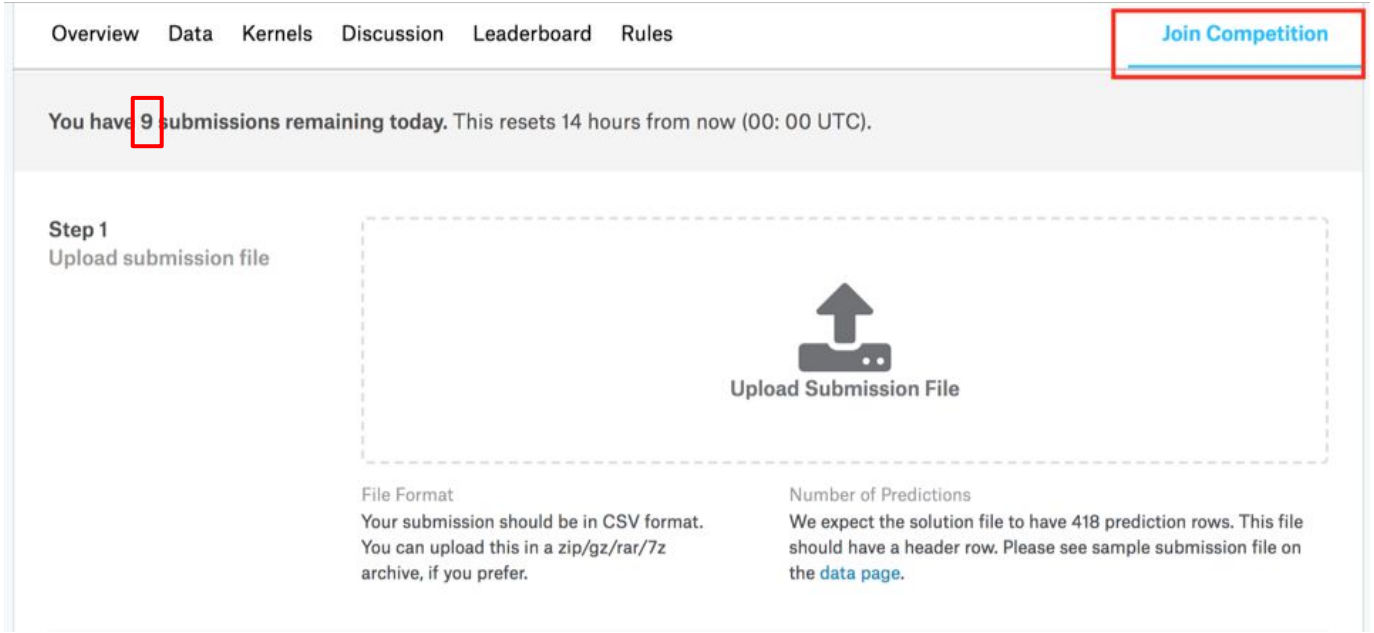


K = 5

People normally use 5 or 10

https://my.oschina.net/Bettyty/blog/751627

Titanic: How to submit



Warning: There is a limit you can submit per day.

WORKSHOP 2

Workshop 2: House Price Prediction

Kaggle Page: https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Problem type: Regression

