機器學習導論

Lecture 6 Kaggle 競賽介紹

Kaggle

- Kaggle 是一個資料科學競賽的平台，很適合大家實踐課堂上所學習的機器學習內容

kaggle Search Competitions Datasets Notebooks Discussion Courses •••

Competitions

Documentation InClass

General | InClass | Hosted

Sort by Grouped

All Categories Search competitions

1 Entered Competition

Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started · Ongoing · tutorial, tabular data, binary classification

Knowledge
15,569 teams

14 Active Competitions

2019 Data Science Bowl
Uncover the factors to help measure how young children learn
Featured · Code Competition · 2 months to go · video games, children, learning, education

$160,000
979 teams

企業或政府會在Kaggle上發佈競賽問題，通常是他們真正面臨的問題，也會提供獎金

2019 Data Science Bowl

Uncover the factors to help measure how y

Booz Allen Hamilton · 979 teams · 2 months to g 2 months to go until merger deadline)

當你提交你的比賽結果後，就可以在 Leaderboard 上看到自己的成績和排名

Overview Data Notebooks Discussion Leaderboard Rules

Join Competition

Public Leaderboard Private Leaderboard

This leaderboard is calculated with approximately 14% of the test data.
The final results will be based on the other 86%, so the final standings may be different.

Raw Data Refresh

In the money Gold Silver Bronze

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	x0x0w1			0.557	45	2d
2	poteman			0.553	46	9h
3	narsil			0.549	17	2d
4	Janey			0.546	33	7h
5	Looking for side job			0.545	19	1d

Kaggle

你也可以在 Discussion (討論區)中，和其他參賽者相互討論，吸取經驗！

Featured Code Competition

2019 Data Science Bowl

Uncover the factors to help measure how young children learn

DATA SCIENCE BOWL
Passion. Curiosity. Purpose.

Presented by Booz | Allen | Hamilton & kaggle

$160,000
Prize Money

Booz Allen · Booz Allen Hamilton · 979 teams · 2 months to go (2 months to go until merger deadline)

Overview | Data | Notebooks | Discussion | Leaderboard | Rules | New Topic

164 topics | Follow | Sort by Hotness

All | Mine | Upvoted | Search topics

85	Measure Up! media types introduction Cosimo Felline 18 days ago	last comment by Mc 3h ago	30
9	TweetChat - 2019 Data Science Bowl, November 14, 13:00 - 14:0... Josette_BoozAllen 10 days ago	last comment by Josette_BoozAllen 10d ago	0
38	Web-version of the PBS KIDS Measure Up! App Josette_BoozAllen 19 days ago	last comment by Robert Tacbad 5d ago	7

Kaggle 獎牌

- 成績優秀者可以獲得獎牌，這也是一種資歷喔！

Competition Medals

Competition medals are awarded for top competition results. The number of medals awarded per competition varies depending on the size of the competition. Note that InClass, playground, and getting started competitions do not award medals.

	0-99 Teams	100-249 Teams	250-999 Teams	1000+ Teams
Bronze	Top 40%	Top 40%	Top 100	Top 10%
Silver	Top 20%	Top 20%	Top 50	Top 5%
Gold	Top 10%	Top 10	Top 10 + 0.2%*	Top 10 + 0.2%*

* (Top 10 + 0.2%) means that an extra gold medal will be awarded for every 500 additional teams in the competition. For example, a competition with 500 teams will award gold medals to the top 11 teams and a competition with 5000 teams will award gold medals to the top 20 teams.

Kaggle

讓我們從入門級的 鐵達尼生存競賽問題
開始Kaggle 之旅吧

kaggle | Search | Competitions | Datasets | Notebooks | Discussion | Courses | •••

Competitions

Documentation | InClass

General | InClass | Hosted

Sort by Grouped

All Categories | Search competitions

1 Entered Competition

Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started · Ongoing · tutorial, tabular data, binary classification

