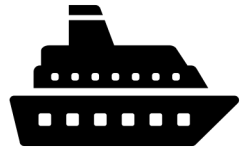
Kaggle

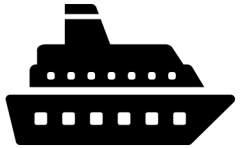
Titanic – Machine Learning From Disaster

code



內容

1. 資料前處理 (Missing Value)
2. 特徵工程 (Feature Engineering)
3. 資料視覺化與分析 (Data Visualization and Analysis)
4. 模型訓練 (Training Model)
5. 模型評估 (Model Evaluation)
6. 超參數調整 (Hyperparameter Tuning)



資料分析

- 這個專案共有 10 個 Features，分別為：Pclass、Name、Sex、Age、SibSp、Parch、Ticket、Fare、Cabin、Embarked，我們這些變數來預測乘客是否在船難中成功存活 (Survived)

	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

資料分析

- 檢查資料中是否存在遺漏值，稍後要進行補值
- training data 有遺漏值之變數：Age、Cabin、Embarked
- testing data 有遺漏值之變數：Age、Fare、Cabin

training data

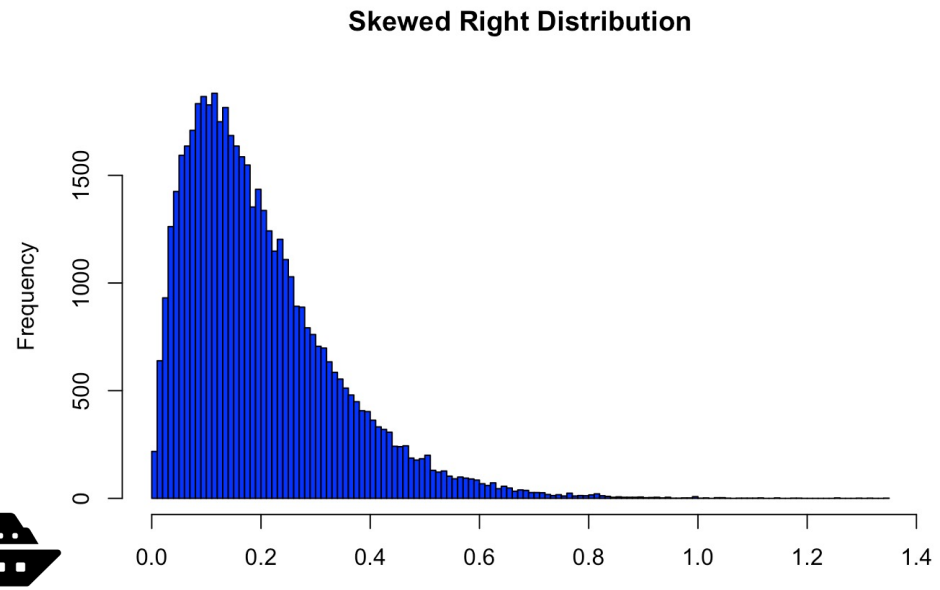
```
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    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
```

testing data

```
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
```

資料分析

- 觀察資料的分佈狀態，從中位數和平均數可以看出，好幾個變數是右偏分配，例如：Age，若分配偏態過於極端，稍後在做特徵工程時，要特別注意



training data

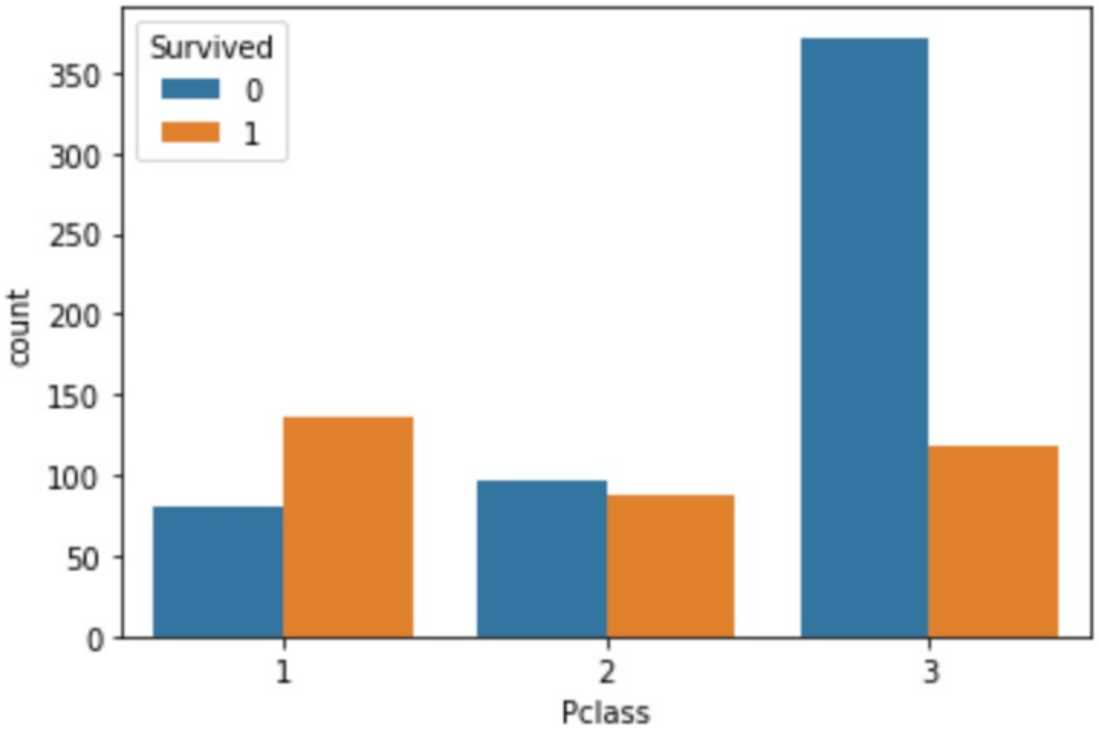
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

