Group 5 Project 2

Titanic : Machine Learning From Disaster

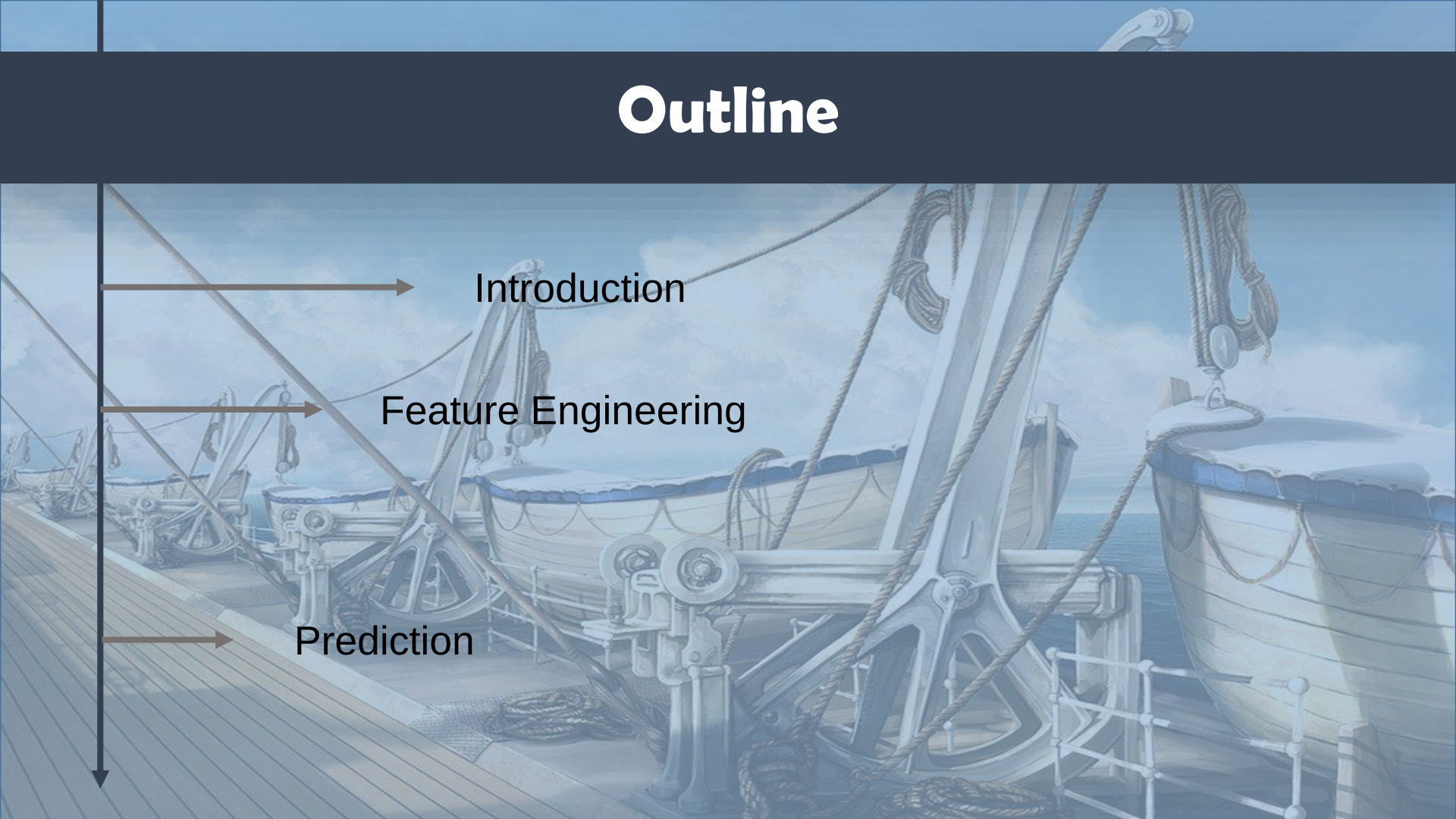
植科所 碩一 R06b42001 謝誌紘
新聞所 碩一 R06342017 羊敏丹
圖資系 大四 B03106014 黃彥鈞

