

# Kaggleの紹介

AI&機械学習しよう！（Do2dle）勉強会

阿部 泰之

# アジェンダ

1. Kaggleとは
2. Kaggleの特徴
3. まずはじめにやるのはどれがおすすめか
4. Kaggleで有効な手法の紹介

# 自己紹介



- 阿部 泰之 / Hiroyuki Abe
- twitter / @taki\_tflare
- <https://tflare.com>
- 業務エンジニア  
(生命保険 主に保険金支払)
- Kaggle初めたばかり



# Kaggleとは

**Kaggleは企業や研究家が、データを投稿し、統計家やデータ分析家がその最適化モデルを競い合うサイトです。**

**(高額賞金付きのコンペが実施されることがあり  
現在合計賞金1,200,000ドルのコンペ実施中  
＊約1億3千万円)**

kaggle

Search kaggle



Competitions

Datasets

Kernels

Discussion

Jobs



Sign In

# The Home of Data Science & Machine Learning

Kaggle helps you learn, work, and play

[Create an account](#)

or

[Host a competition](#)

---

## Competitions ›

Climb the world's most elite  
machine learning  
leaderboards

[Want to host a competition?](#)

---

## Datasets ›

Explore and analyze a  
collection of high quality  
public datasets

---

## Kernels ›

Run code in the cloud and  
receive community feedback  
on your work

# Kaggleの特徴

**Competitions**

**Kernels**

**Discussion**

**Datasets**

# Kaggleの特徴

**Competitions**

**Kernels**

**Discussion**

**Datasets**



# Competitions

## Welcome to Kaggle Competitions

Challenge yourself with real-world machine learning problems



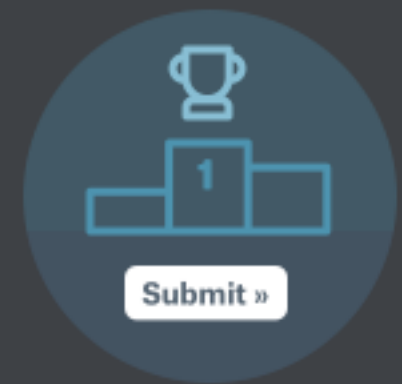
### New to Data Science?

Get started with a tutorial on our most popular competition for beginners, [Titanic: Machine Learning from Disaster](#).



### Build a Model

Get the data & use whatever tools or methods you prefer to make predictions.



### Make a Submission

Upload your prediction file for real-time scoring & a spot on the leaderboard.

[Learn more](#)[🎓 InClass](#)



# Competitions

コンペです。

Kaggleの特徴は、

- ・ 実業務で使用しているデータがダウンロードできる
- ・ 参考になるコードが実行可能な形で置かれている
- ・ 議論も行われている
- ・ チームが組める

上記により、参加しなくても参考になるデータ、コードが利用できるのが特徴


# Competitions

Do2dle

16 active competitions

All Categories

Search competitions




**Passenger Screening Algorithm Challenge**

Improve the accuracy of the Department of Homeland Security's threat recognition algorithms

Featured · 2 months to go · terrorism, image, object detection

\$1,500,000

312 teams




**Zillow Prize: Zillow's Home Value Prediction (Zestimate)**

Can you improve the algorithm that changed the world of real estate?

Featured · 3 months to go · housing, real estate · Entered

\$1,200,000

3,706 teams




**Cdiscount's Image Classification Challenge**

Categorize e-commerce photos

Featured · 2 months to go · multiclass classification

\$35,000

205 teams




**Porto Seguro's Safe Driver Prediction**

Predict if a driver will file an insurance claim next year.

Featured · 2 months to go · tabular, binary classification

\$25,000

1,580 teams



**Web Traffic Time Series Forecasting**

Forecast future traffic to Wikipedia pages

Research · a month to go · time series, internet, tabular, forecasting

\$25,000

1,095 teams




**Text Normalization Challenge - English Language**

Convert English text from written expressions into spoken forms

Research · a month to go · languages, linguistics, text · Entered

\$25,000

351 teams



**Text Normalization Challenge - Russian Language**

Convert Russian text from written expressions into spoken forms

Research · a month to go · languages, linguistics, text · Entered

\$25,000

144 teams

下記の1～2が用意されており、3～7を実施する

1. 実施内容の決定



2. データ入手



3. データ前処理



4. 手法選択



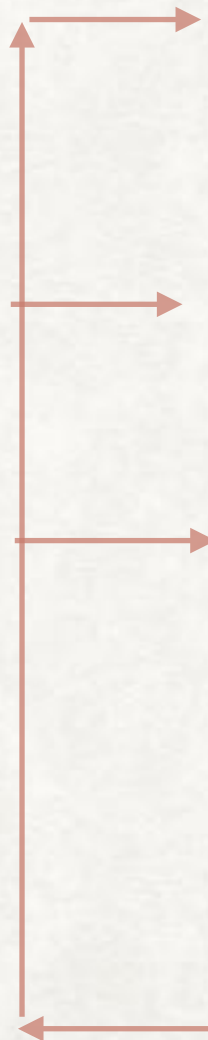
5. ハイパーパラメータ選択



6. モデルの学習



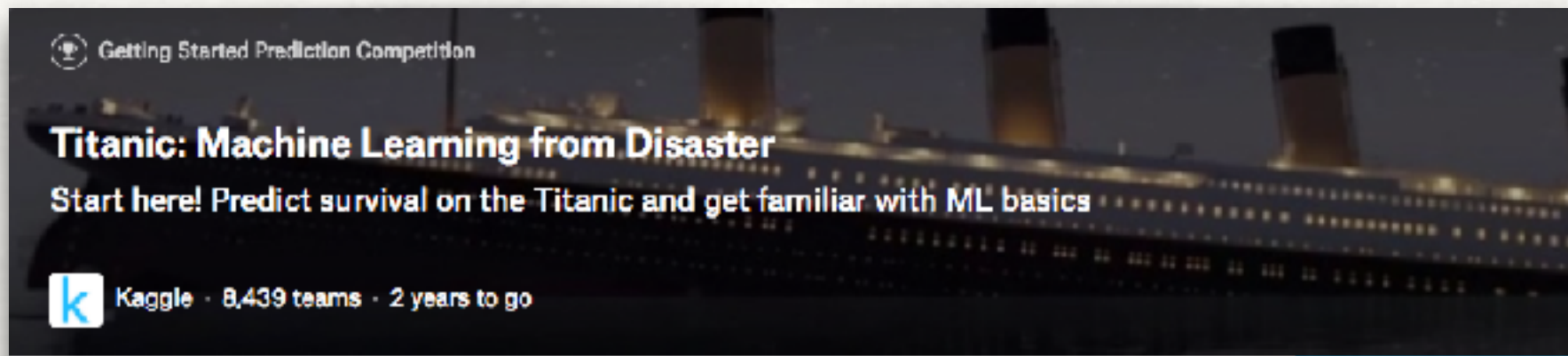
7. モデルの評価





# Competitions


Do2dle



Getting Started Prediction Competition

## Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

 Kaggle · 8,439 teams · 2 years to go

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#)

[Overview](#)

[Description](#)

[Evaluation](#)

[Frequently Asked Questions](#)

[Tutorials](#)

### Start here if...

You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.

