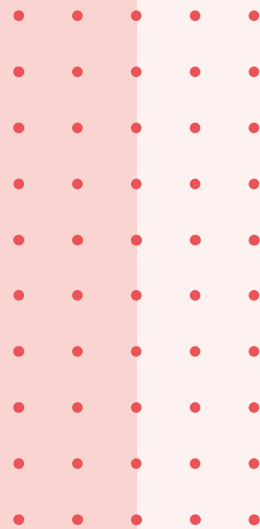


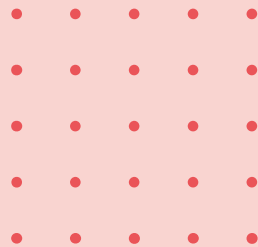
Data Mining Midterm Report

組長：B0928007 余明昌

組員：B0928015 艾思嘉、B0928024 莊靜修



目錄



01. 題目背景

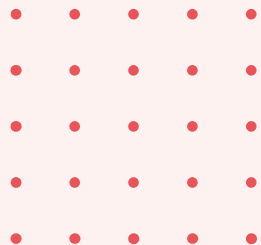
03. 資料集預處理

02. 資料集內容

04. 預計使用模型及方法

01. 題目背景

題目背景介紹及研究目的



Titanic - Machine Learning from Disaster

- 1912年鐵達尼號沉船事件，2224名乘客和船員中有1502人死亡。
- 乘客中有著各式各樣的人，從票種、性別、年齡等各種不同因素皆可能影響這些人的生存機會，因此某些特定群體比其他群體之生存率更高。

Titanic - Machine Learning from Disaster

➤ 研究目的：

透過這些資訊以機器學習的方式判斷人們在沉船後是否能夠生存。



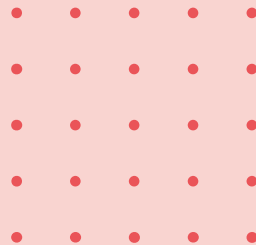
02. 資料集內容

介紹資料集內容

1. train.csv
2. test.csv
3. gender_submission.csv



Data Set



1. train.csv

- 模型訓練資料
- 訓練前將先做預處理再丟入模型訓練

2. test.csv

- 進行預測之資料
- 缺少Survived欄位，該欄位即為我們要去預測之結果

Data Set – train.csv

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
...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows x 12 columns

訓練資料集含有Survived欄位

Data Set – test.csv

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
...
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C
418 rows x 11 columns											


測試資料集可以明顯發現並沒有Survived欄位

Data Set

3. Gender_submission.csv

➤ 只包含兩欄位，繳交回Kaggle之資料集

1. PassengerID
2. Survived



	PassengerId	Survived
0	892	0
1	893	1
2	894	0
3	895	0
4	896	1
...
413	1305	0
414	1306	1
415	1307	0
416	1308	0
417	1309	0

418 rows x 2 columns

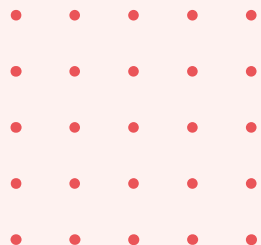
Kaggle

➤ 透過我們預測之資料，檢查乘客ID並對應是否存活，進而計算預測之準確率

03. 資料集預處理

介紹資料集預計處理方式

➤ train.csv



訓練資料集處理

1. 由電影可知，老弱婦孺先上逃生艇，因此年齡及性別為重要變數。
2. 將某些欄位從numeric的型態轉換為nominal型態。

ex.年齡大小按照範圍區間給予分類（老人、小孩、青少年等）

3. 性別為重要變數，因此將姓名中的 Mr. Ms. 稱謂分離出來當做特徵。
4. 適當刪除用不到之資料欄位。

訓練資料集處理－試做結果

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Lname	NamePrefix
0	1	0	3	male	Student	1	0	1_quartile	N	Braund,	Mr.
1	2	1	1	female	Adult	1	0	4_quartile	C	Cumings,	Mrs.
2	3	1	3	female	Young Adult	0	0	1_quartile	N	Heikkinen,	Miss.
3	4	1	1	female	Young Adult	1	0	4_quartile	C	Futrelle,	Mrs.
4	5	0	3	male	Young Adult	0	0	2_quartile	N	Allen,	Mr.

欄位從numeric的型態轉換為
nominal型態，給予分類。

將姓名中Mr Mrs
Miss 特別拉出

➤ 這邊有試著先刪掉一些可能無用之欄位，Ticket、Name、Embarked 等。

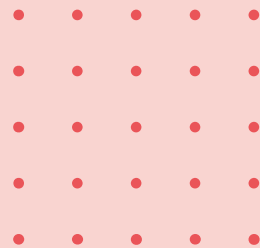
04.

預計使用模型及方法

介紹預計使用之機器學習模型及方法



預計使用之模型及方法



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.8354	0.8815	0.7219	0.8199	0.7670	0.6406	0.6445	0.1420
lr	Logistic Regression	0.8334	0.8854	0.7329	0.8065	0.7675	0.6383	0.6404	0.6000
lda	Linear Discriminant Analysis	0.8314	0.8750	0.7541	0.7860	0.7693	0.6367	0.6374	0.1160
ridge	Ridge Classifier	0.8274	0.0000	0.7380	0.7869	0.7611	0.6264	0.6276	0.0840
ada	Ada Boost Classifier	0.8194	0.8309	0.7542	0.7603	0.7565	0.6132	0.6138	0.1400
xgboost	Extreme Gradient Boosting	0.8154	0.8495	0.7432	0.7636	0.7528	0.6055	0.6062	0.1680
knn	K Neighbors Classifier	0.8133	0.8635	0.7112	0.7795	0.7423	0.5964	0.5994	0.1360
rf	Random Forest Classifier	0.8114	0.8545	0.7272	0.7620	0.7434	0.5944	0.5955	0.2280
et	Extra Trees Classifier	0.8113	0.8348	0.7111	0.7721	0.7395	0.5920	0.5941	0.2220
svm	SVM - Linear Kernel	0.8073	0.0000	0.7327	0.7529	0.7388	0.5867	0.5906	0.0840
dt	Decision Tree Classifier	0.7953	0.7735	0.7168	0.7378	0.7255	0.5625	0.5643	0.1060
dummy	Dummy Classifier	0.6245	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1100
qda	Quadratic Discriminant Analysis	0.4256	0.5033	0.7927	0.3947	0.5035	0.0087	-0.0003	0.1080
nb	Naive Bayes	0.4016	0.6860	0.9572	0.3819	0.5458	0.0197	0.0536	0.1220

- 利用Pycaret 這個套件，同時用多個演算法先做初步的模型比較
- 可以看到各個演算法模型的準確率差異
- 預計挑選2-3個準確率較優異的驗算法進行期末專題製作



Thanks