2nd degree families and mojority voting

Erik Bruin

- Grouping variable
- “Second degree” family
- 5 predictors
- 3 Models

Outline



Introduction - background

1912 鐵達尼號事件…

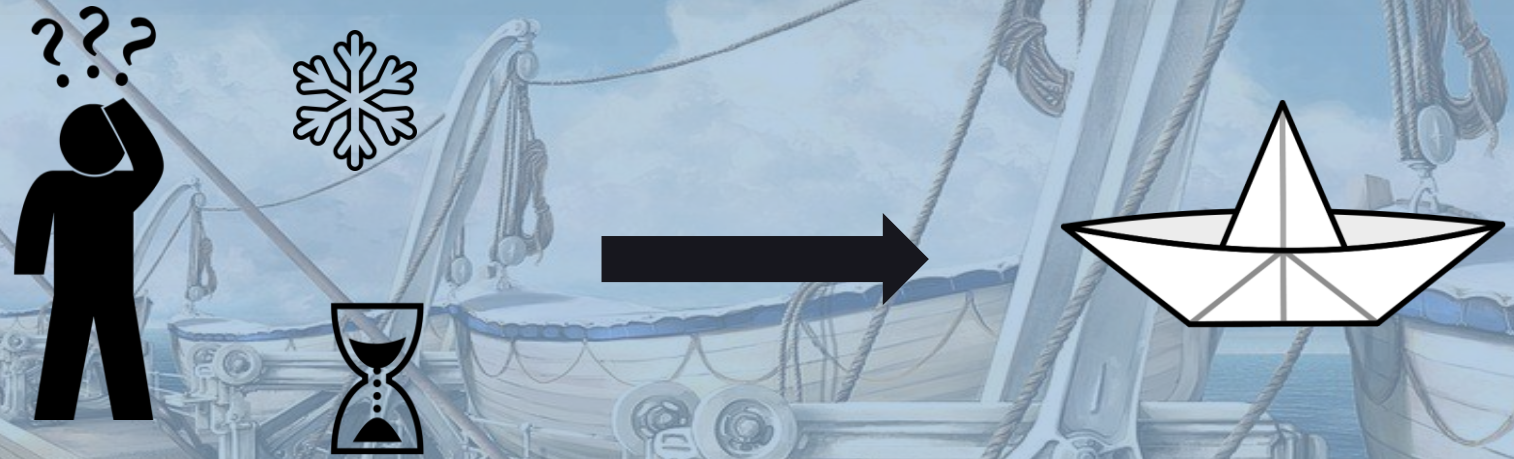
2224

1172

DEAD!!

1052

Introduction - background



任務 : 使用機器學習工具來預測哪些旅客能生還下來

Introduction - dataset

Training data

用於建模
資料結果
(參考標準)