### Competition Description

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

### Practice Skills

- Binary classification
- Python and R basics

# Titanic: Machine Learning from Disaster

- Kaggleのチュートリアル
- 乗客がタイタニックの沈没を生き延びたかどうかを予測し、この精度を競う
- 訓練用データ (891行 × 12列のcsv)      データに一部欠損あり
- テストデータ (418行 × 11列のcsv)      データに一部欠損あり

# Titanic: Machine Learning from Disaster

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

詳細は一から始める機械学習（Kaggleで学ぶ機械学習）

<https://speakerdeck.com/tflare/machine-learning-to-learn-at-kaggle>



# Titanic: Machine Learning from Disaster

- PassengerId : データにシーケンシャルでついている番号
- Survived : 生存 (0 = No, 1 = Yes)      訓練用データにのみ存在
- Pclass : チケットのクラス (1 = 1st, 2 = 2nd, 3 = 3rd)
- Name : 名前
- Sex : 性別
- Age : 年齢

# Titanic: Machine Learning from Disaster

- SibSp : タイタニック号に乗っていた兄弟と配偶者の数
- Parch : タイタニック号に乗っていた両親と子どもの数
- Ticket : チケット番号
- Fare : 旅客運賃
- Cabin : 船室番号
- Embarked : 乗船場 (C = Cherbourg, Q = Queenstown, S = Southampton)

# Kernels

Do2dle

## Welcome to Kaggle Kernels

The best place to explore data science results and share your own work

```
# Load libraries
library(xgboost)
library(Matrix)
input_dir = "
```

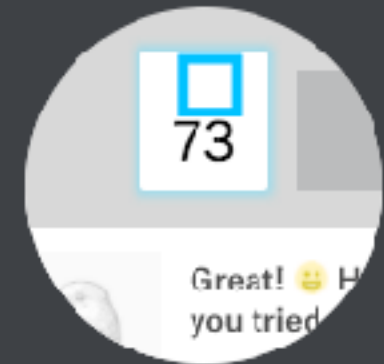
### Code

Skip the download. Kernels is preloaded with the most common data science languages and libraries.



### Learn

Gain exposure to new tools and techniques. The "hottest" kernels showcase the best work on Kaggle.



### Mentor

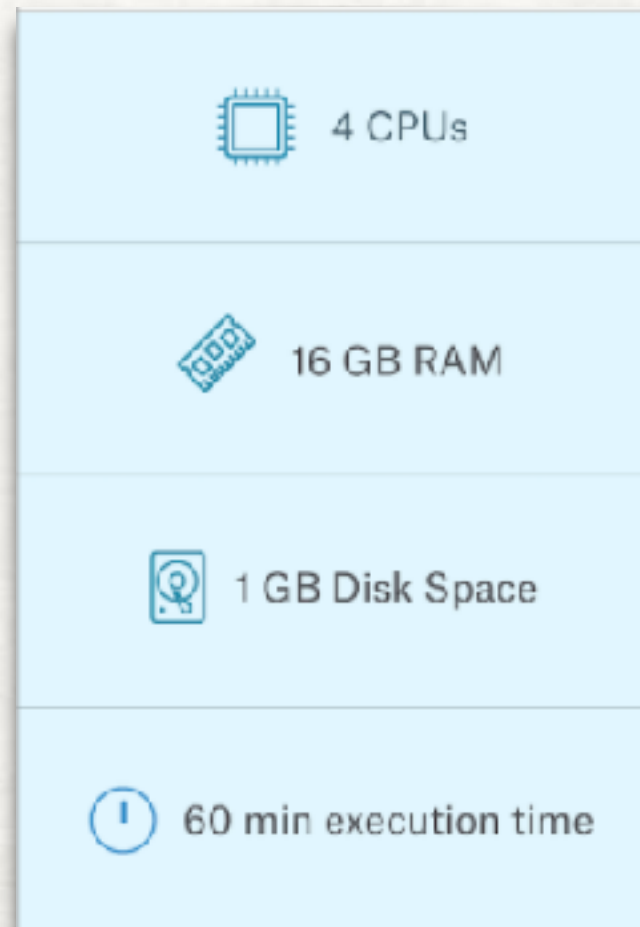
Give back by sharing what you know. You can answer questions and leave feedback on others' code and results.

[New Kernel](#)