Knowledge
15,569 teams

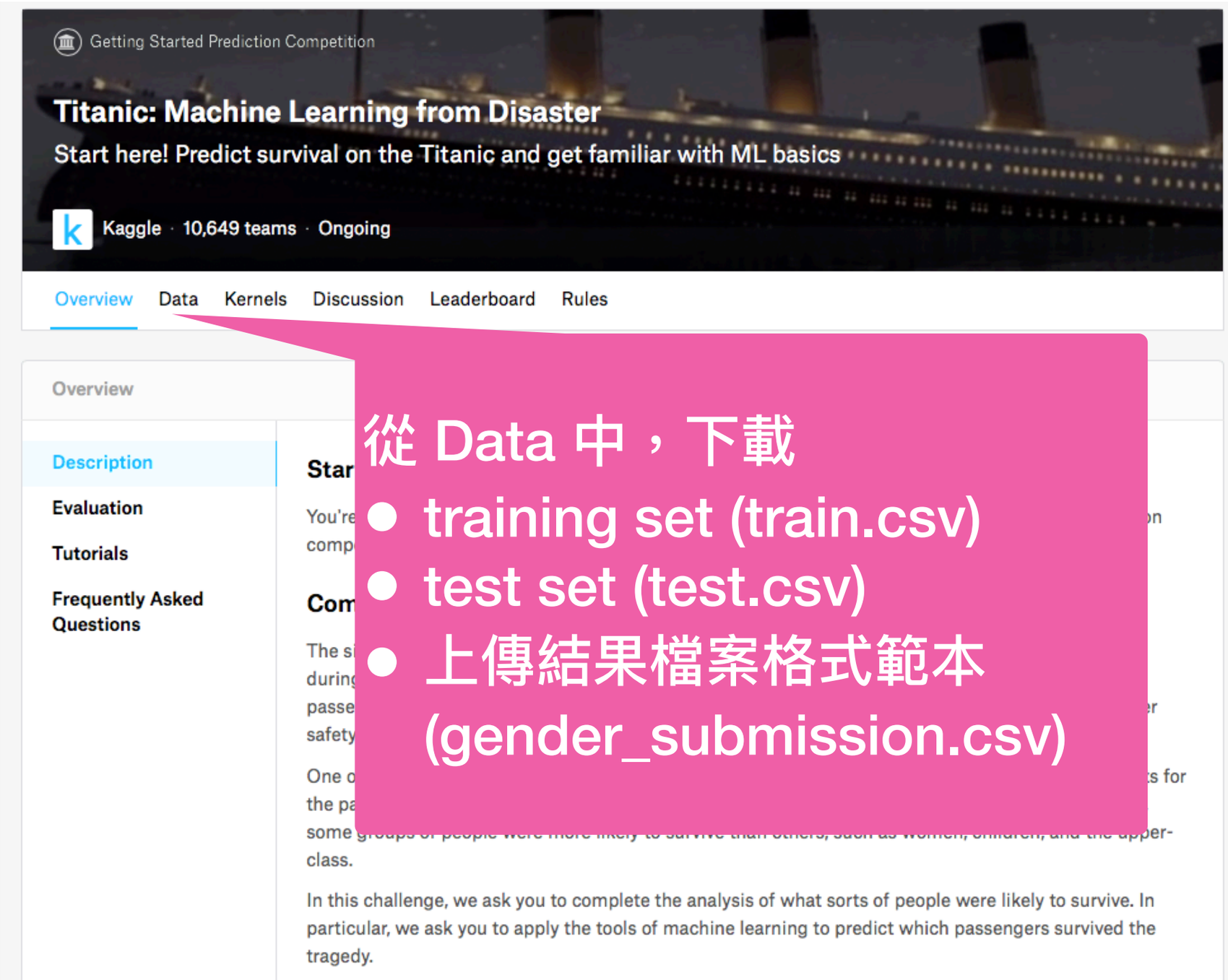
14 Active Competitions

2019 Data Science Bowl
Uncover the factors to help measure how young children learn
Featured · Code Competition · 2 months to go · video games, children, learning, education

$160,000
979 teams

Kaggle

- **Titanic: Machine Learning from Disaster**
  https://www.kaggle.com/c/titanic



從 Data 中，下載

- training set (train.csv)
- test set (test.csv)
- 上傳結果檔案格式範本 (gender_submission.csv)

Kaggle

- 下載數據集並觀察數據

```python
from sklearn import preprocessing
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# Loading the data
train = pd.read_csv("train.csv")
test = pd.read_csv("test.csv")
submit = pd.read_csv("gender_submission.csv")
# Observing the data
train.info()
test.info()
```

```
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
```

有missing values

Kaggle

- 由於要對整體資料做一些觀察，所以先合併資料。因為合併後index重複，因此將index重新設定

```
data = train.append(test)
data
data.reset_index(inplace=True, drop=True)
```