Test data

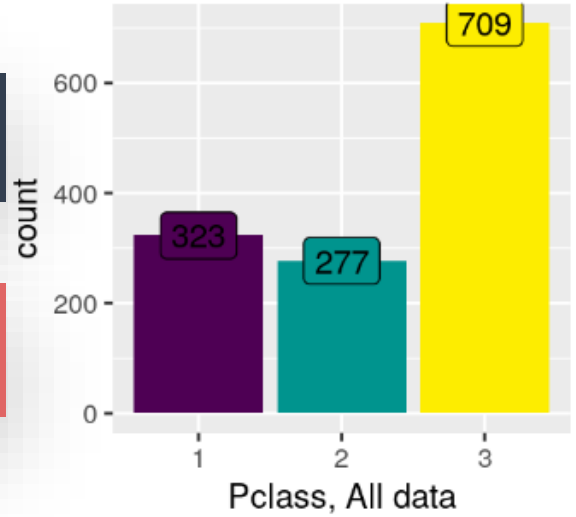
預測結果

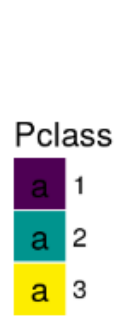
Introduction - dataset

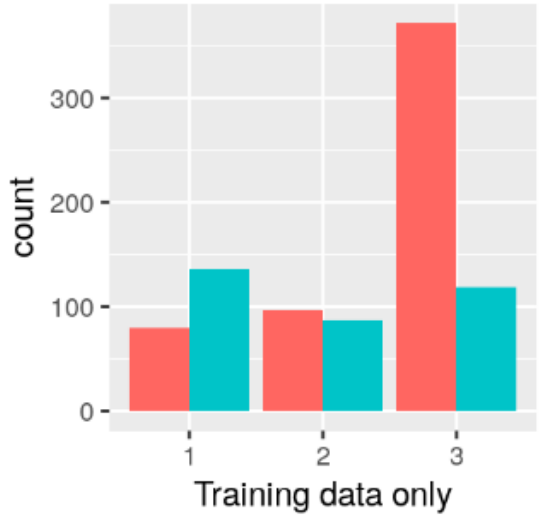
變數	變數定義	註	NA in Test Data
survival	生存與否	0 = 死亡 , 1 = 生存	418
pclass	船票等級	1 = 1st , 2 = 2nd , 3 = 3rd	0
sex	性別		0
age	年齡		263
sibsp	兄弟姊妹或配偶人數	未婚夫妻不算在內	0
parch	父母或小孩人數	保母不算在內	0
ticket	船票號碼		0
fare	票價(每張票)		1
cabin	船艙號碼		1014
embarked	登船港口	C = Cherbourg, Q = Queenstown, S = Southampton	2

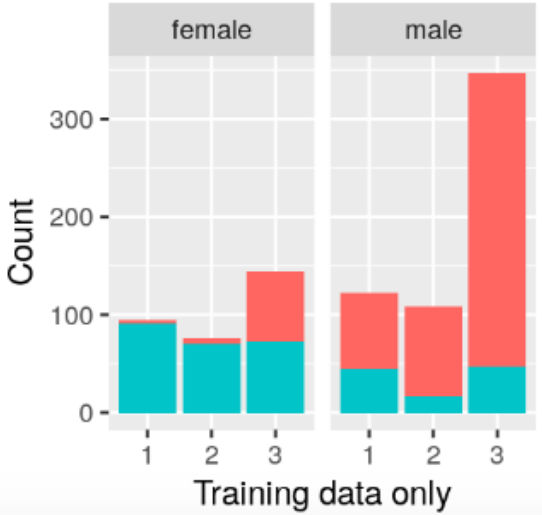
Variables

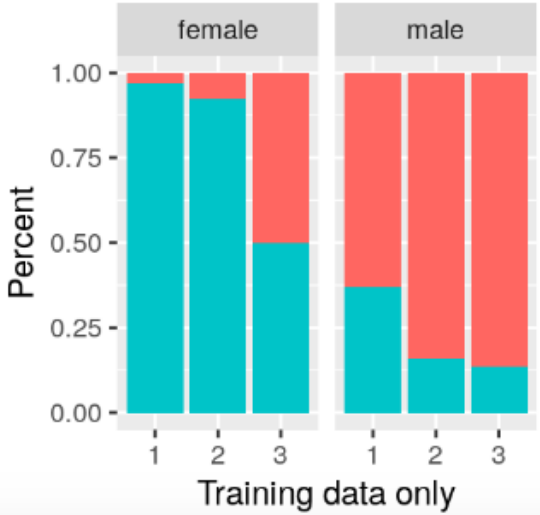
PclassSex











Variables

GroupSize

1. 建立變數 Fsize

- 將名稱欄位“Surname”和“Title”切割出來
- 算出旅客父母小孩、親戚配偶的數量
- 將家族數量和姓結合，變成Family Size。

Variables

GroupSize

2. 解決問題

- 將“Surname”和“Title”從旅客資訊裡的Title欄位切割出來。
- 算出旅客父母小孩、親戚配偶的數量
- 將家族數量和姓結合，變成Family Size。
- 旁系血親、姻親該如何算?