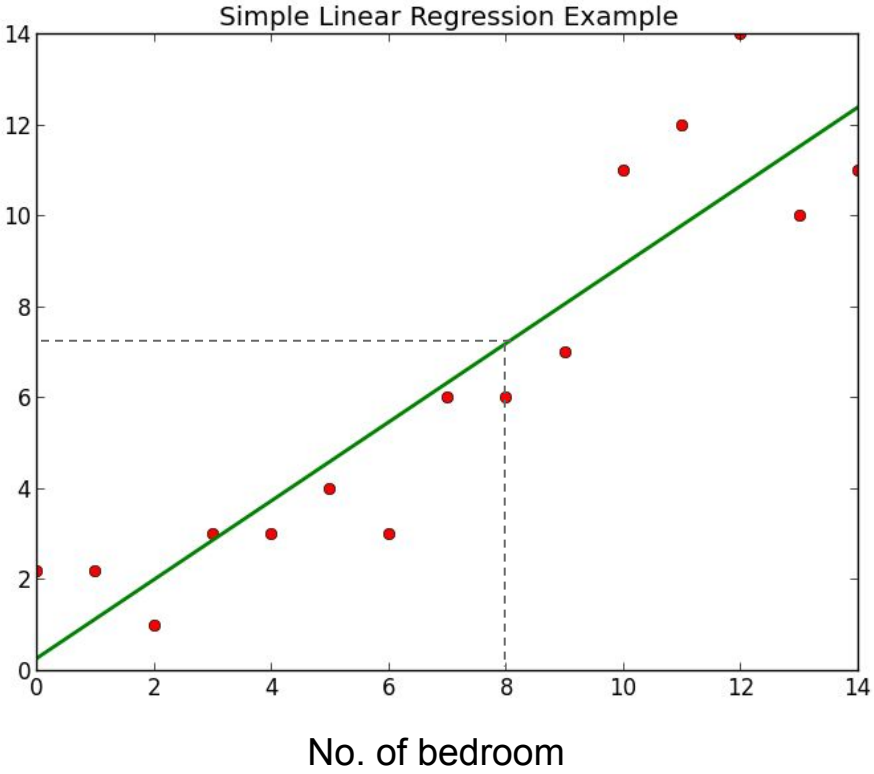
Regression VS Classification problem?

Regression: Target variable = continuous number (e.g. 1, 1.2, 1.5)

Classification: Target variable = categorical (e.g. Yes = 1, No = 0)

Regression Problem

House Price
(1 = $10,000)



One data point

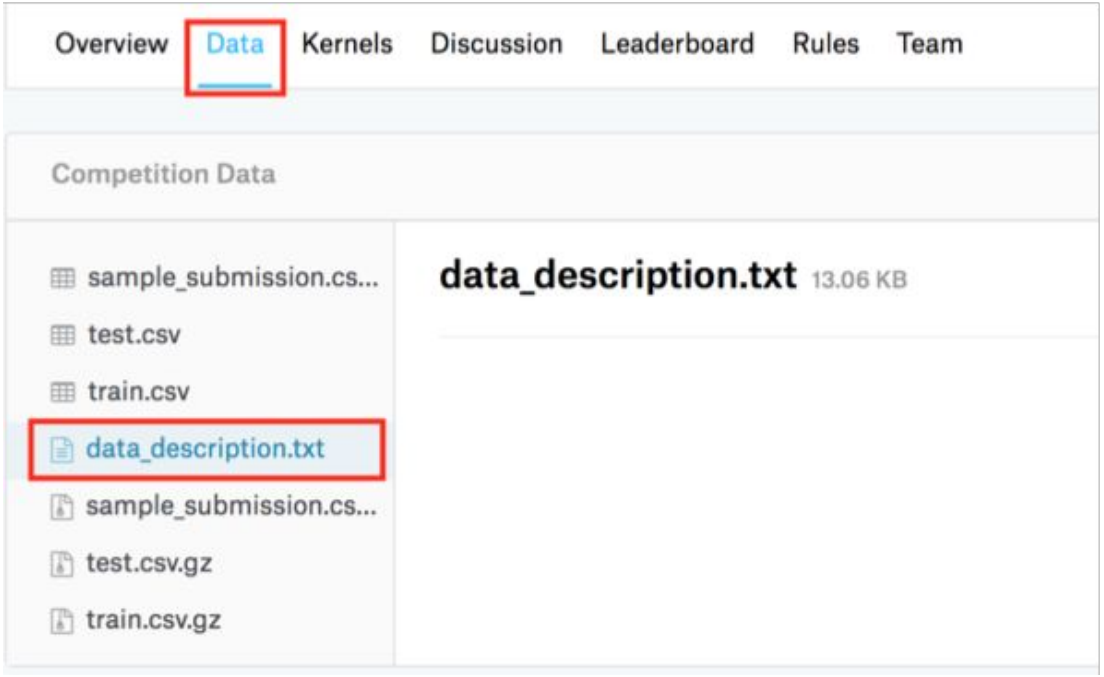
Regression line

Example:
When No. of bedroom = 8,
Predicted Price = $77,000

https://austingwalters.com/using-svd-to-obtain-regression-lines/

Workshop 2: House Price Prediction

Data Dictionary can be found in: Data > data_description.txt



Workshop 2: House Price Prediction

Question: What is the house sale price from its quality?