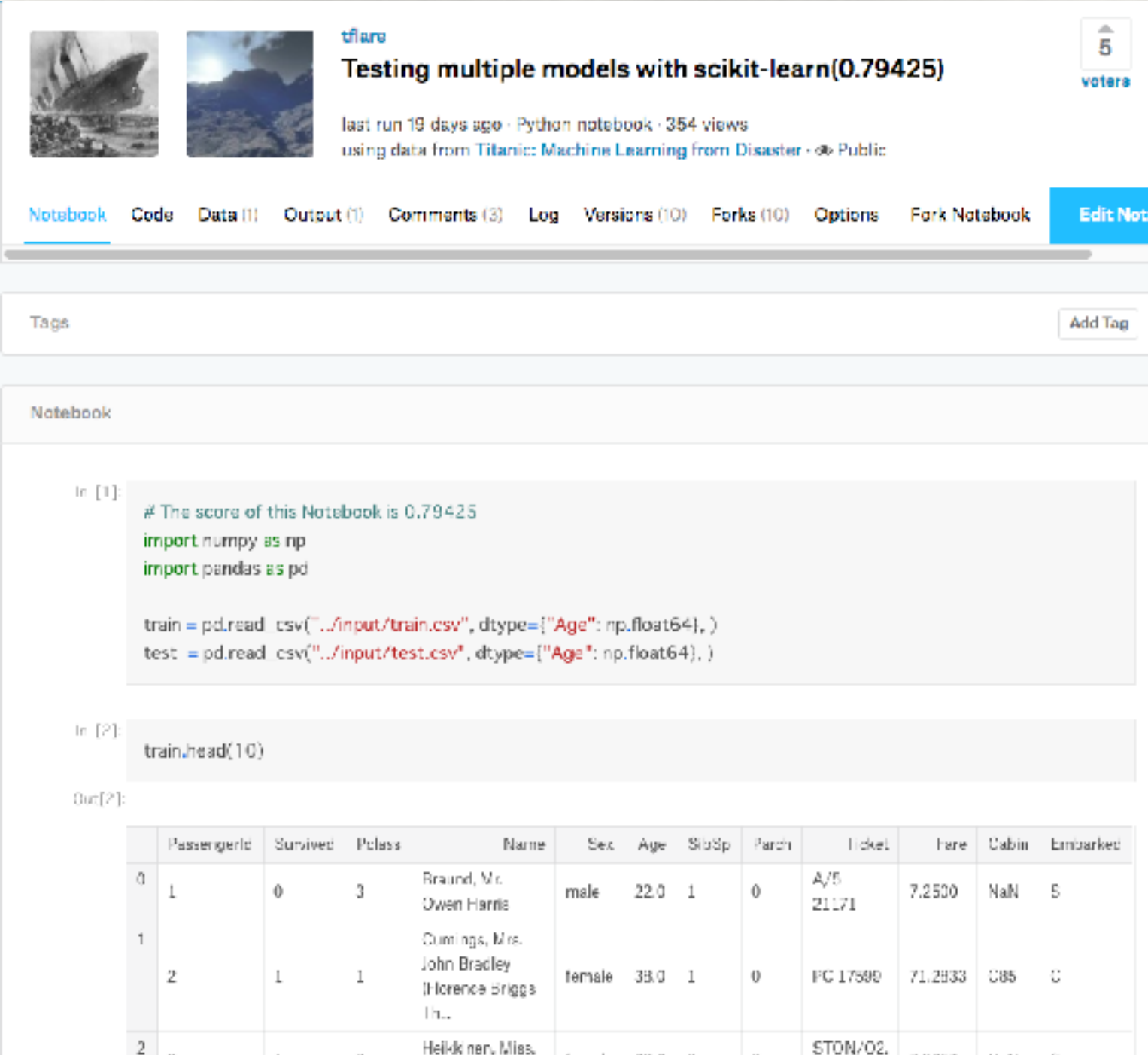
# Kernels

環境が整ったクラウドでコードを実行可能、  
他の人のコードもKernelをForkすることで実行可能  
カーネルに対してコメントすることが可能  
実行時間は最大60分まで



# Kernels

Do2dle



**tflare**  
**Testing multiple models with scikit-learn(0.79425)**  
last run 19 days ago · Python notebook · 354 views  
using data from [Titanic: Machine Learning from Disaster](#) · Public

5 voters

Notebook Code Data (1) Output (1) Comments (3) Log Versions (10) Forks (10) Options Fork Notebook Edit Notebook

Tags  Add Tag

Notebook

```
In [1]: # The score of this Notebook is 0.79425
import numpy as np
import pandas as pd

train = pd.read_csv("../input/train.csv", dtype={"Age": np.float64}, )
test = pd.read_csv("../input/test.csv", dtype={"Age": np.float64}, )

In [2]: train.head(10)
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cummings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	53.1000	NaN	S

<https://www.kaggle.com/tflare/testing-multiple-models-with-scikit-learn-0-79425>

# Kernels

```
In [3]: train_corr = train.corr()  
train_corr
```

Out[3]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.035144	0.036847	-0.057527	-0.001652	0.012658
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307
Pclass	-0.035144	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500
Age	0.036847	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067
SibSp	-0.057527	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651
Parch	-0.001652	0.081629	0.018443	-0.189119	0.414838	1.000000	0.215225
Fare	0.012658	0.257307	-0.549500	0.096067	0.159651	0.215225	1.000000

```
In [4]: #-1.0 to -0.7 Strong negative correlation  
#-0.7 to -0.4 Negative correlation  
#-0.4 to -0.2 Weak negative correlation  
#-0.2 to +0.2 There is no correlation  
#+0.2 to +0.4 Weak positive correlation  
#+0.4 to +0.7 Positive correlation  
#+0.7 to +1.0 Strong Positive correlation
```

```
In [5]: def correct_data(train_data, test_data):  
  
    # Make missing values for training data from test data as well  
    train_data.Age = train_data.Age.fillna(test_data.Age.median())  
    train_data.Fare = train_data.Fare.fillna(test_data.Fare.median())  
  
    test_data.Age = test_data.Age.fillna(test_data.Age.median())  
    test_data.Fare = test_data.Fare.fillna(test_data.Fare.median())  
  
    train_data = correct_data_common(train_data)  
    test_data = correct_data_common(test_data)
```



# Kernels

```
In [7]: from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC, LinearSVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.neural_network import MLPClassifier

from sklearn.model_selection import cross_val_score

predictors = ["Pclass", "Sex", "Age", "SibSp", "Parch", "Fare", "Embarked"]

models = []

models.append(("LogisticRegression", LogisticRegression()))
models.append(("SVC", SVC()))
models.append(("LinearSVC", LinearSVC()))
models.append(("KNeighbors", KNeighborsClassifier()))
models.append(("DecisionTree", DecisionTreeClassifier()))
models.append(("RandomForest", RandomForestClassifier()))
rf2 = RandomForestClassifier(n_estimators=100, criterion='gini',
                           max_depth=10, random_state=0, max_features=None)
models.append(("RandomForest2", rf2))
models.append(("MLPClassifier", MLPClassifier(solver='lbfgs', random_state=0)))

results = []
names = []
for name, model in models:
    result = cross_val_score(model, train_data[predictors], train_data["Survived"], cv=3)
    names.append(name)
    results.append(result)

for i in range(len(names)):
    print(names[i], results[i].mean())
```

# Kernels

```
In [8]:  
alg = rf2  
  
alg.fit(train_data[predictors], train_data["Survived"])  
  
predictions = alg.predict(test_data[predictors])  
  
submission = pd.DataFrame({  
    "PassengerId": test_data["PassengerId"],  
    "Survived": predictions  
})  
  