Variables

GroupSize

3. 分組
GroupSize

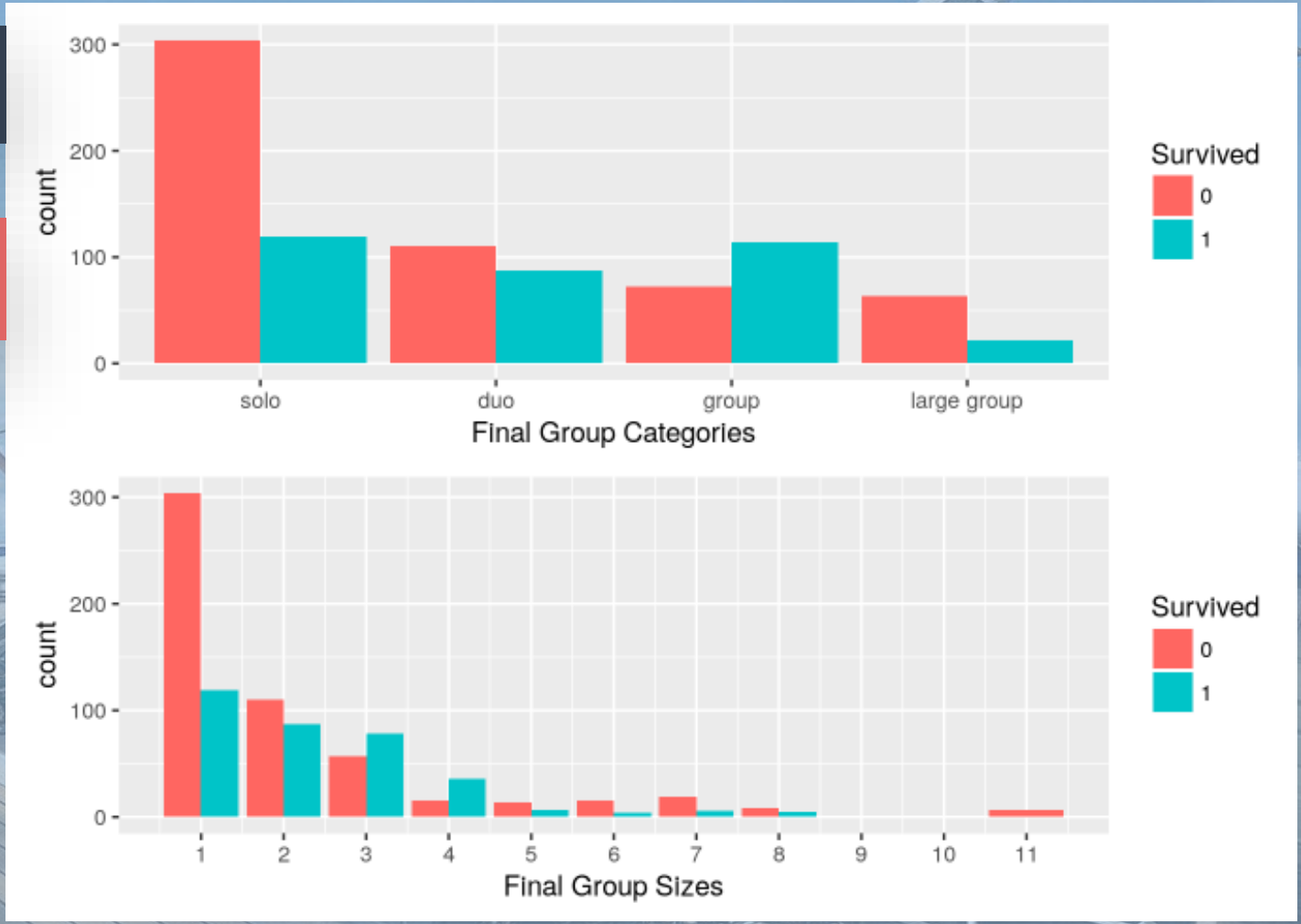
- Fsize = 1 ，則為 Solo
- Fsize = 2 , 則為Duo
- Fsize = 3 / 4 , 則為Group
- Fsize = 5人以上，則為Large Group