Step-by-step guide:

1. Load data
2. Explore data
3. Data cleaning
4. Model training

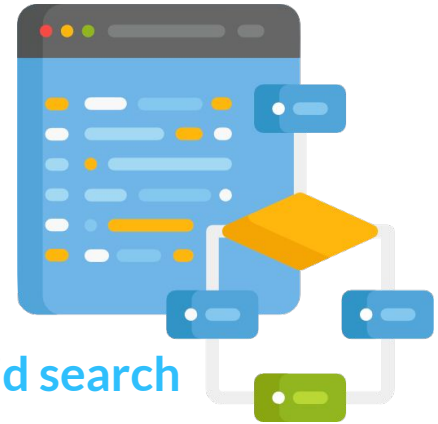
First step of ML:

Find the target variable

Let’s get to the code

Structure of "workshop2-houseprice.R"

1. Load data
2. Explore data
   a. Summary
   b. Correlation
3. Clean data
   a. (Add any function you would like)
4. Train model using **random forest** & optimize using **grid search**
5. Use model to predict the test data
6. Export file to submit to Kaggle



House Price: Evaluation Metric = RMSE

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum\left(\hat{Y}_i - Y_i\right)^2}$$

Sum (for each data point):
N = No. of data point
Y^ = Predicted price
Y = Real price

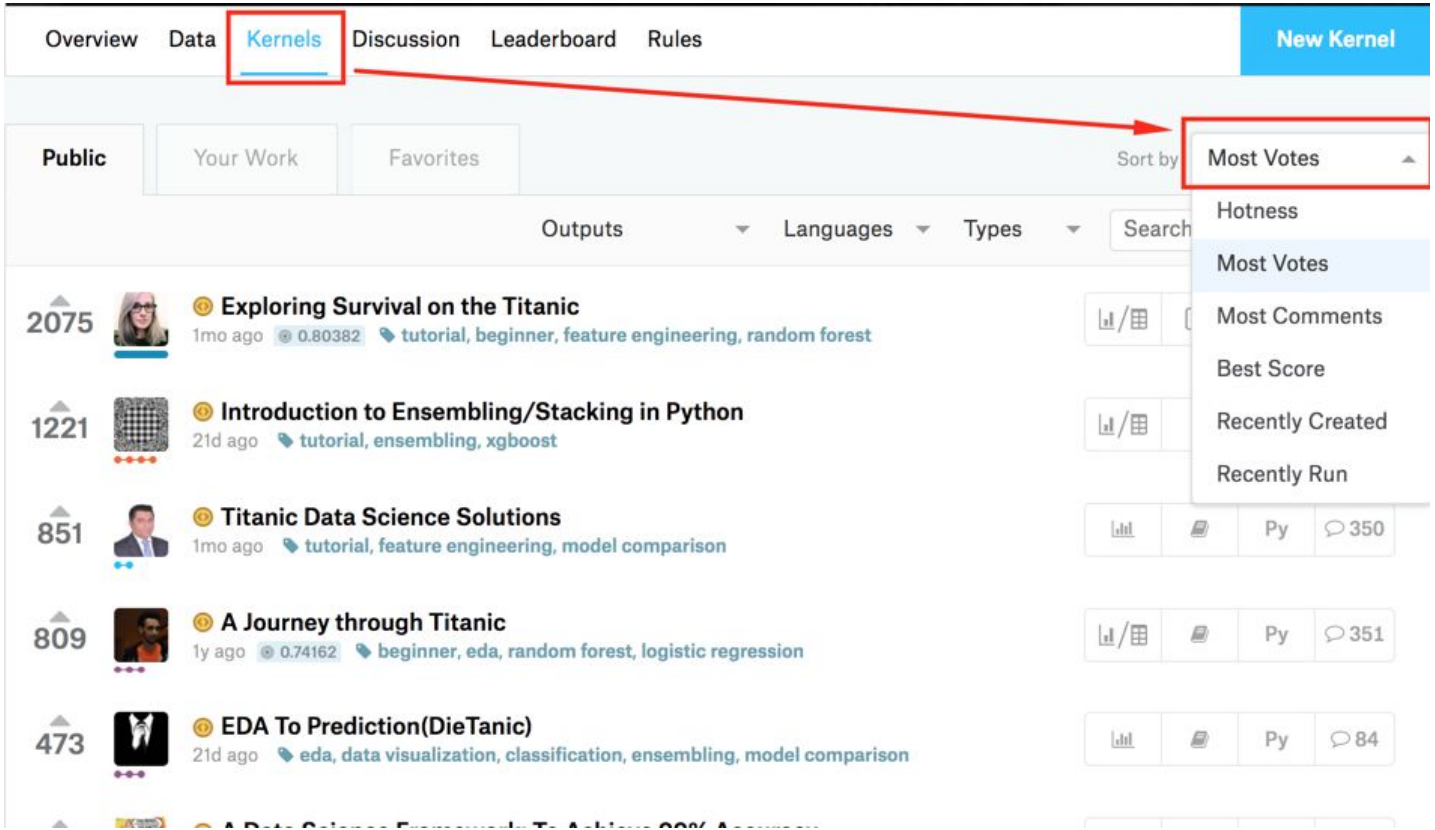
RMSE = How big is the prediction error?