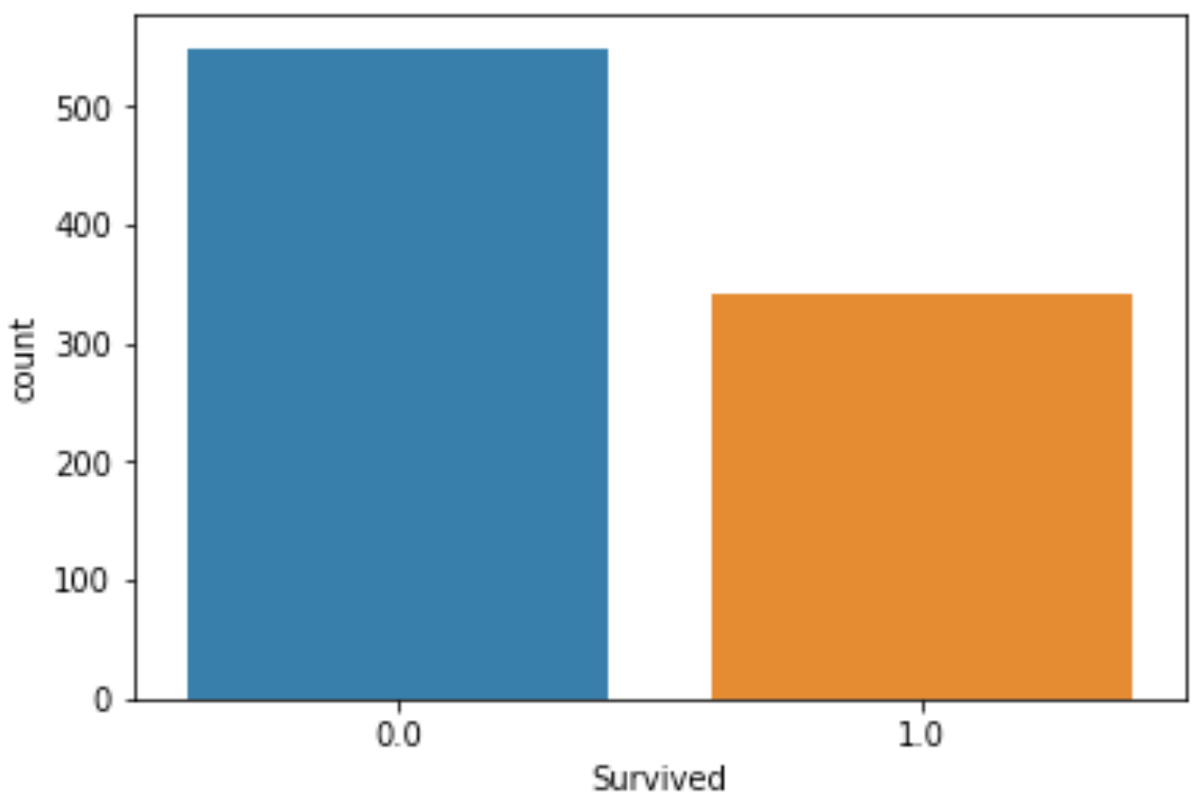
Kaggle

資料分析

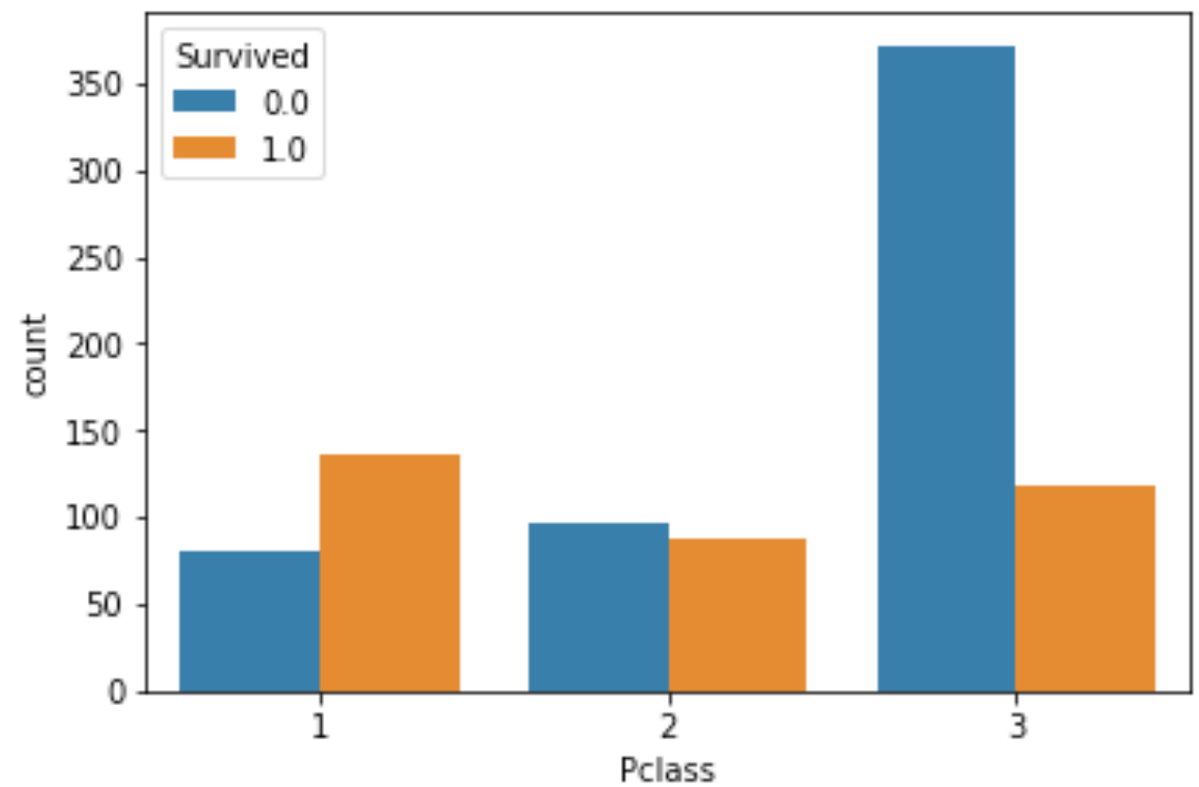
```
%matplotlib inline
sns.countplot(data['Survived'])
sns.countplot(data['Pclass'], hue=data['Survived'])
```