submission.to_csv('submission.csv', index=False)
```

Comments (3)

All Comments

Sort by

Hotness



Click here to enter a comment...



KarthickVadivel • Posted on Version 4 • a month ago • Options • Reply

1



very nice..... :). In filling missing values for testing data , you should take median from training data rather than test data


























tflarc • Posted on Version 5 • a month ago • Options • Edit • Reply

0

Thank you. The result increased from 0.77990 to 0.79426. I also updated the code.

# Discussion

Do2dle

28,887 topics		Sort by Most Votes	
All	Mine	Upvoted	All Categories Topics
		Search topics	
670		 This is insane discrimination and insult to our international com... Sergey Mushinskiy 4mo ago in Passenger Screening Algorithm Challenge	last comment by Tili 2d ago 144
297		 Data Scientist Hero Gilberto Titericz Junior 1y ago in Kaggle Forum	last comment by Rio Harapan Ps. 1mo ago 136
244		 1st PLACE - WINNER SOLUTION - Gilberto Titericz & Stanislav ... Gilberto Titericz Junior 2y ago in Otto Group Product Classification Challen...	last comment by Menelaos Kanakis 2mo ago 88
224		 The 'Magic' (Leak) feature is attached Μαριος Μιχαηλιδης Kazanova 6mo ago in Two Sigma Connect...	last comment by Dmitry Ukrainskiy 5mo ago 120
186		 1st place solution Maximilien@DAMI 4mo ago in Quora Question Pairs	last comment by TCH 2mo ago 51
178		What tools do people generally use to solve problems? Shravana Aadith R B 5y ago in Getting Started	last comment by mtax 11d ago 214
178		 3rd-Place Solution Overview sjv 2mo ago in Instacart Market Basket Analysis	last comment by SeanZhang 8d ago 66
157		 Beat the benchmark with less than 1MB of memory. tintgu 3y ago in Click-Through Rate Prediction	last comment by Swapnil 2y ago 184
149		 Score 0.53778 (or 0.52879) using StackNet Μαριος Μιχαηλιδης Kazanova 7mo ago in Two Sigma Connect...	last comment by Sameh Faidi 5mo ago 151
144		 #1 Dexter's Lab winning solution raddar 1y ago in BNP Paribas Cardif Claims Management	last comment by Ayush Raj Singh 1y ago 75
142		 Proper validation framework CV 0.32x -> LB 0.32x raddar 5mo ago in Sberbank Russian Housing Market	last comment by Pietro Marinelli 3mo ago 41
138		 Analysis of Duplicate Variables   Correlated Variables (large post)	last comment by 45



# Discussion

議論ができる場です。

出題元からの説明

過去のコンペでは成績優秀者の説明、コードへのリンクなどがあります。

# Leaderboard

順位が確認できる場所です。

- この順位はCompetitions毎に決められたルールで設定されます。  
そのため、最終的な順位と異なる場合があります。

**Titanic: Machine Learning from Disaster**の場合は以下です。

「This leaderboard is calculated with approximately 50% of the test data.