testing data

	PassengerId	Pclass	Age	SibSp	Parch	Fare
count	418.000000	418.000000	332.000000	418.000000	418.000000	417.000000
mean	1100.500000	2.265550	30.272590	0.447368	0.392344	35.627188
std	120.810458	0.841838	14.181209	0.896760	0.981429	55.907576
min	892.000000	1.000000	0.170000	0.000000	0.000000	0.000000
25%	996.250000	1.000000	21.000000	0.000000	0.000000	7.895800
50%	1100.500000	3.000000	27.000000	0.000000	0.000000	14.454200
75%	1204.750000	3.000000	39.000000	1.000000	0.000000	31.500000
max	1309.000000	3.000000	76.000000	8.000000	9.000000	512.329200

資料分析

- 怕一開始使用所有的 Features 去訓練模型會 overfitting，因此打算先訓練一個 basic model，此模型的 accuracy 也能當作一個基準值
- 選擇兩個直覺上會影響存活率的 features，分別為艙等、性別，兩者皆無遺漏值

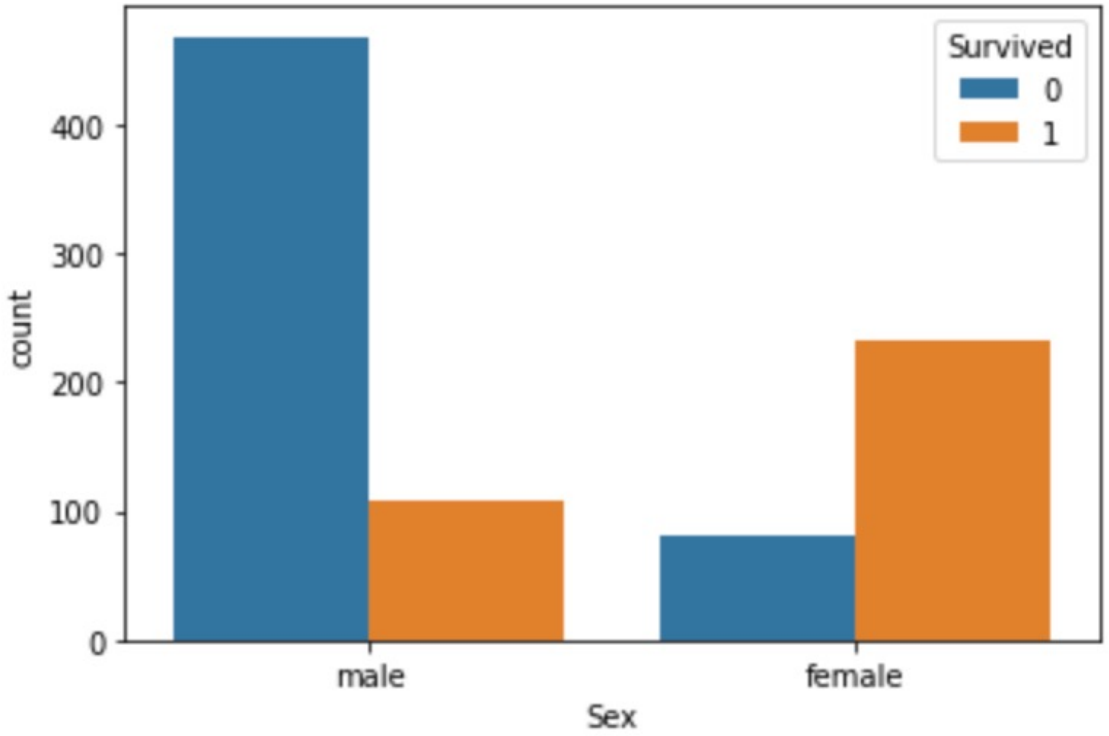


- 船艙分三個等級，高到低為 1 -> 3，越高級的船艙存活率越高，符合直覺

特徵工程

- 男性死亡數遠遠大於女性，直覺來說應該是讓女性先逃難

- 對性別做 One-Hot Encoding，把 male 轉換為 1、female 為 0



模型訓練

- 把 training data 按照 8 : 2 的比例切成 training set、validation set，使用 validation accuracy 挑選模型而不是用 training accuracy ，這樣才不會有 accuracy 因樣本外 testing data 而失真的問題

