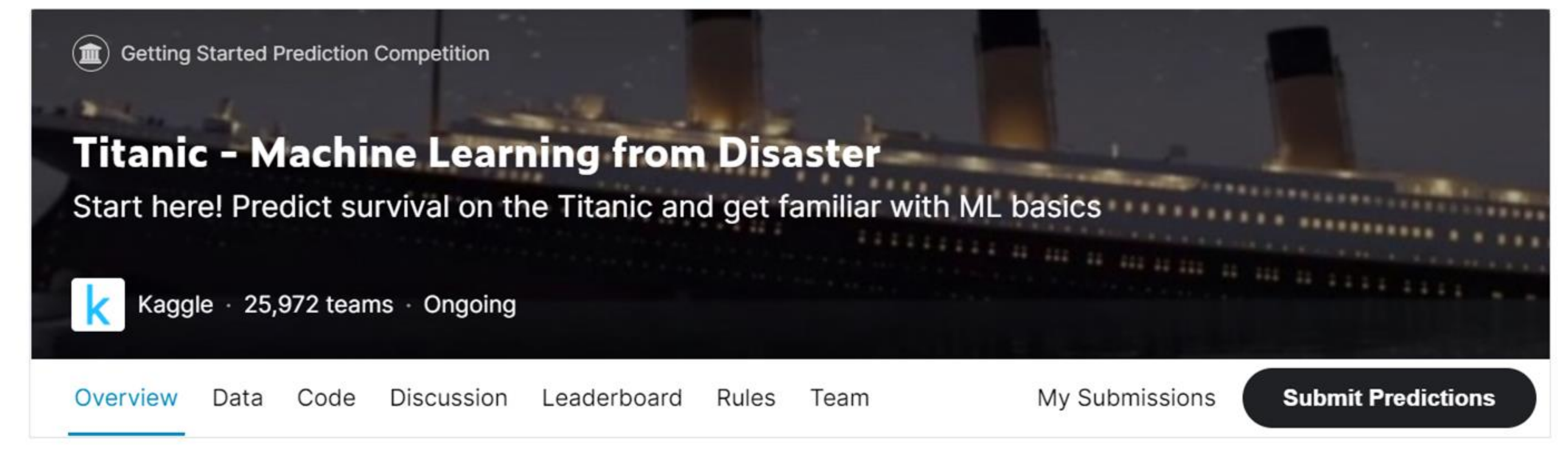
國立臺北大學
National Taipei University

Group 1
Case Study-Titanic


Course: 資料探勘 Data mining
Member: 鄧詠薇 711036115(組長)
林奕衜 710836102(副組長)

kaggle

Google



題目:

利用機器學習建立一個模型來預測在鐵達尼的乘客是否會存活

變數介紹

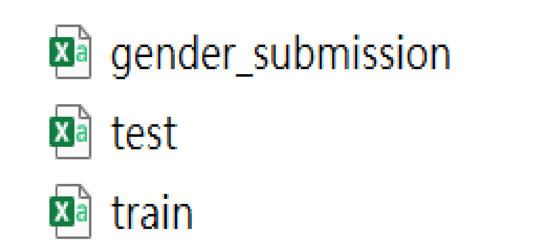
資料處理

模型使用

分析結果

變數介紹

- PassengerID
- Survived-survived(1),no survived(0)
- Pclass-upper(1),middle(2),lower(3)
- SibSp-兄弟姊妹與配偶數
- Parch-父母與小孩數
- Name
- Sex
- Age
- Ticket
- Fare
- Cabin-船艙
- Embarked-登船地點
(C = Cherbourg, Q = Queenstown, S = Southampton)

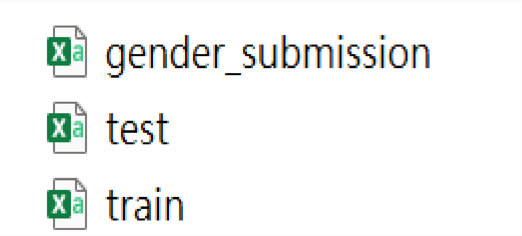


變數介紹

資料處理

模型使用

分析結果



1. 了解資料集的組成
2. 算出變數空值數
3. 合併 train&test set
4. 補空值(age,embarked)
5. Encoding 及Standardization
6. 資料拆分
7. 查看相關係數(correlation coefficient)