More RMSE = less accurate model
RMSE 0 = No difference
Less RMSE = more accurate model

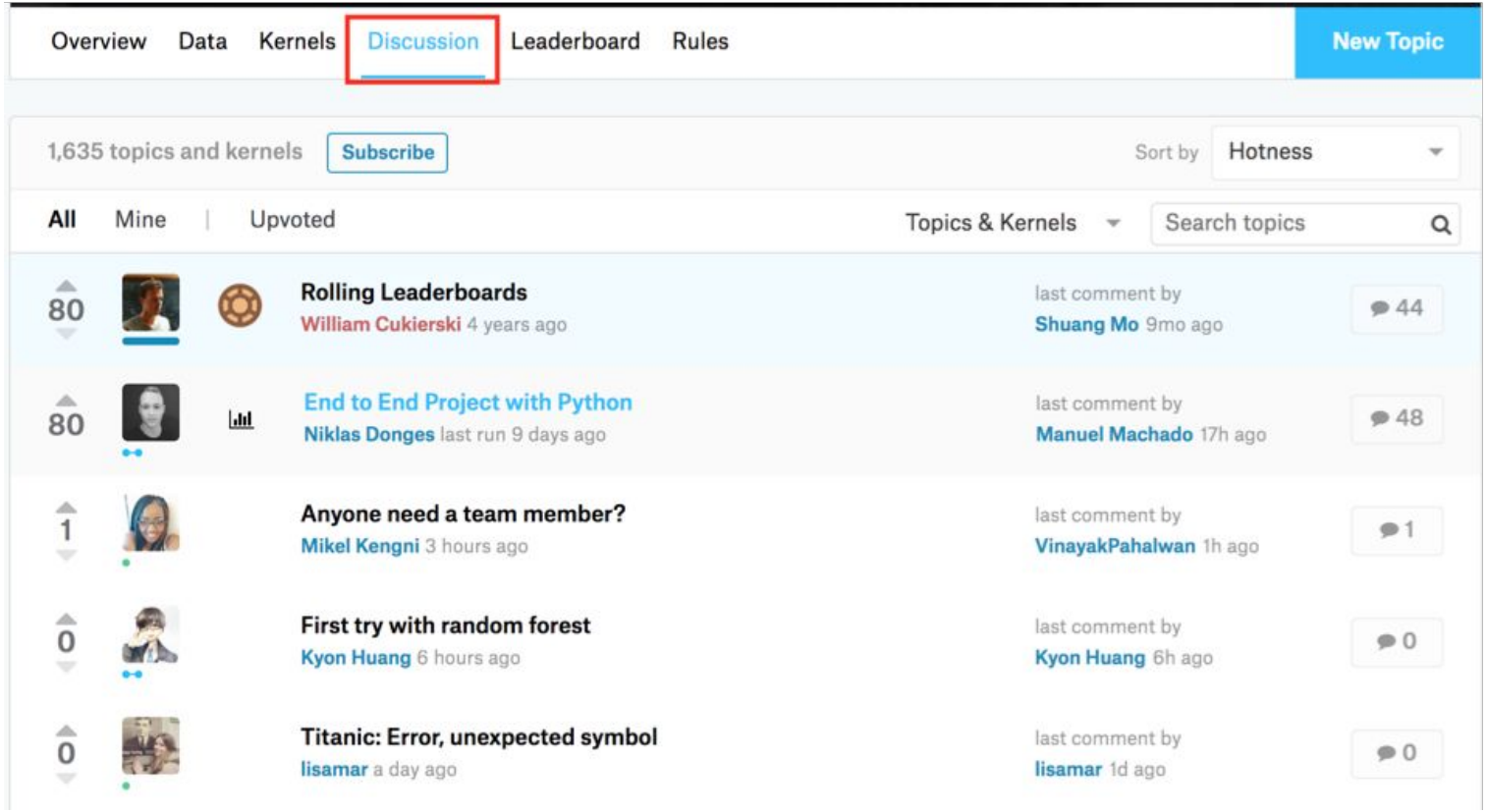
Mourice, S. , Rweyemamu, C. , Tumbo, S. and Amuri, N. (2014) Maize Cultivar Specific Parameters for Decision Support System for Agrotechnology Transfer (DSSAT) Application in Tanzania. American Journal of Plant Sciences, 5, 821-833. doi: 10.4236/ajps.2014.56096.