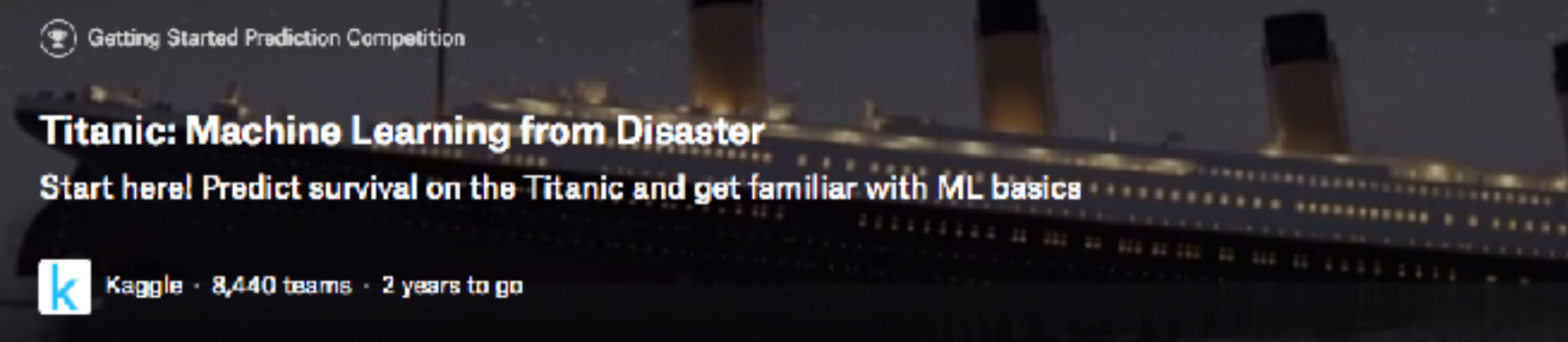
The final results will be based on the other 50%, so the final standings may be different.」

上記により、最適化しすぎるのではなく、一般化することを狙っていると思われます。

- 提出できる回数は限度があります。（Zillow Prizeは1日5回まで）

# Leaderboard


Do2dle



Getting Started Prediction Competition

## Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

 Kaggle · 8,440 teams · 2 years to go

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#)

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
submission.csv	a month ago	1 seconds	0 seconds	0.77511




**Complete**

[Jump to your position on the leaderboard](#)

[Public Leaderboard](#) [Private Leaderboard](#)

This leaderboard is calculated with approximately 50% of the test data.  
The final results will be based on the other 50%, so the final standings may be different.

[Raw Data](#) [Refresh](#)

#	Δ1w	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	SwarnenduMallick			1.00000	1	2mo
2	—	Nikhil Mathew			1.00000	41	2mo
3	—	Shahnab			1.00000	1	2mo



# Kaggleの特徴

**Competitions**

**Kernels**

**Discussion**

**Datasets**

# Datasets

Do2dle

kaggle



Competitions

Datasets

Kernels

Discussion

Jobs

...



## Welcome to Kaggle Datasets

The best place to discover and seamlessly analyze open data



### Discover

Use the search box to find open datasets on everything from government, health, and science to popular games and dating trends.



### Explore

Execute, share, and comment on code for any open dataset with our in-browser analytics tool, [Kaggle Kernels](#). You can also download datasets in an easy-to-read format.



### Create a Dataset

Contribute to the open data movement and connect with other data enthusiasts by clicking "[New Dataset](#)" to publish an open dataset of your own.

[Learn More](#)

[New Dataset](#)

# Datasetsの3つの機能

Datasetsでは、以下3つの機能があります。

- データセットを探す
- データセットについて探究する  
(ダウンロードする、コードを書く、コメントをする)
- 新しくデータセットを作る

具体的なデータセットを通じてみてみましょう。




# Complete FIFA 2017 Player dataset (Global)

Featured Dataset

## Complete FIFA 2017 Player dataset (Global)




15k+ players, 50+ Attributes per player from the latest EA Sports Fifa 17

 Soumitra Agarwal • last updated 6 months ago

102

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Activity](#) [Download \(6 MB\)](#) [New Kernel](#)

Tags: [video games](#) [association football](#) [small](#) [featured](#)

Top Contributors	Kernels	Discussion
 Selfish Gene 1st	<a href="#">Analyzing Soccer Player Faces</a> 131 votes run 18 days ago	<a href="#">Exploring FIFA 2017 dataset</a> 19 replies 13 days ago