4. 加入其他可能

- 用船票去推估一起訂票的朋友。

Variables

GroupSize



Variables

FarePP

・回顧：

Fare欄位有1個NA值。

Embarked欄位有2個NA值

・解決：

依照登船的城市、

船艙得出Fare Per Person.

・問題：

免費登船？

統計結果傾斜？

Variables

IsChildP12

・回顧：

- 年齡欄位有263個NA值。

・解決：

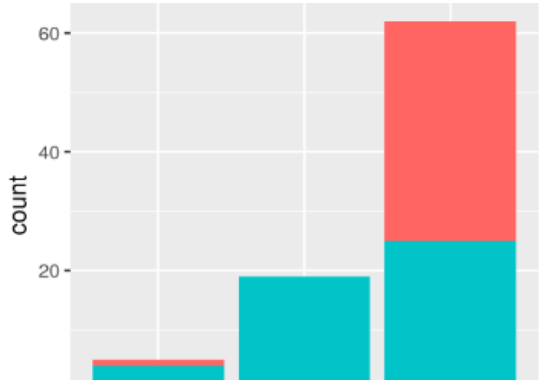
以Title和Pclass重要的預測值，採用Mice和Linear Regression，推估年齡。

・發現：

- 14.5歲以下的兒童存活率明顯的高。
- 從船艙資料分析，發現P3的結果與P1、P2差異大，且考量P3很多闕漏值。

・解決：

去掉P3的資料。



Variables--AnySurvivors

・發現：

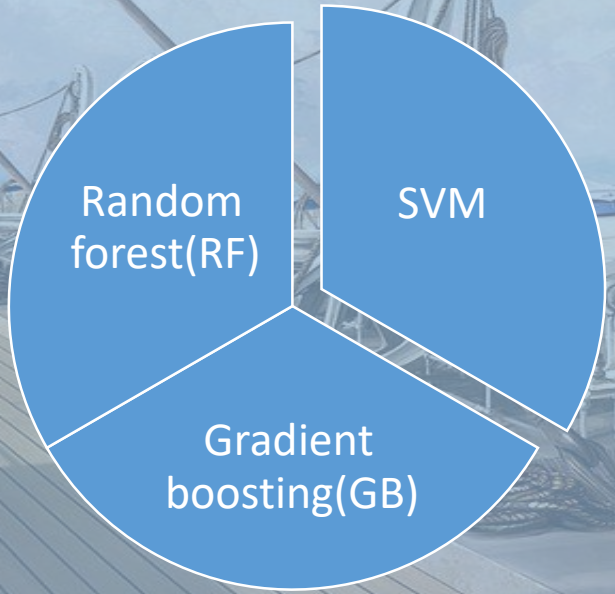
發現有同樣Ticket的人們，若有一個成員存活，其他成員存活機率也會變大。

・其他變數們：

- Cabin和Embarked的闕漏值太多，因此捨棄不用來做預測的變數。

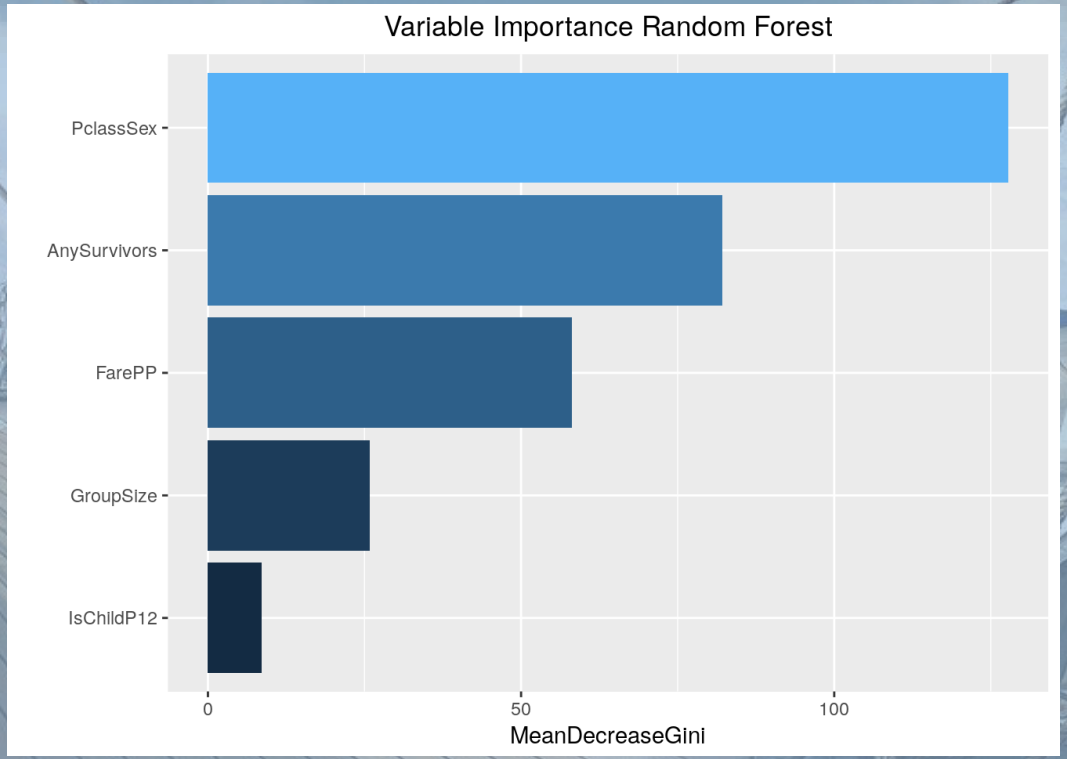
Prediction – Machine Learning

Models :



		Machine Learning	預測準確率
Training Data Set	Data Cleaning	Random Forest	85%
	Select Variable	SVM	82%
		GB	85%