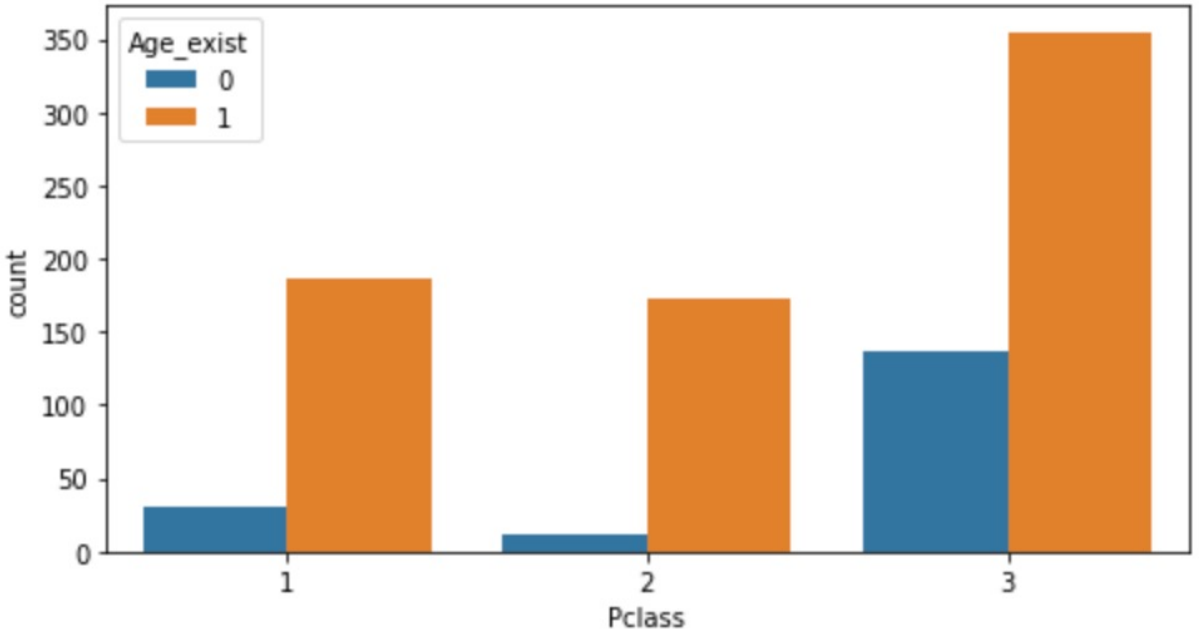
- 使用 Random Forest Model，此時的 parameter 不做過多調整，validation accuracy 為 0.788，Kaggle 上測試 score = 0.76555，若之後增加變數，accuracy 反而下降，就要注意是否有 overfitting 或是變數不具解釋能力的問題

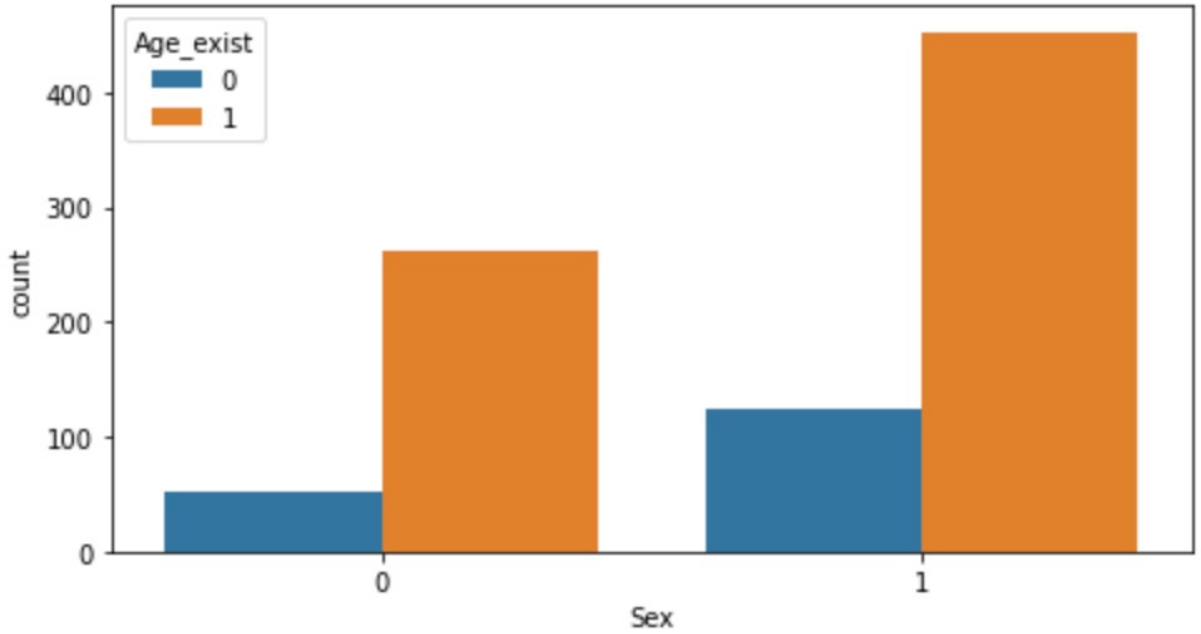
```
basic_forest = RandomForestClassifier(n_estimators=300,   # number of decision trees
                                      min_samples_split=20,
                                      n_jobs=2,   # number of cores for parallelism
                                      random_state=2)   # seed used by the random number generator

basic_forest.fit(X_train, y_train)

y_pred = basic_forest.predict(X_valid)
print('Accuracy (basic forest): %.3f' % accuracy_score(y_valid, y_pred))
```

```
Accuracy (basic forest): 0.788
```