 Hitesh palamada 2nd	<a href="#">Exploring FIFA 2017 dataset</a> 28 votes run a month ago	<a href="#">Analyzing Soccer Player Fa...</a> 38 replies 17 days ago
 Soumitra Agarwal 3rd	<a href="#">Making a Super Team I</a> 9 votes run 4 months ago	<a href="#">the "rating" column</a> 0 replies 24 days ago

<https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global>

# データセットの内容

Do2dle

## The dataset for people who double on Fifa and Data Science

### Content

- 17,000+ players
- 50+ attributes per player ranging from ball skills aggression etc.
- Player's attributes sourced from EA Sports' FIFA video game series, including the weekly updates
- Players from all around the globe
- URLs to their homepage
- Club logos
- Player images male and female
- National and club team data

### Weekly Updates would include :

- Real life data (Match events etc.)
- The fifa generated player dataset
- Betting odds
- Growth

# データセットの内容

Do2dle

ClubNames.csv

FullData.csv

NationalNames....

PlayerNames.csv

ClubPictures.zip

Antalyaspor.png

Arsenal.png

AS Monaco.png

Atletico Madrid.png

Blackburn Rovers.png

Bor. Dortmund.png

Bor. Mgladbach.png

Chelsea.png

Crystal Palace.png

Deport. Alava.png

(37 total)

Pictures.zip

Aaron Ramsey.png

Adam Lallana.png

Adil Rami.png

Adn.png

Adrien Rabiot.png

Adrien Silva.png

Aduriz.png

Alan Dzagoev.png

Alejandro Gmez.png

Aleksandar Kolarov....

## FullData.csv

1.33 MB • Updated 6 months ago

Download

About this file

Details of all the active players fifa 17

Preview (first 100 rows)

Column Metadata

36

Name	Nationality	National_Position	National_Kit	Club	Club_Position	Club_Kit	Club_
Cristiano Ronaldo	Portugal	LS	7.0	Real Madrid	LW	7.0	07/01
Lionel Messi	Argentina	RW	10.0	FC Barcelona	RW	10.0	07/01
Neymar	Brazil	LW	10.0	FC Barcelona	LW	11.0	07/01
Luis Suárez	Uruguay	LS	9.0	FC Barcelona	ST	9.0	07/11,
Manuel Neuer	Germany	GK	1.0	FC Bayern	GK	1.0	07/01
De Gea	Spain	GK	1.0	Manchester Utd	GK	1.0	07/01
Robert Lewandowski	Poland	LS	9.0	FC Bayern	ST	9.0	07/01
Gareth Bale	Wales	RS	11.0	Real Madrid	RW	11.0	09/01
Zlatan Ibrahimović	Sweden			Manchester Utd	ST	9.0	07/01
Thibaut Courtois	Belgium	GK	1.0	Chelsea	GK	13.0	07/26
Jérôme Boateng	Germany	RCB	17.0	FC Bayern	Sub	17.0	07/14,
Eden Hazard	Belgium	LF	10.0	Chelsea	LW	10.0	07/01
Luka Modrić	Croatia			Real	RCM	19.0	08/01



# データセットの機能

Do2dle

データセット毎に、Kernels、Discussionがあつて、実行可能なコード、ディスカッションが行われている。

