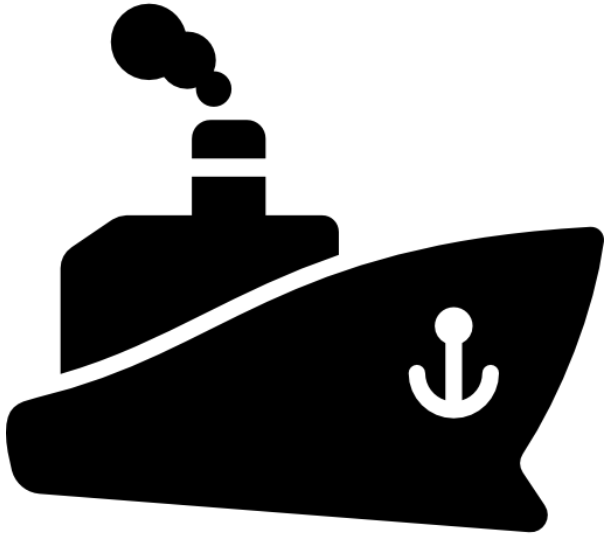
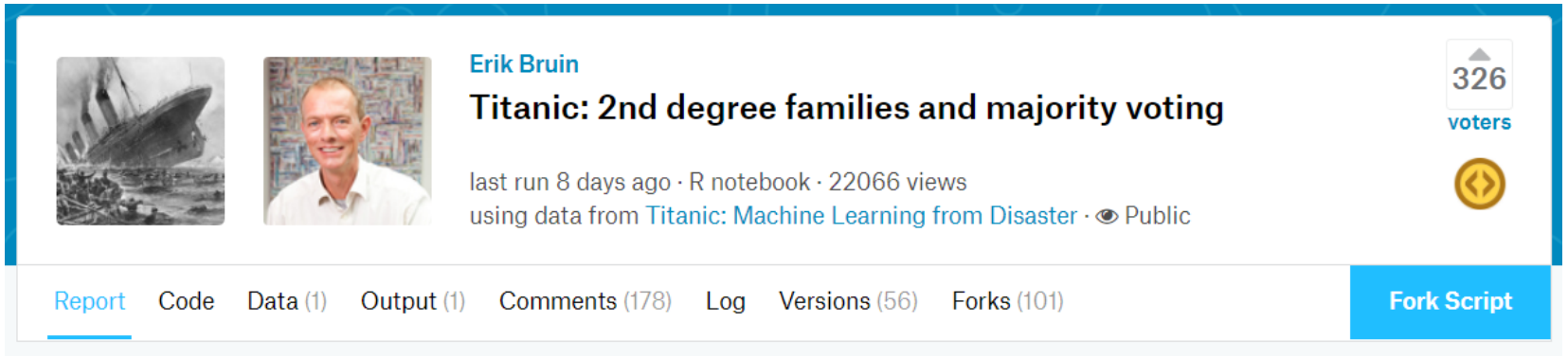
캐글 뽀개기



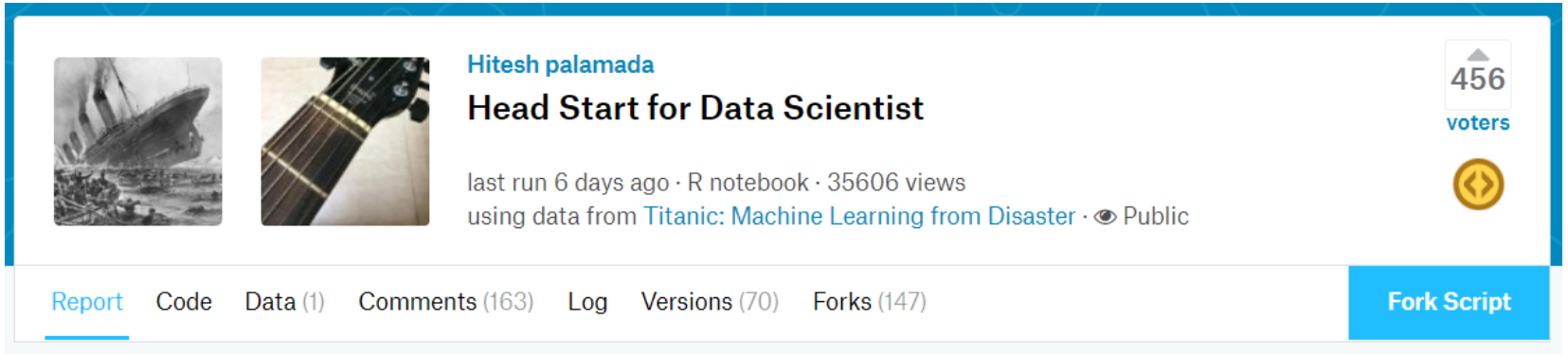
Titanic : Machine Learning from Disaster