資料處理

step 1- 了解資料集的組成

train set

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

test set

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Pclass       418 non-null    int64
 2   Name         418 non-null    object
 3   Sex          418 non-null    object
 4   Age          332 non-null    float64
 5   SibSp        418 non-null    int64
 6   Parch        418 non-null    int64
 7   Ticket       418 non-null    object
 8   Fare         417 non-null    float64
 9   Cabin        91 non-null     object
 10  Embarked     418 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.0+ KB
```

step 2- 算出變數空值數

```
PassengerId : 0
Survived : 0
Pclass : 0
Name : 0
Sex : 0
Age : 177
SibSp : 0
Parch : 0
Ticket : 0
Fare : 0
Cabin : 687
Embarked : 2
```

step 3- 合併 train&test set

```
# 將train 、test set合併，一同進行Encoding及standardization
df_all = pd.concat([df_train_nosurvived,test],axis=0)
df_all = df_all.reset_index(drop=True)
df_all
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  1309 non-null   int64
 1   Pclass       1309 non-null   int64
 2   Name         1309 non-null   object
 3   Sex          1309 non-null   object
 4   Age          1046 non-null   float64
 5   SibSp        1309 non-null   int64
 6   Parch        1309 non-null   int64
 7   Ticket       1309 non-null   object
 8   Fare         1309 non-null   float64
 9   Cabin        295 non-null    object
 10  Embarked     1307 non-null   object
dtypes: float64(2), int64(4), object(5)
memory usage: 112.6+ KB
```

step 4- 補空值(Age,Embarked)

- Embarked-經過google 人名後發現兩人皆由southampton上船，因此填上"S"。
- Age- 查看與各變數的相關係數值，發現與Pclass有最高的相關係數，因此利用Pclass區分男女，取中位數來填補空值。

step 4- 補空值(Age,Embarked)

	Feature 1	Feature 2	Correlation Coefficient
3	Age	Age	1.000000
9	Age	Pclass	0.408106
12	Age	SibSp	0.243699
17	Age	Fare	0.178328
20	Age	Parch	0.150917
30	Age	PassengerId	0.028814