The screenshot shows the Do2dle interface for a dataset. At the top, it says 'Featured Dataset' and 'Complete FIFA 2017 Player dataset (Global)' with a subtitle '15k+ players, 50+ Attributes per player from the latest EA Sports Fifa 17'. The creator is 'Soumitra Agarwal' and it was 'last updated 6 months ago'. There are 102 upvotes. Below this is a navigation bar with 'Overview' (selected), 'Data', 'Kernels', 'Discussion', and 'Activity'. To the right are 'Download (6 MB)' and a 'New Kernel' button. A 'Tags' section includes 'video games', 'association football', 'small', and 'featured'. The main content area is divided into three columns: 'Top Contributors', 'Kernels', and 'Discussion'. Each column has a list of items with their respective scores and dates.

Top Contributors	Kernels	Discussion
Selfish Gene 1st	<a href="#">Analyzing Soccer Player Faces</a> 131 votes run 18 days ago	<a href="#">Exploring FIFA 2017 dataset</a> 19 replies 13 days ago
Hitesh palamada 2nd	<a href="#">Exploring FIFA 2017 dataset</a> 28 votes run a month ago	<a href="#">Analyzing Soccer Player Fa...</a> 38 replies 17 days ago
Soumitra Agarwal 3rd	<a href="#">Making a Super Team I</a> 9 votes run 4 months ago	<a href="#">the "rating" column</a> 0 replies 24 days ago

# どのようにすすめればよいか

まずは、Titanic: Machine Learning from Disasterをやり  
あとは気になったCompetitionsをやるです。  
私は以下をやっています。

- Zillow Prize: Zillow's Home Value Prediction (Zestimate)
- Text Normalization Challenge - English Language
- Text Normalization Challenge - Russian Language

まだやっていませんが、以下はデータ量が少なくて始めやすそうです

- Porto Seguro's Safe Driver Prediction

# どのようにすすめればよいか

- Cdiscount's Image Classification Challenge

上記はTrainデータが58.19 GB、Testデータが14.53 GBあります。  
データ量の多い、画像関連のCompetitionsは性能が高いGPUがない場合、上位に行くことは困難です。

(AWS、GCP等でGPUを搭載しているインスタンスを借りても良いですが)

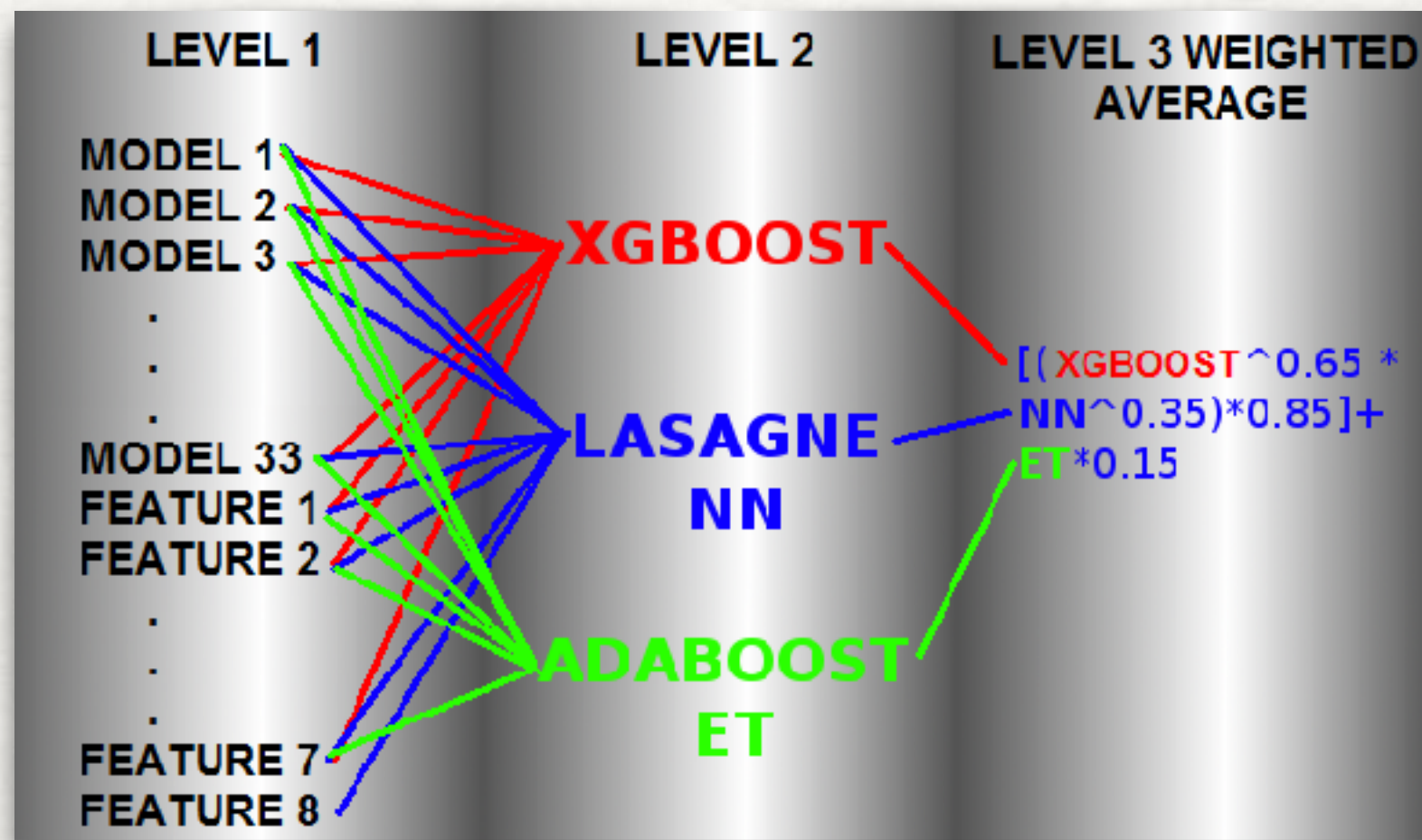


# Kaggleで有効な手法の紹介

Otto Group Product Classification Challenge

商品の特徴（93種類）から商品をカテゴリ分けする課題    trainデータ6万    testデータ14万

以下は優勝者のモデルです。



1st PLACE - WINNER SOLUTION - Gilberto Titericz & Stanislav Semenov

<https://www.kaggle.com/c/otto-group-product-classification-challenge/discussion/14335>

# Kaggleで有効な手法の紹介

33モデルを組み合わせて精度を実現しています。

機械学習の研究・実践ではシンプルな手法が良いが、

Kaggleでは勝てればなんでもよい。

→ 他の人との差別化のために、複数モデルの組み合わせを行うことがよくある。

# Kaggleで有効な手法の紹介

## ノーフリーランチ定理

どのような問題に対しても一番性能が良い万能アルゴリズムはない。

複数の識別器を組み合わせて、性能を上げる方法（アンサンブル学習）の検討が前から行われている。



# Kaggleで有効な手法の紹介

**Otto Group Product Classification Challenge**  
でも使われており、今でも使われることが多い

以下2つの説明をします。

- **Xgboost**
- **Stacked generalization**

# Xgboostとは

**Gradient Boosting（勾配ブースティング）の高速な実装です。**

**Kaggleで最も人気のある機械学習手法です。**

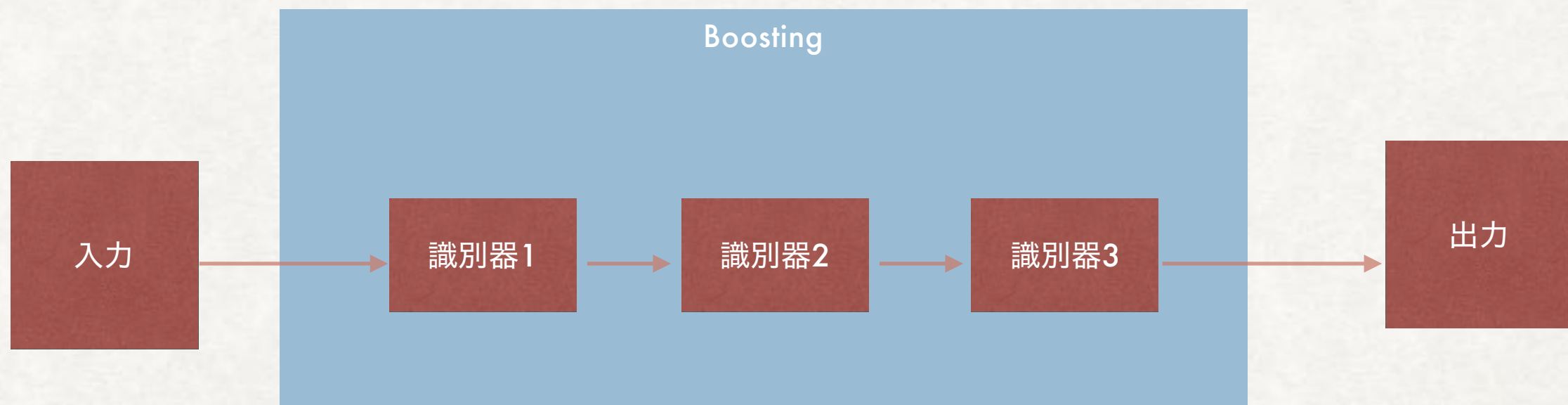