特徵工程

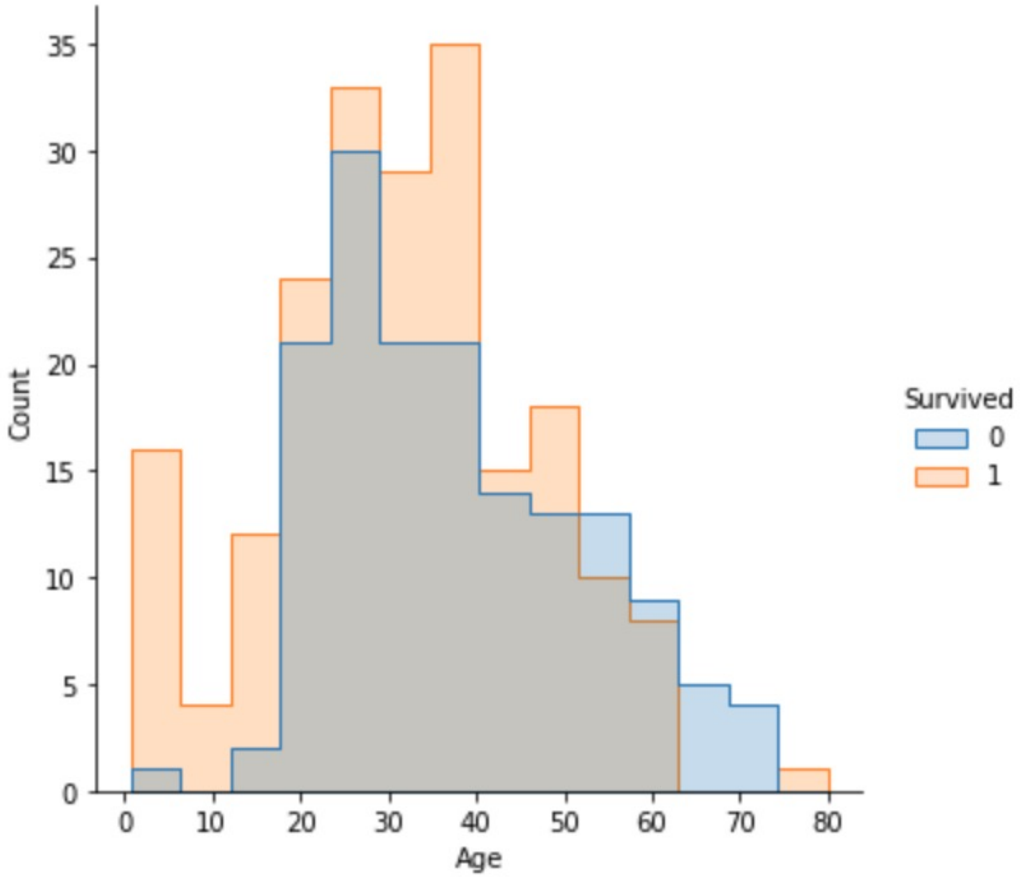
- 討論年齡變數，約有 15% 的遺漏值
- 年齡的遺漏值大部分來自 3 等船艙
- 男性遺漏值的佔比約比女性高 5%





特徵工程

- 因年齡的遺漏值大部分來自 3 等船艙，若要觀察年齡與存活之間的相關性，要先排除艙等 3，以免因遺漏值過多而失真



- 由右圖可看出， 17 歲以前，存活的人數明顯高於死亡人數，有可能是因為會先救小孩

- 我要以 17 歲當分界點，創造一個二元變數，會比直接使用真實年齡更適合，因若年紀大於 17 以後，探討真實年齡無太大意義

資料前處理

- 年齡的遺漏值約有 15%，若直接使用中位數補值，可能會影響預測結果
- 因年齡與 Name 中稱謂 (Title) 有一定程度的關聯，我決定借助稱謂進行補值
- 把 Title 做分類，可分為先生 (Mr)，罕見稱謂 (Rare)，小男孩 (Master)，小姐 (Miss)， 女士 (Mrs)
- 若為遺漏值，則按照每個 Title 年齡中位數進行補值，才不會有小男孩被補值成先生的問題發生

特徵工程

- 補值完後創到一個二元變數：Age_17，大於 17 歲設為 1，反之設為 0

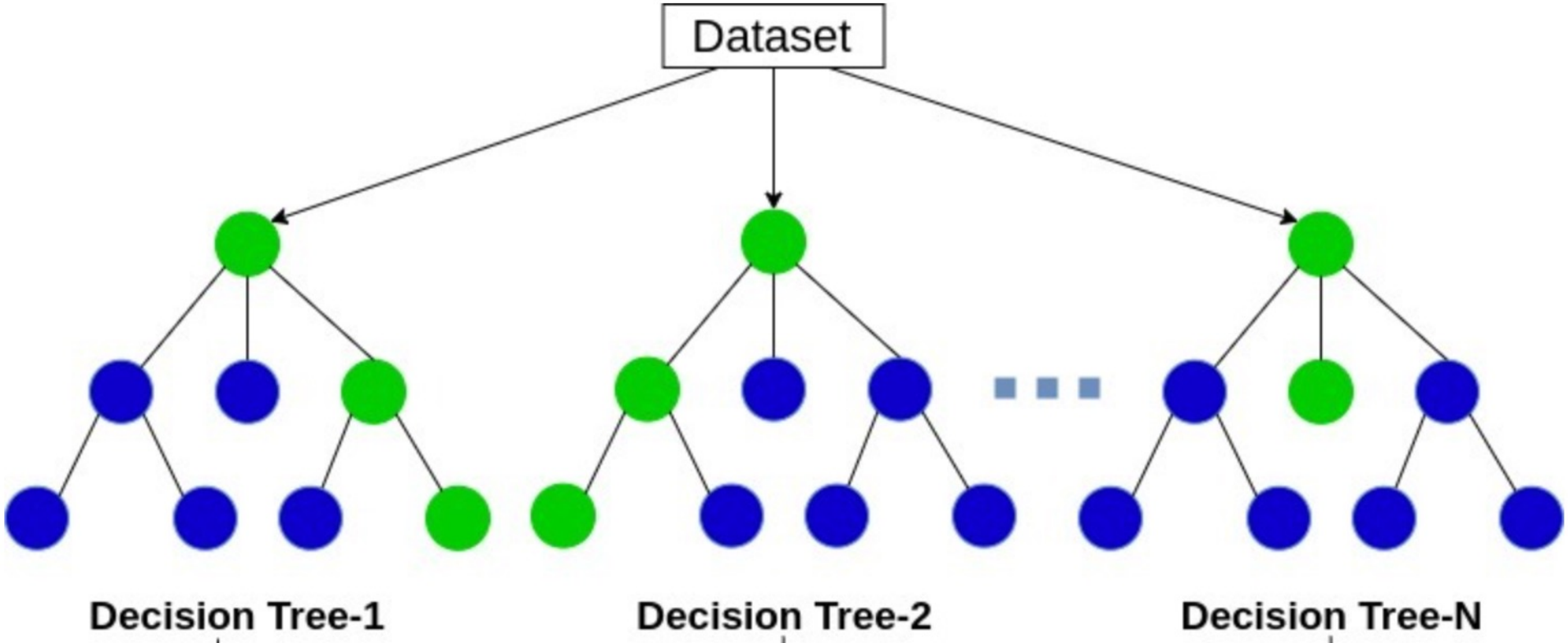
- 可以看到 Moran, Mr. James 的 Age 為 NaN，已被補值

Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title	all_Age	Age_17
3	Braund, Mr. Owen Harris	1	22.0	1	0	A/5 21171	7.2500	NaN	S	Mr	22.0	1
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	PC 17599	71.2833	C85	C	Mrs	38.0	1
3	Heikkinen, Miss. Laina	0	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	Miss	26.0	1
1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	113803	53.1000	C123	S	Mrs	35.0	1
3	Allen, Mr. William Henry	1	35.0	0	0	373450	8.0500	NaN	S	Mr	35.0	1
3	Moran, Mr. James	1	NaN	0	0	330877	8.4583	NaN	Q	Mr	30.0	1
1	McCarthy, Mr. Timothy J	1	54.0	0	0	17463	51.8625	E46	S	Mr	54.0	1
3	Palsson, Master. Gosta Leonard	1	2.0	3	1	349909	21.0750	NaN	S	Master	2.0	0
3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	0	27.0	0	2	347742	11.1333	NaN	S	Mrs	27.0	1
2	Nasser, Mrs. Nicholas (Adele Achem)	0	14.0	1	0	237736	30.0708	NaN	C	Mrs	14.0	0
3	Sandstrom, Miss. Marguerite Rut	0	4.0	1	1	PP 9549	16.7000	G6	S	Miss	4.0	0

模型訓練

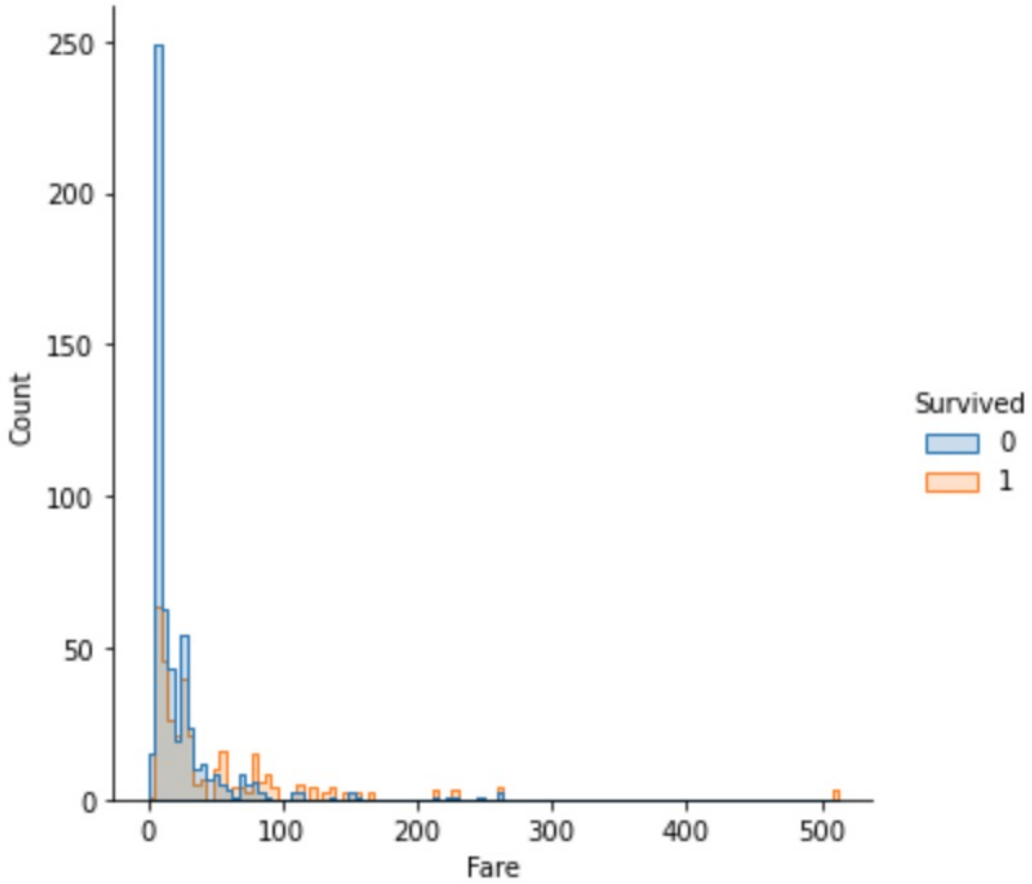
- 一樣把 training data 按照 8 : 2 的比例切成 training set、validation set