類別數據

觀察兩類別的比例是否差別很大



觀察船票等級和生存的關係



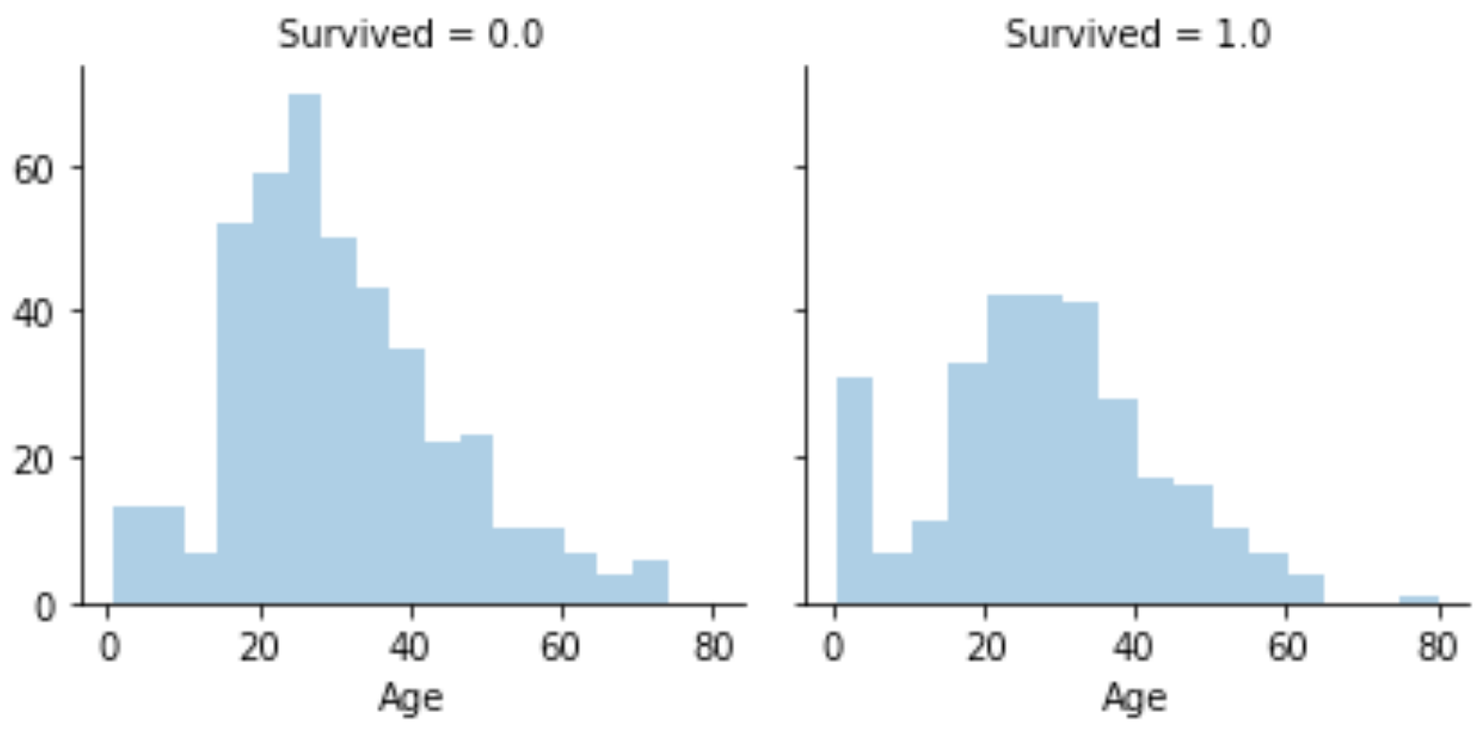
Kaggle

- 資料分析

```
g = sns.FacetGrid(data,col='Survived')
g.map(sns.distplot, 'Age', kde=False)
```