How to get the most of Kaggle?

Learn from the “Most Votes” Kernel



Participate in the forum discussion



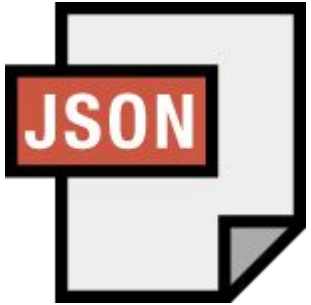
Invite friends to do Kaggle together

Facebook Group:
Thai Data Scientists & Kagglers:

https://www.facebook.com/groups/thaidsml

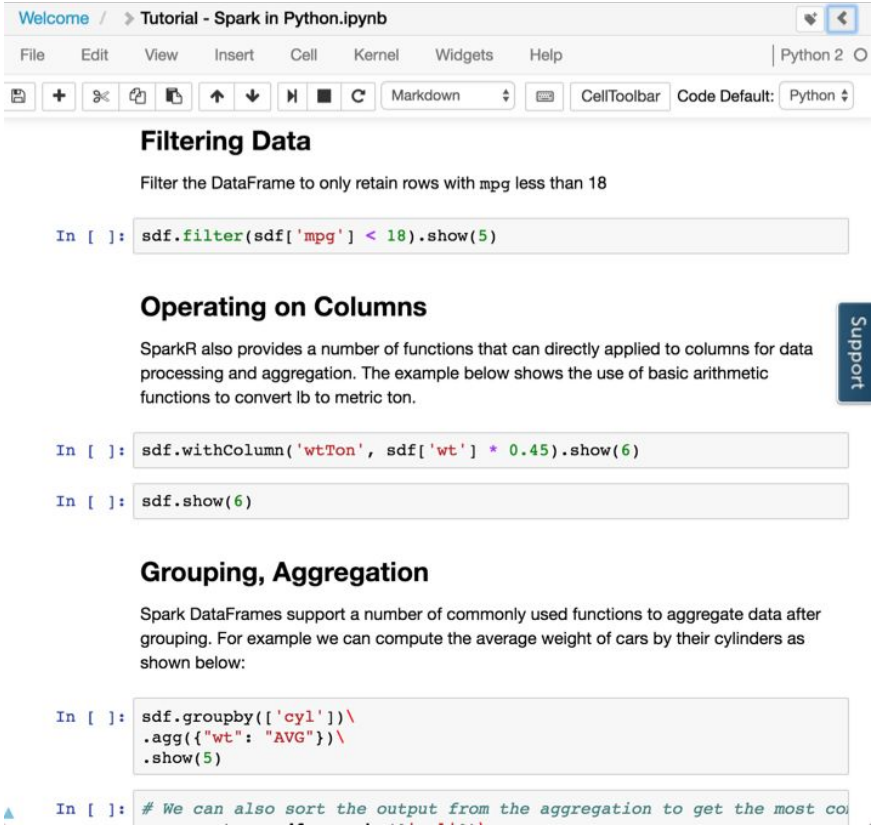
We will discuss & update in this group :)