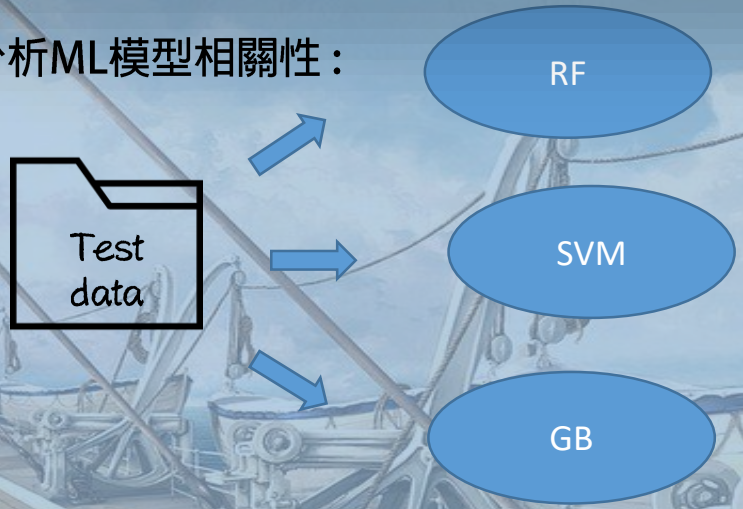
Prediction – RF Variable Importance

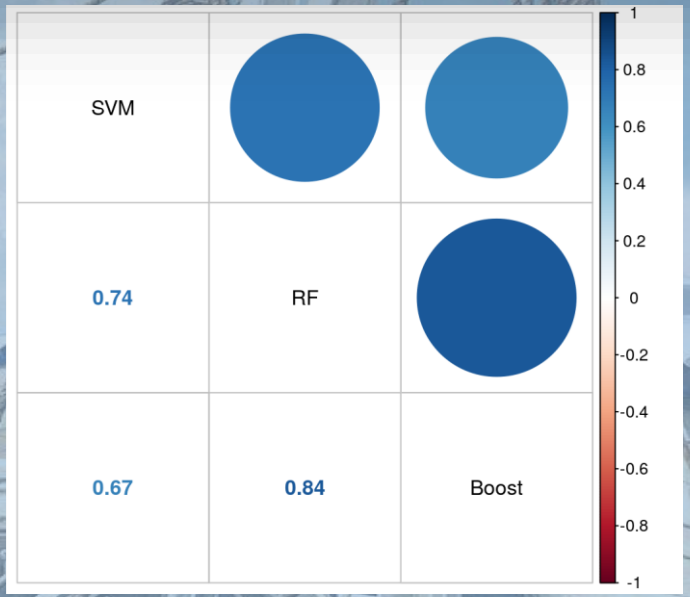


從RF上面可以看到PclassSex這個特徵是非常重要的!!!