- 使用 Random Forest Model， parameter 不做過多調整，validation accuracy 為 0.793，Kaggle 上測試 score = 0.77 ，皆進步了 1 % 左右



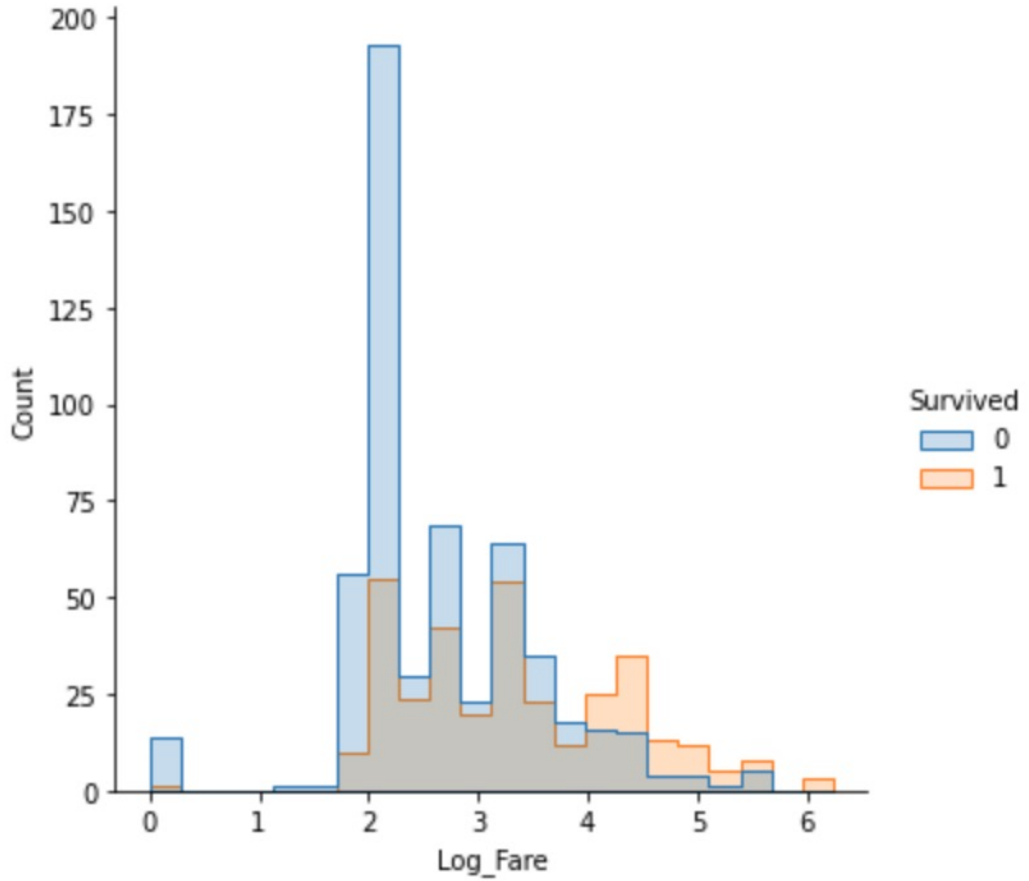
特徵工程

- 票價這個 feature 只有一個遺漏值，直接以中位數進行補值即可
- 下圖為票價的分配，非常廣斜票，呈現右偏分配，圖形不易看出重點



特徵工程

- 票價呈現右偏分配，圖形不易看出重點，因此我對其取 log，讓票價服從常態分配