```
Median age of Pclass 1 females: 36.0
Median age of Pclass 1 males: 42.0
Median age of Pclass 2 females: 28.0
Median age of Pclass 2 males: 29.5
Median age of Pclass 3 females: 22.0
Median age of Pclass 3 males: 25.0
Median age of all passengers: 28.0
```

step 5 -Encoding & Standardization

1. 使用SKlearn中的 One Hot Encoding 將特徵轉為數字

Color
Red
Red
Yellow
Green

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

2. 標準化各變數，再把train & test set 分開

	Survived
Pclass	-0.338481
Age	-0.058635
SibSp	-0.035322
Parch	0.081629
Fare	0.257307
Sex_female	0.543351
Sex_male	-0.543351
Embarked_C	0.168240
Embarked_Q	0.003650
Embarked_S	-0.155660
PassengerId	-0.005007
Survived	1.000000

step 6- 資料拆分

-把 train set 資料集依照 8:2分為 訓練集(train set)及開發集(dev set)

step 7- 查看相關係數 (correlation coefficient)

-查看survived與其他變數的相關係數

變數介紹

資料處理

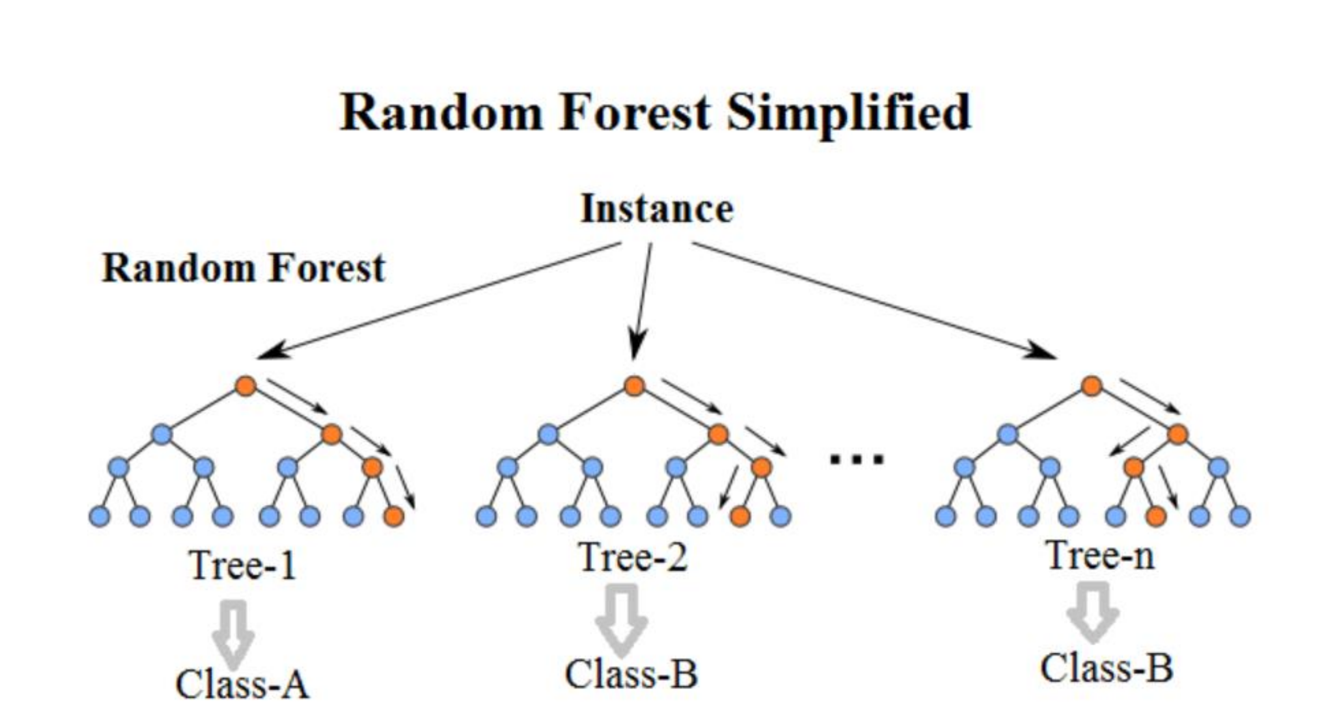
模型使用

分析結果

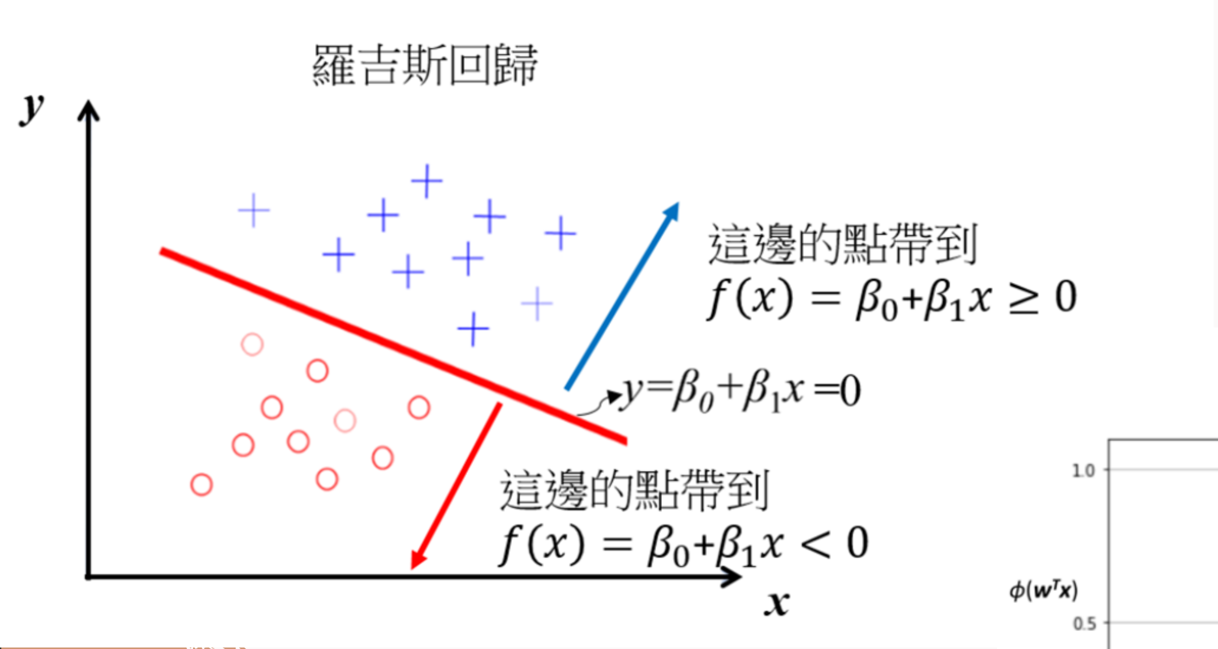
模型使用

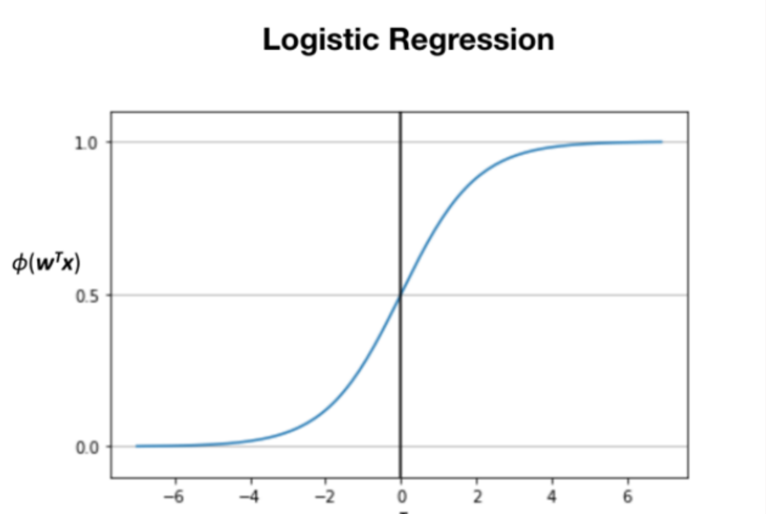
1. Random Forest
2. Logistic Regression
3. Support Vector Machine(SVM)

Random Forest

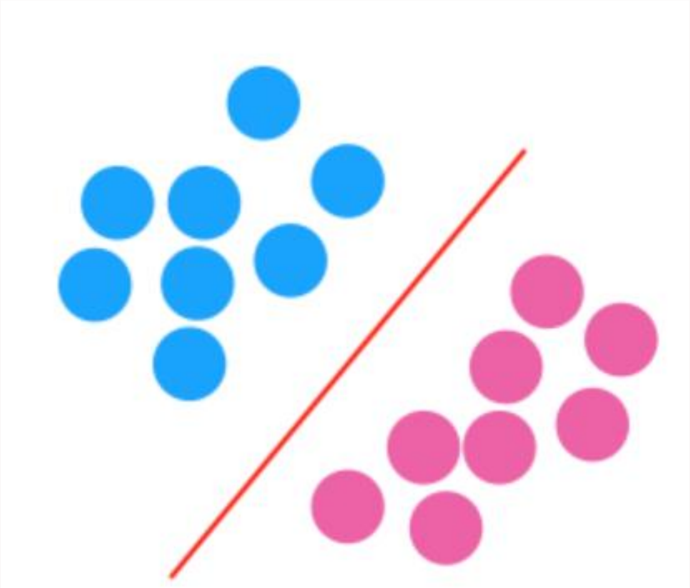


Logistic Regression

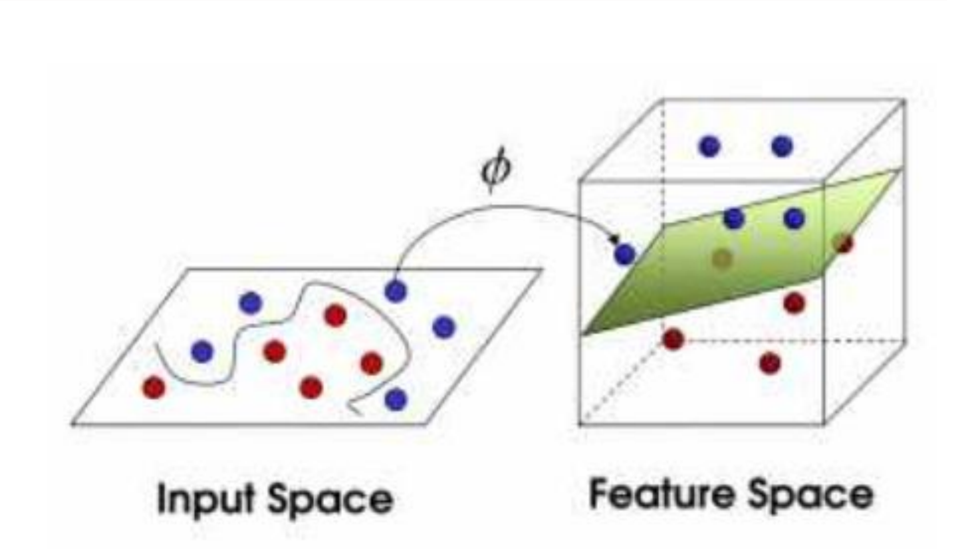


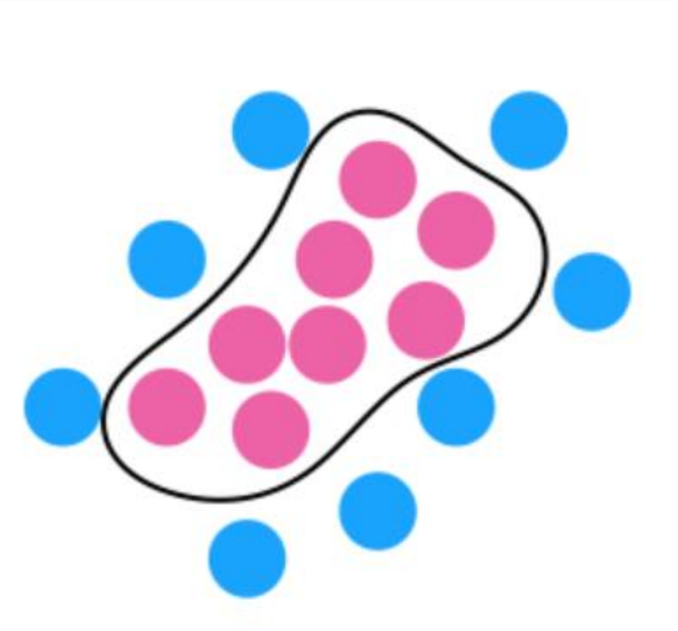


Support Vector Machine(SVM)



Support Vector Machine(SVM)





變數介紹

資料處理

模型使用

分析結果

分析結果

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

1. Random Forest→ 0.82

2. Logistic Regression→ 0.79

3. Support Vector Machine(SVM)→ 0.79

Dev Set

1. Random Forest→ 0.78

2. Logistic Regression→ 0.77

3. Support Vector Machine(SVM)→ 0.77

Test Set

8730/27740 (31%)

8563	NTPUIM-Day	0.77751	16	~10s

Your Best Entry ↑

Ranking

國立臺北大學
National Taipei University

Code：
https://colab.research.google.com/drive/1vJdNdwZoHOC8sLxAdOgnmIS4x6l3ALLY?usp=sharing

Data：
https://drive.google.com/drive/folders/1CEFNpF0Rdeai4e2UmwGraqmtNsblCvnX?usp=sharing


Course: 資料探勘 Data mining

Member: 鄧詠薇 711036115(組長)

林奕衛 710836102(副組長)

國立臺北大學
National Taipei University

心得


Course: 資料探勘 Data mining
Member: 鄧詠薇 711036115(組長)
林奕銜 710836102(副組長)

國立臺北大學
National Taipei University

QA

Course: 資料探勘 Data mining

Member: 鄧詠薇 711036115(組長)

林奕銜 710836102(副組長)

Thank You