**利点：早い、外れ値や欠損値に強い、線形分離不可能パターンに強い**

**欠点：線形分離可能パターンに弱い**

# Boostingとは

**Boosting（ブースティング）は、複数の弱識別器を用意して、学習を直列的にし、前の弱識別器の学習結果を参考にしながら一つずつ弱識別器を学習する。...**

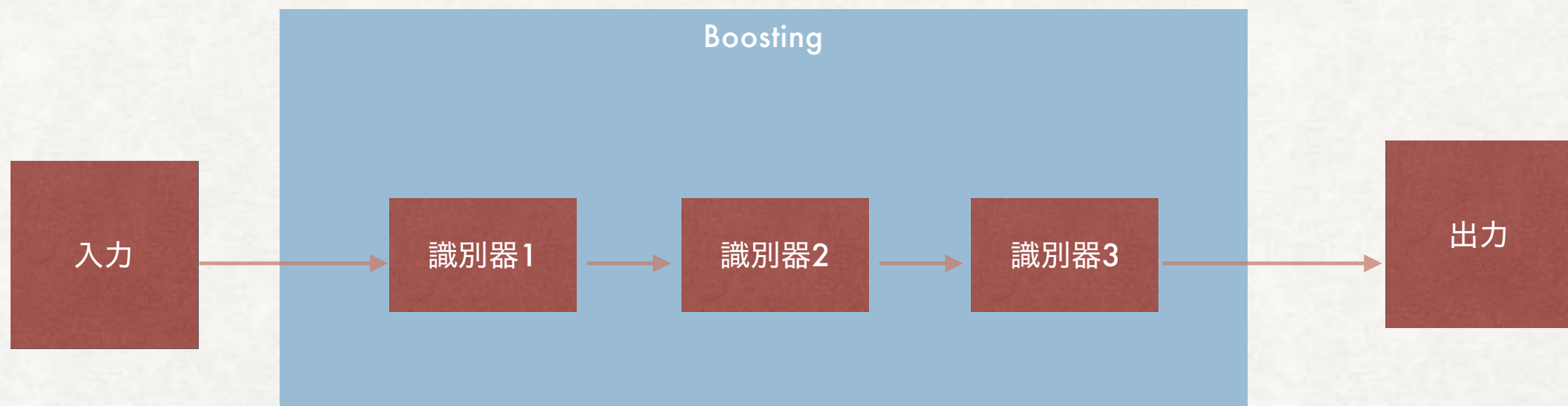
**平井 有三（2012）はじめてのパターン認識 p188 森北出版**





# Gradient Boosting とは

Gradient Boosting（勾配ブースティング）は、簡単に言うとBoostingの各ステップのパラメタ最適化の際に、勾配降下法を用いる方法



# Xgboostの強み

deep learningは画像、音声、テキスト処理などに力を発揮する。

樹木モデル（Xgboostも含む）は表形式のデータ処理に力を発揮する。

上記は両方が重要であり、状況に応じて使い分ける必要がある。

# もっと Gradient Boosting

A Kaggle Master Explains Gradient Boosting

<http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/>

XGBoostについては以下を参照ください。

Introduction to Boosted Trees (公式ドキュメントと元になった資料)

<http://xgboost.readthedocs.io/en/latest/model.html>

<https://homes.cs.washington.edu/~tqchen/pdf/BoostedTree.pdf>

Github

<https://github.com/dmlc/xgboost>



# 主要なアンサンブル学習

Kaggleではアンサンブル学習がよく使われています。

主要なアンサンブル学習は以下の5つです。

- Voting
- Averaging
- Rank averaging
- Stacked generalization
- Blending

# 主要なアンサンブル学習

Kaggleではアンサンブル学習がよく使われています。

主要なアンサンブル学習は以下の5つです。

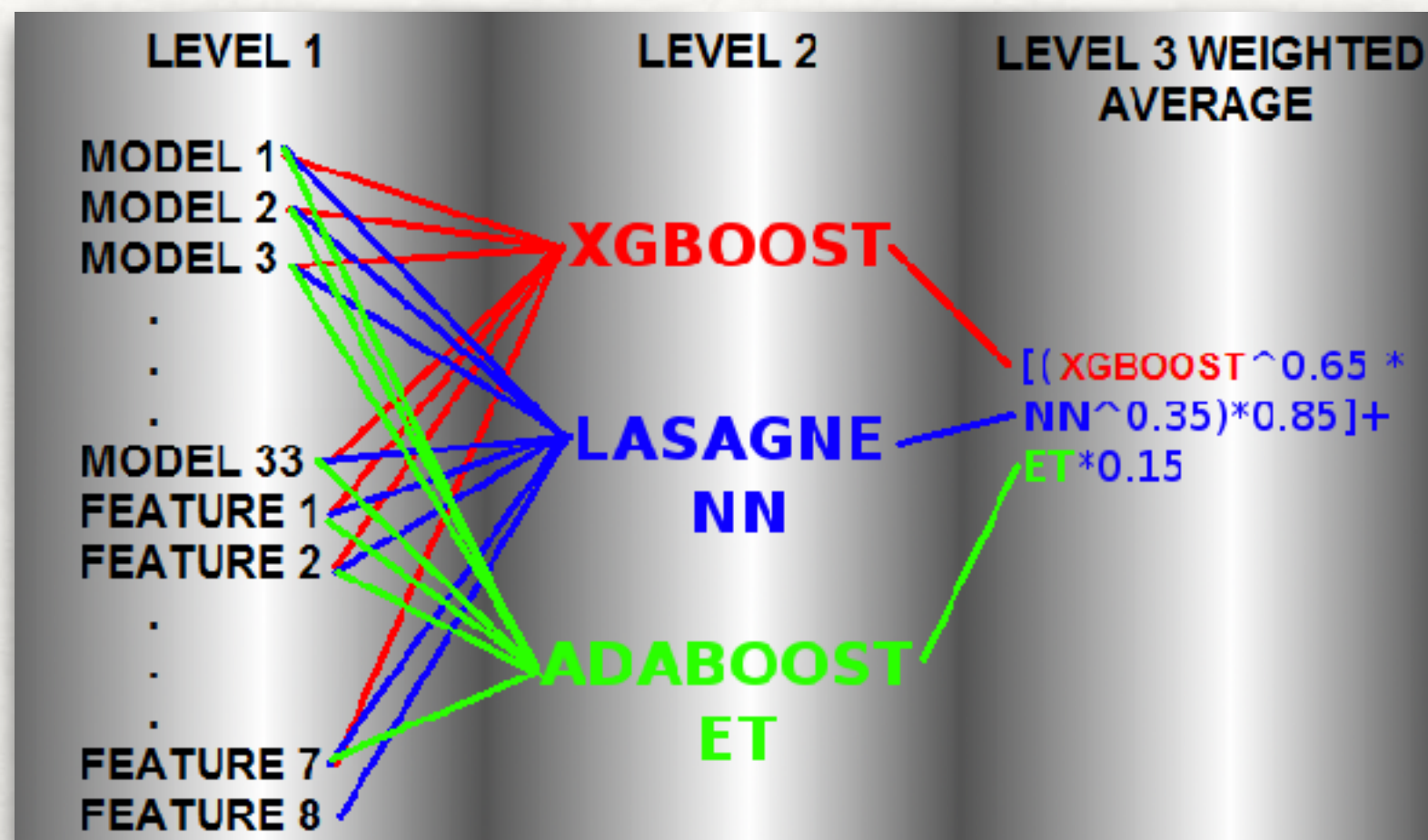
- Voting
- Averaging
- Rank averaging
- Stacked generalization
- Blending

今回は上記の中からよく使われる  
Stacked generalizationについて説明します。

# Stacked generalizationとは

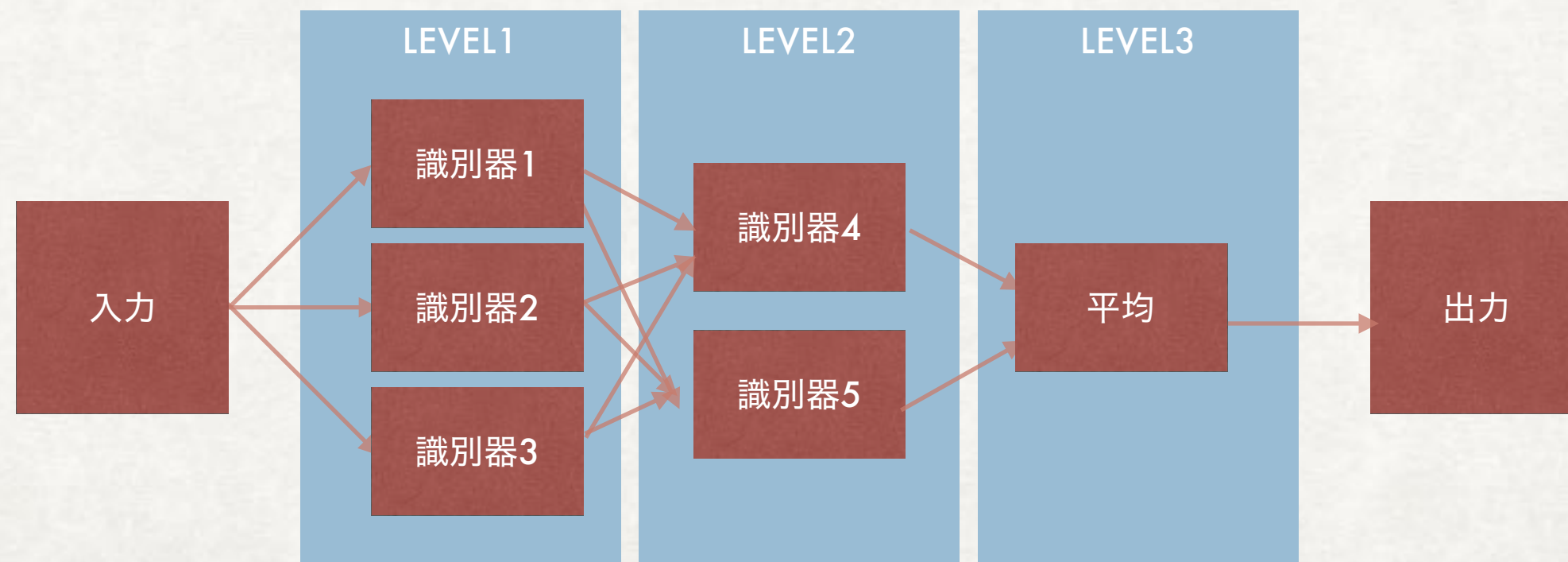
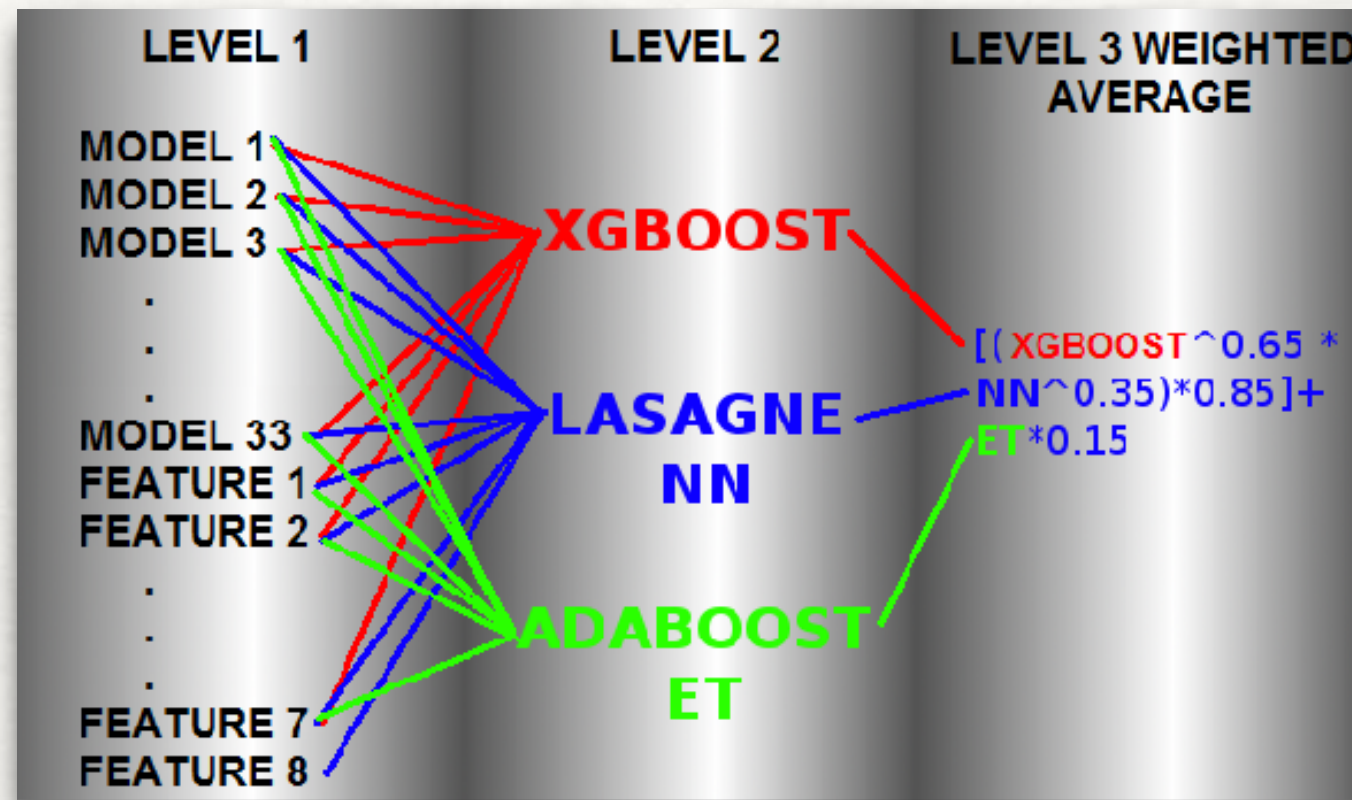
アンサンブル学習の中でKaggleで成績上位者の中でよく使われるのが、  
Stacked generalizationです

Stacked generalizationの基本的な考え方は、識別器を組み合わせ、より  
精度の良いモデルをつくるというもの

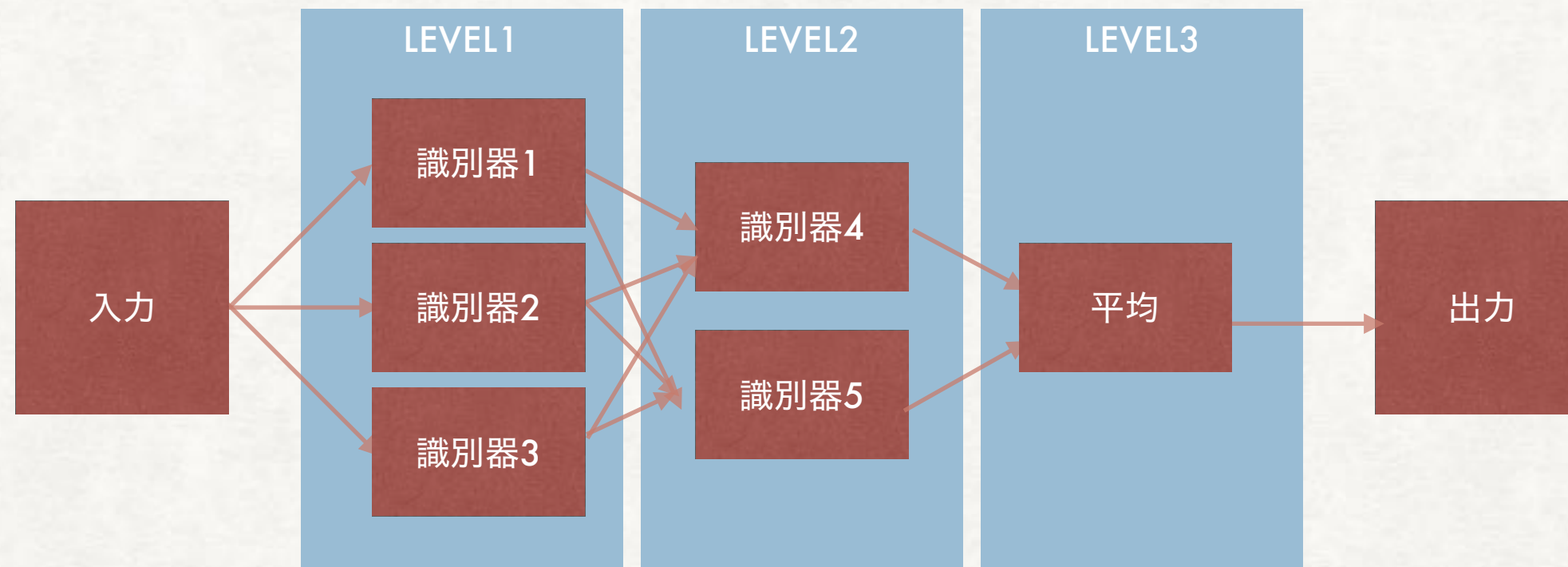




# Stacked generalizationとは



# Stacked generalizationとは



異なる種類の識別器を組み合わせて、より性能の高いモデルを作成する。  
最終的に平均などで結果を決める。

Stacked generalizationは自由度が高い手法で、

Kaggleなどのコンテストで、他の人との差別化を行うために使われている。

# もっとStacked generalization

stacked generalization

<http://puyokw.hatenablog.com/entry/2015/12/12/090000>

論文

<http://www.cs.utsa.edu/~bylander/cs6243/wolpert92stacked.pdf>



# 主要なアンサンブル学習

Kaggleではアンサンブル学習がよく使われています。

主要なアンサンブル学習は以下の5つです。

- Voting
- Averaging
- Rank averaging
- Stacked generalization
- Blending

# 他のアンサンブル手法

## Voting

Original signal:

1110110011

Encoded:

10,3 101011001111101100111110110011

Decoding:

1010110011

1110110011

1110110011

Majority vote:

1110110011

他のアンサンブル手法を一例紹介させていただきました。

興味がある方はKAGGLE ENSEMBLING GUIDEの参照をお願いします。

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

# 参考になるサイト

以下が参考になります。

- No Free Hunch (The Official Blog of Kaggle.com)

WINNERS' INTERVIEWS

<http://blog.kaggle.com/category/winners-interviews/>

- kaggle\_memo

[https://github.com/nejumi/kaggle\\_memo](https://github.com/nejumi/kaggle_memo)