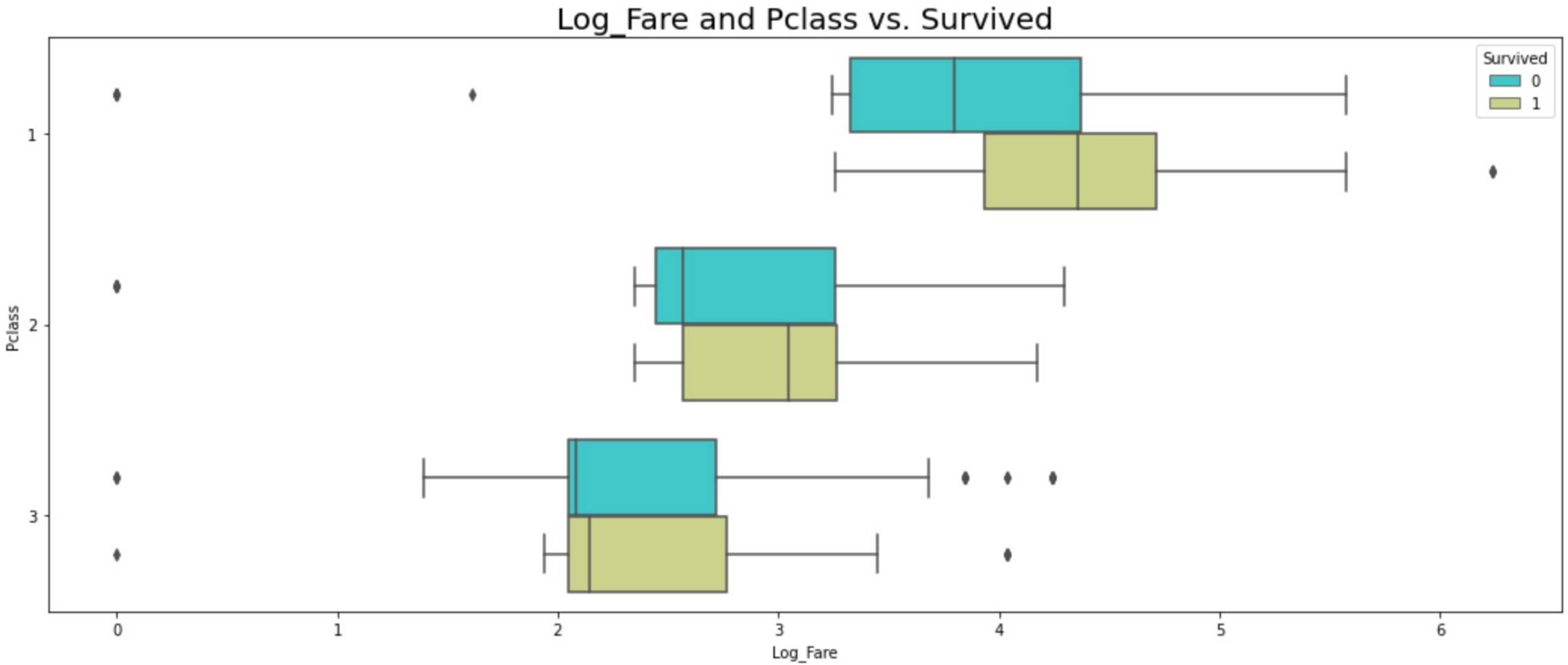


- 從 log(Fare) 的分配可以看出，較低的票價確實有較高的死亡率

- 用取 log 後的票價下去訓練模型，不會改變資料的性質和相關性，但會壓縮變數的尺度，使資料更加平穩

特徵工程

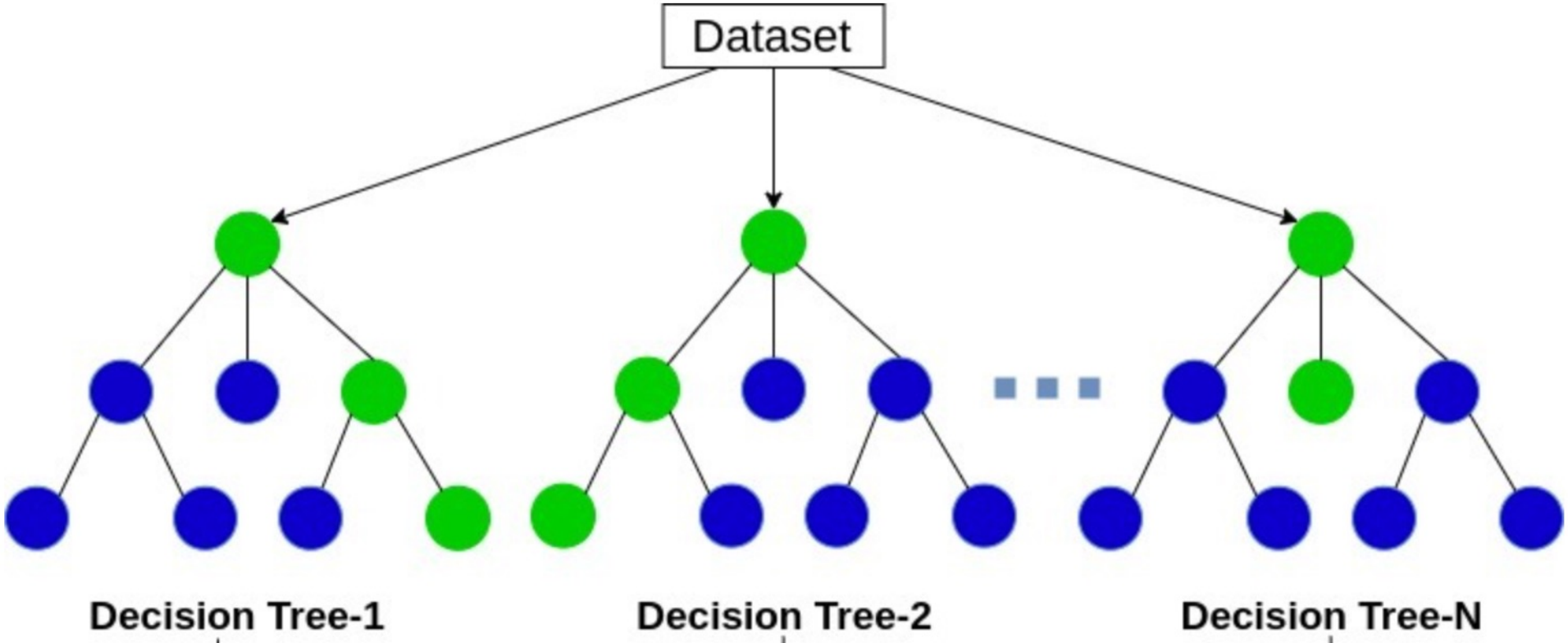
- 從盒鬚圖更可以明顯看出，存活下來的乘客確實平均而言付出較高的票價，故決定放入 log(票價) 這個變數