數值數據

觀察年齡和生存的關係

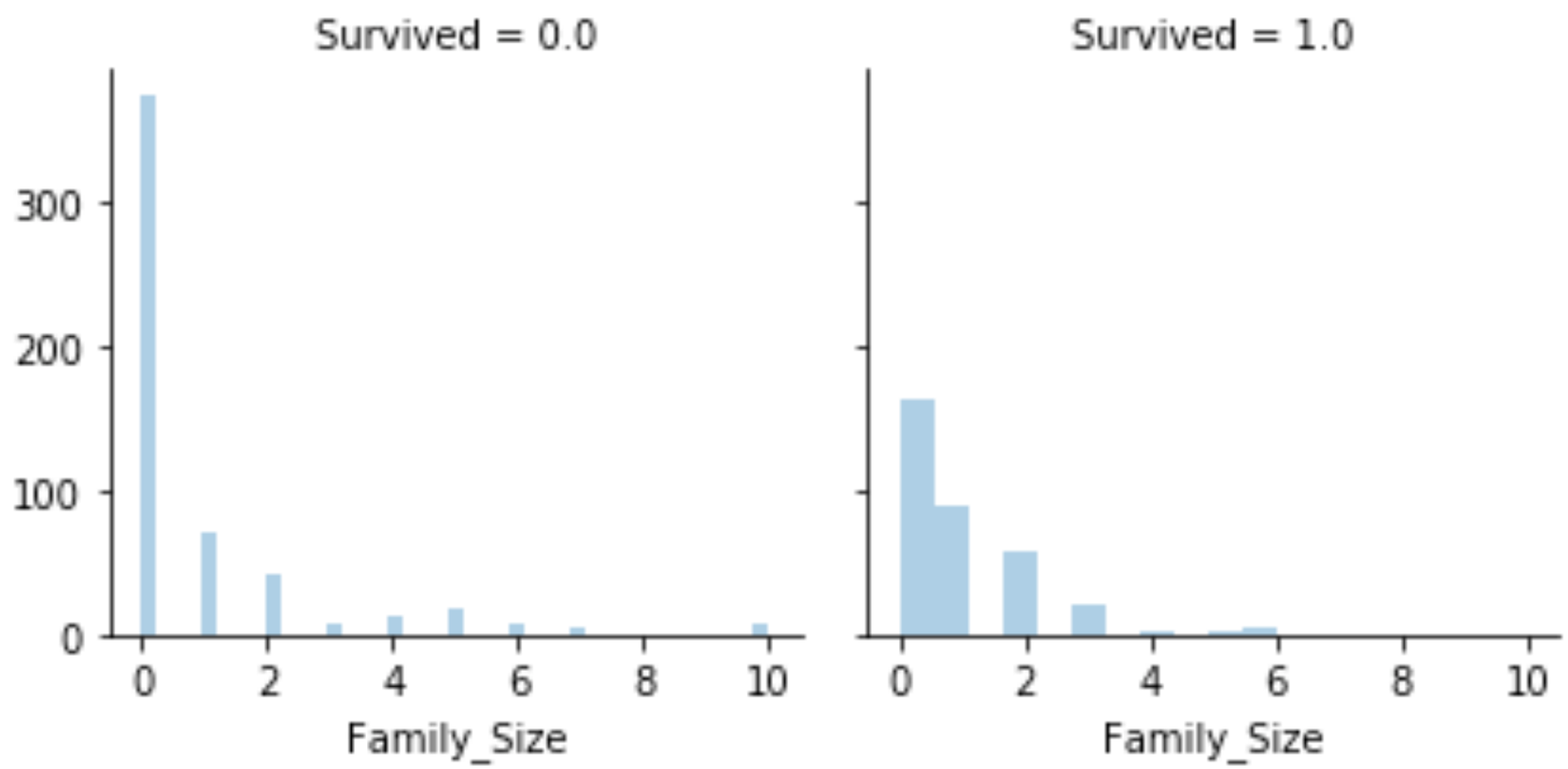


Kaggle

定義新的 feature: Family_Size = Parch + SibSp

- 資料轉換

```
data['Family_Size'] = data['Parch']+data['SibSp']
g = sns.FacetGrid(data,col='Survived')
g.map(sns.distplot, 'Family_Size', kde=False)
```



Kaggle

- 特徵工程

Name
Braund, Mr. Owen Harris
Cumings, Mrs. John Bradley (Florence Briggs Th...
Heikkinen, Miss. Laina
Futrelle, Mrs. Jacques Heath (Lily May Peel)
Allen, Mr. William Henry
Moran, Mr. James

「姓名」不能直接拿來預測，但其中的「稱謂」可能會跟是否生存有關

Kaggle

- 特徵工程

```
data['Title1'] = data['Name'].str.split(", ", expand=True)[1]
data['Title1'].head(3)
```

```
0                                Mr. Owen Harris
1    Mrs. John Bradley (Florence Briggs Thayer)
2                                 Miss. Laina
Name: Title1, dtype: object
```

```
data['Title1'] = data['Title1'].str.split(".", expand=True)[0]
data['Title1'].head(3)
```

```
0      Mr
1     Mrs
2    Miss
Name: Title1, dtype: object
```

```
data['Title1'].unique()
```

```
array(['Mr', 'Mrs', 'Miss', 'Master', 'Don', 'Rev', 'Dr', 'Mme', 'Ms',
       'Major', 'Lady', 'Sir', 'Mlle', 'Col', 'Capt', 'the Countess',