Play around with free data sets



- Government Data - Data.gov
- World Health Organisation: http://apps.who.int/gho/data/node.home
- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.html
- Data World - Data.world

Document your learning with Notebook

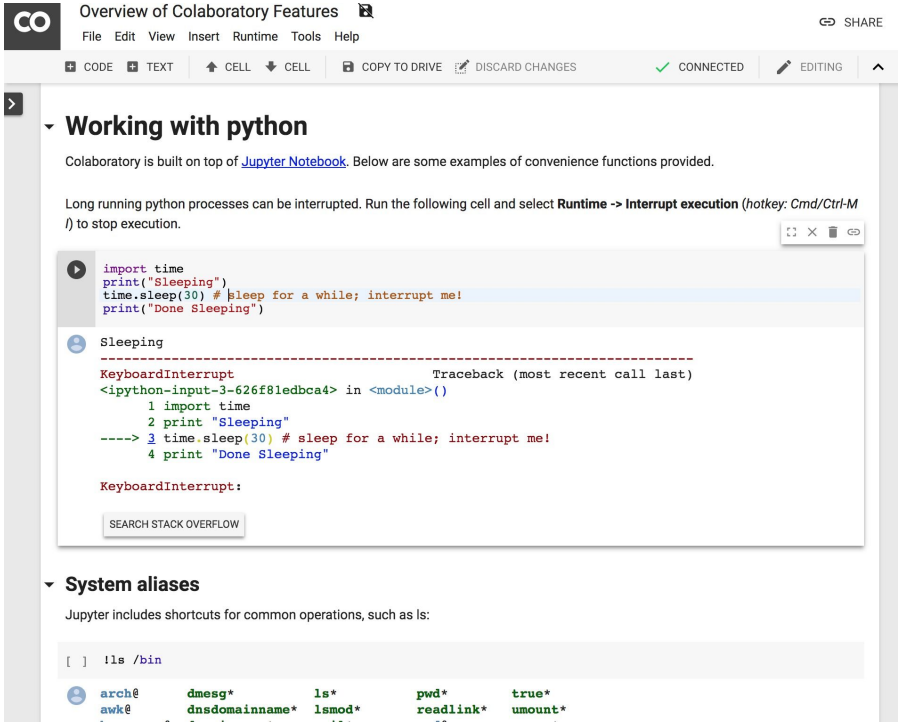


Jupyter Notebook

https://datascientistworkbench.com/

- Free
- Same as “Kernel” on Kaggle
- Offline (with Anaconda) or Online
- Support Python with Anaconda installation https://anaconda.org/
- Support R with IRKernel

Document your learning with Notebook



Google Colaboratory

https://colab.research.google.com/notebook

- Free
- Online
- Support Python 2 and 3
- Shareable link
- Collaboration with people (like Google Docs)

Congratulations!

You have became **Kagglers**!

RECAP

- Kaggle is a good way to practice machine learning, data analysis skills
- Kaggle’s goal: Turn data into prediction results
- Data dictionary is important for understanding the data set
- Find the target variable first

RECAP: Building Model

How to build a prediction model

1. Prepare features and target variables
2. Use train data to train the model
3. Feed test data into the model to get prediction results
4. Optimize the model
5. Submit to Kaggle