模型訓練

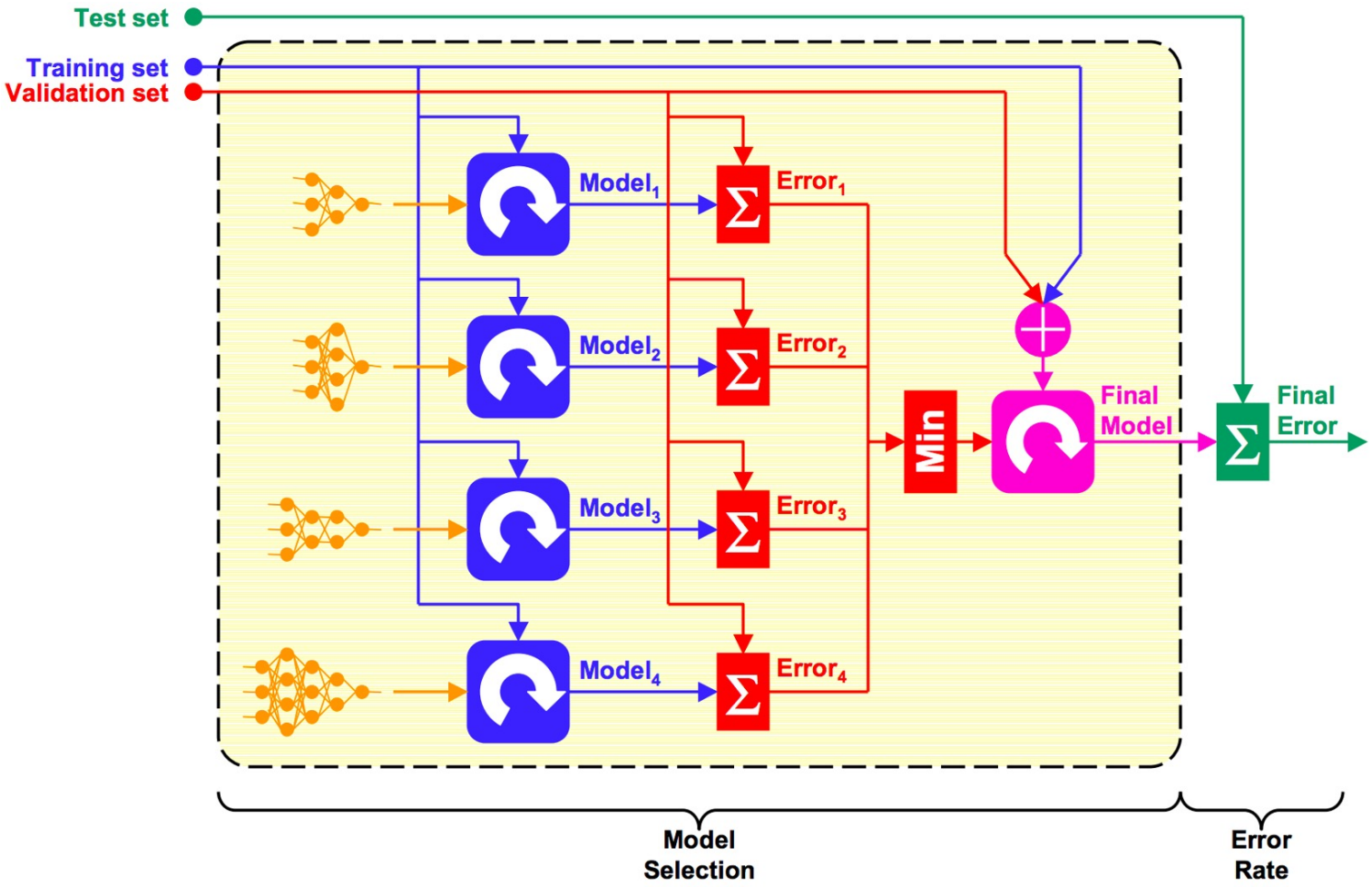
- 一樣把 training data 按照 8 : 2 的比例切成 training set、validation set

- 使用 Random Forest Model， parameter 不做過多調整，validation accuracy 為 0.832，進步了 4 %



模型訓練

- 此時訓練出來的模型之 validation accuracy 為目前最高的，我決定把 training data 和 validation data 合併，再訓練一次模型



- 在 Kaggle 上測試 score 為 0.78，又進步了 1 %

超參數調整

- 最後使用 Grid Search Function 對 Hyperparameters 進行調整

- 為了避免 overfitting，我只調整了先前使用的參數

```
param_min_samples_split = [14, 16, 18, 20, 22]
param_n_estimators = [300, 400, 500]

forest_grid = RandomForestClassifier(random_state=2, n_jobs=2)

param_grid = [{'min_samples_split':param_min_samples_split,
               'n_estimators':param_n_estimators}]

gs = GridSearchCV(estimator=forest_grid,
                  param_grid=param_grid,
                  scoring='accuracy')

gs = gs.fit(X, y)
print(gs.best_score_)
print(gs.best_params_)
```

```
0.8238277572029377
{'min_samples_split': 14, 'n_estimators': 400}
```

- 在 Kaggle 上測試 score 為 0.783，又進步了 0.3 %

模型評估

- 透過完整的步驟，最終使預測 accuracy 上升 5 %
- 最後使用 Random Forest 計算 Feature Importance，發現票價和性別在模型中的重要性非常高
- 往後延伸可以使用這兩個變數，搭配其他未使用到的 Features 進行預測