       'Jonkheer', 'Dona'], dtype=object)
```

Kaggle

- 特徵工程
  - 將稱謂與其他特徵作分析

```
pd.crosstab(data['Title1'],
            data['Sex']).T.style.background_gradient(cmap='summer_r')
```

Title1	Capt	Col	Don	Dona	Dr	Jonkheer	Lady	Major	Master	Miss	Mlle	Mme	Mr	Mrs	Ms	Rev	Sir	the Countess
Sex																		
female	0	0	0	1	1	0	1	0	0	260	2	1	0	197	2	0	0	1
male	1	4	1	0	7	1	0	2	61	0	0	0	757	0	0	8	1	0

  - 某些稱謂的乘客很少，所以合併其中的某些稱謂

```
data['Title2'] = data['Title1'].replace(['Mlle','Mme','Ms','Dr','Major','Lady',
    'the Countess','Jonkheer','Col','Rev','Capt','Sir','Don','Dona'],
         ['Miss','Mrs','Miss','Mr','Mr','Mrs','Mrs','Mr','Mr','Mr','Mr','Mr',
          'Mr','Mrs'])
data['Title2'].unique()
```

```
array(['Mr', 'Mrs', 'Miss', 'Master'], dtype=object)
```

Kaggle

- 特徵工程
  - 將票號資訊取出英文代碼(房間位置)的部分，省略後面的號碼，如果只有號碼的票號用 X 表示

```
data['Ticket_info'] = data['Ticket'].apply(lambda x :
    x.replace(".","").replace("/","").strip().split(' ')[0] if
    not x.isdigit() else 'X')
data['Ticket_info'].unique()
sns.countplot(data['Ticket_info'], hue=data['Survived'])
```