Prediction – Machine Learning

分析ML模型相關性：



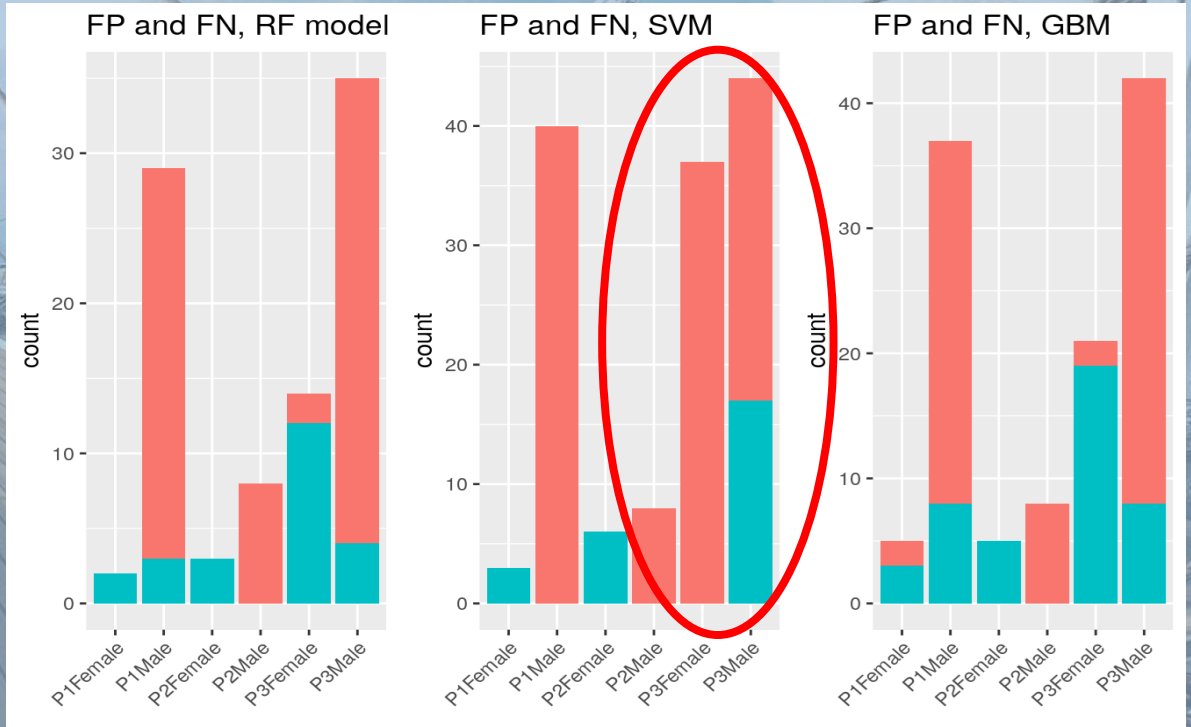


發現：SVM 與 RF 及 Boost

相關性最低(推論是方法的問題)

Prediction – Machine Learning

預測的成功與否在於比較少的False Positive and False Negative Rate



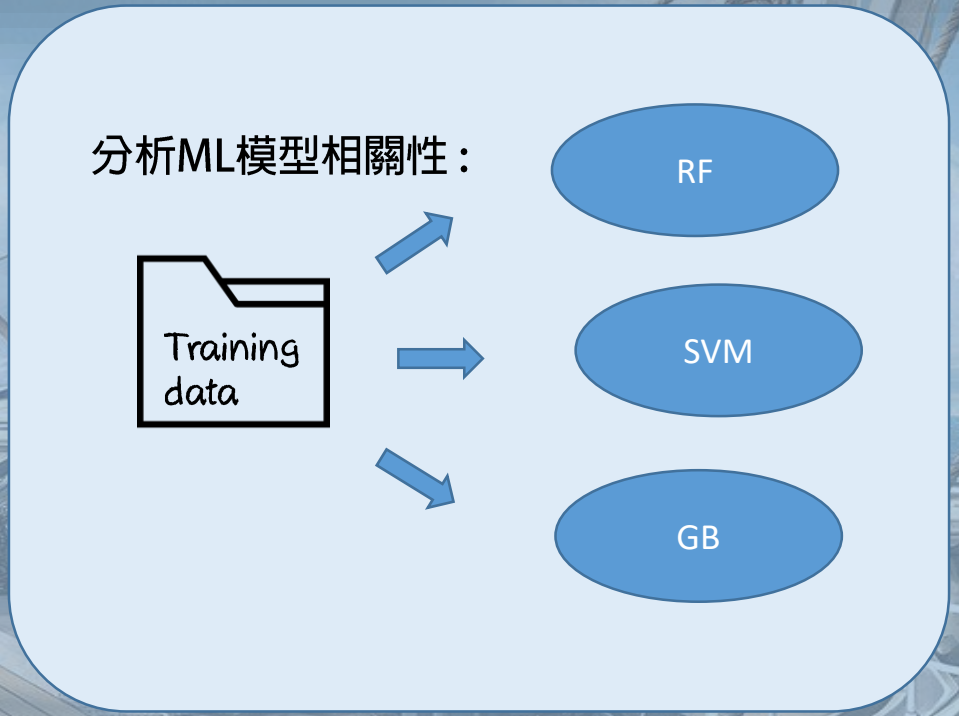
藍色 : False Positive
預測存活但死亡

紅色 : False Negative
預測死亡但存活

預測結果得知用SVM預測P3會有嚴重FP, FN

因此改用GB模型修正並去預測testing Data

Prediction – Machine Learning



Conclusion

只用了5個預測變數和3個ML model…

- 就獲得0.81818分的成績(預測testing data)(top 4% grade)

我們從他的Work上面學到了什麼!!?...

- 作者在資料清理方面真的是太厲害了!!
- 從名字上可以去觀察到親屬關係
- 還有對於資料的觀察十分細微
- 最重要的是: 他的report 上的步驟十分詳細且確實(好像教學文章)


Thanks for listening