Kaggle

- 處理遺失值
  - 登船港口（Embarked）只遺漏少數，補次數最多的“ S ”
  - 費用（Fare) 也只少一筆，直接補上平均值
  - 觀察艙等（Cabin) 的資料後，只取出最前面的英文字母，剩下的用NoCabin來表示

```
data['Embarked'] = data['Embarked'].fillna('S')
data['Fare'] = data['Fare'].fillna(data['Fare'].mean())
data['Cabin'].head(10)
data["Cabin"] = data['Cabin'].apply(lambda x : str(x)[0] if not pd.isnull(x)
else 'NoCabin')
data["Cabin"].unique()
sns.countplot(data['Cabin'], hue=data['Survived'])
data.info()
```

Kaggle

- 將類別資料轉成整數

```
data['Sex'] = data['Sex'].astype('category').cat.codes
data['Embarked'] = data['Embarked'].astype('category').cat.codes
data['Pclass'] = data['Pclass'].astype('category').cat.codes
data['Title1'] = data['Title1'].astype('category').cat.codes
data['Title2'] = data['Title2'].astype('category').cat.codes
data['Cabin'] = data['Cabin'].astype('category').cat.codes
data['Ticket_info'] = data['Ticket_info'].astype('category').cat.codes
```

Kaggle

- 利用隨機森林來推測年齡

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import RandomForestRegressor
dataAgeNull = data[data["Age"].isnull()]
dataAgeNotNull = data[data["Age"].notnull()]
remove_outlier = dataAgeNotNull[(np.abs(dataAgeNotNull["Fare"]-
        dataAgeNotNull["Fare"].mean())>(4*dataAgeNotNull["Fare"].std()))|
                                (np.abs(dataAgeNotNull["Family_Size"]-
 dataAgeNotNull["Family_Size"].mean())>(4*dataAgeNotNull["Family_Size"].std()))
                        ]
rfModel_age = RandomForestRegressor(n_estimators=2000,random_state=42)
ageColumns = ['Embarked', 'Fare', 'Pclass', 'Sex', 'Family_Size', 'Title1',
              'Title2','Cabin','Ticket_info']
rfModel_age.fit(remove_outlier[ageColumns], remove_outlier["Age"])

ageNullValues = rfModel_age.predict(X= dataAgeNull[ageColumns])
dataAgeNull.loc[:,"Age"] = ageNullValues
data = dataAgeNull.append(dataAgeNotNull)
data.reset_index(inplace=True, drop=True)
```

Kaggle

- 留下需要的特徵

```
dataTrain = data[pd.notnull(data['Survived'])].sort_values(by=["PassengerId"])
dataTest = data[~pd.notnull(data['Survived'])].sort_values(by=["PassengerId"])

dataTrain.columns

dataTrain = dataTrain[['Survived', 'Age', 'Embarked', 'Fare',  'Pclass', 'Sex',
                       'Family_Size', 'Title2','Ticket_info','Cabin']]
dataTest = dataTest[['Age', 'Embarked', 'Fare', 'Pclass', 'Sex', 'Family_Size',
                     'Title2','Ticket_info','Cabin']]
dataTrain
```

Kaggle

- 利用隨機森林來預測存活率

```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(criterion='gini',
                            n_estimators=1000,
                            min_samples_split=12,
                            min_samples_leaf=1,
                            oob_score=True,
                            random_state=1,
                            n_jobs=-1)

rf.fit(dataTrain.iloc[:, 1:], dataTrain.iloc[:, 0])
print("%.4f" % rf.oob_score_)
```

Kaggle

- 產生 Submit 檔

```
pd.concat((pd.DataFrame(dataTrain.iloc[:, 1:].columns, columns = ['variable']),
           pd.DataFrame(rf.feature_importances_, columns = ['importance'])),
          axis = 1).sort_values(by='importance', ascending = False)[:20]

rf_res = rf.predict(dataTest)
submit['Survived'] = rf_res
submit['Survived'] = submit['Survived'].astype(int)
submit.to_csv('submit.csv',index=False)
```

Overview Data Notebooks Discussion Leaderboard Rules Team My Submissions **Submit Predictions**

You have 9 submissions remaining today. This resets 14 hours from now (00: 00 UTC).

Step 1
Upload submission file

記得要上傳

File Format
Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

Number of Predictions
We expect the solution file to have 418 prediction rows. This file should have a header row. Please see sample submission file on the data page.

Step 2
Describe submission

Briefly describe your submission

Ref: https://medium.com/@yehjames/資料分析-機器學習-第4-1講-kaggle競賽-鐵達尼號生存預測-前16-排名-a8842fea7077