김선빈　조명장　류성균

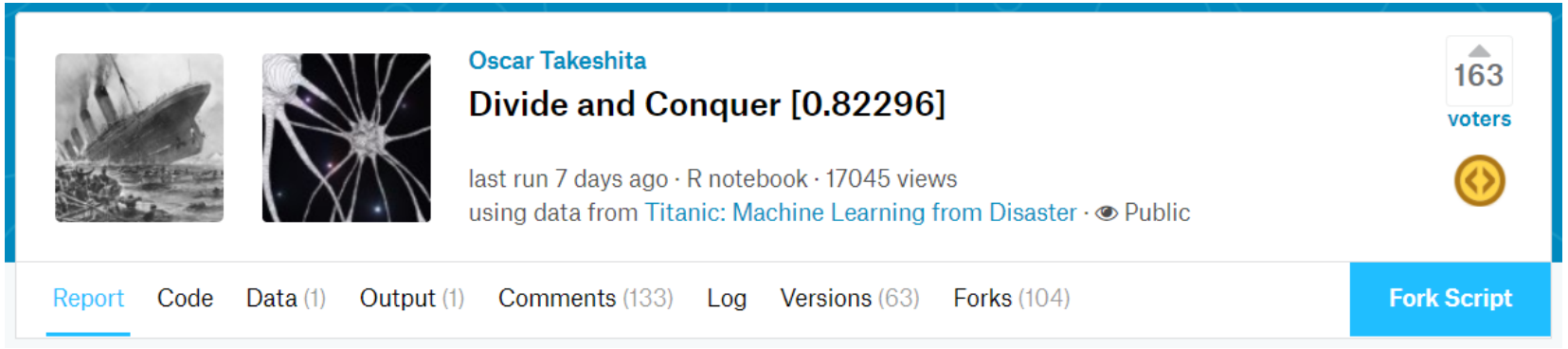
참고한 커널



숨겨진 그룹 찾기



기초 데이터 탐색



cross validation
& public score
결과가 비슷함 => 과적합 방지

Goal : 승객들의 생존여부 맞추기

생존 1

죽음 0

Data Dictionary

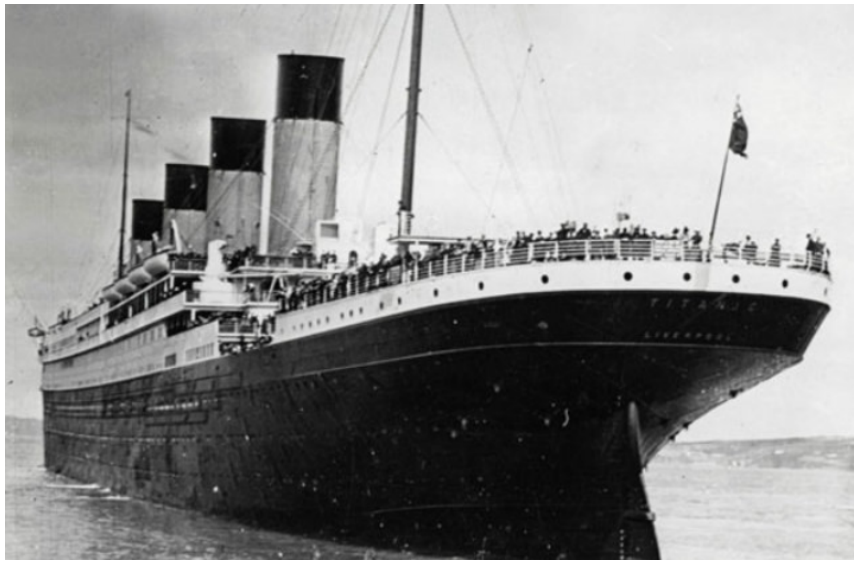
sibsp
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

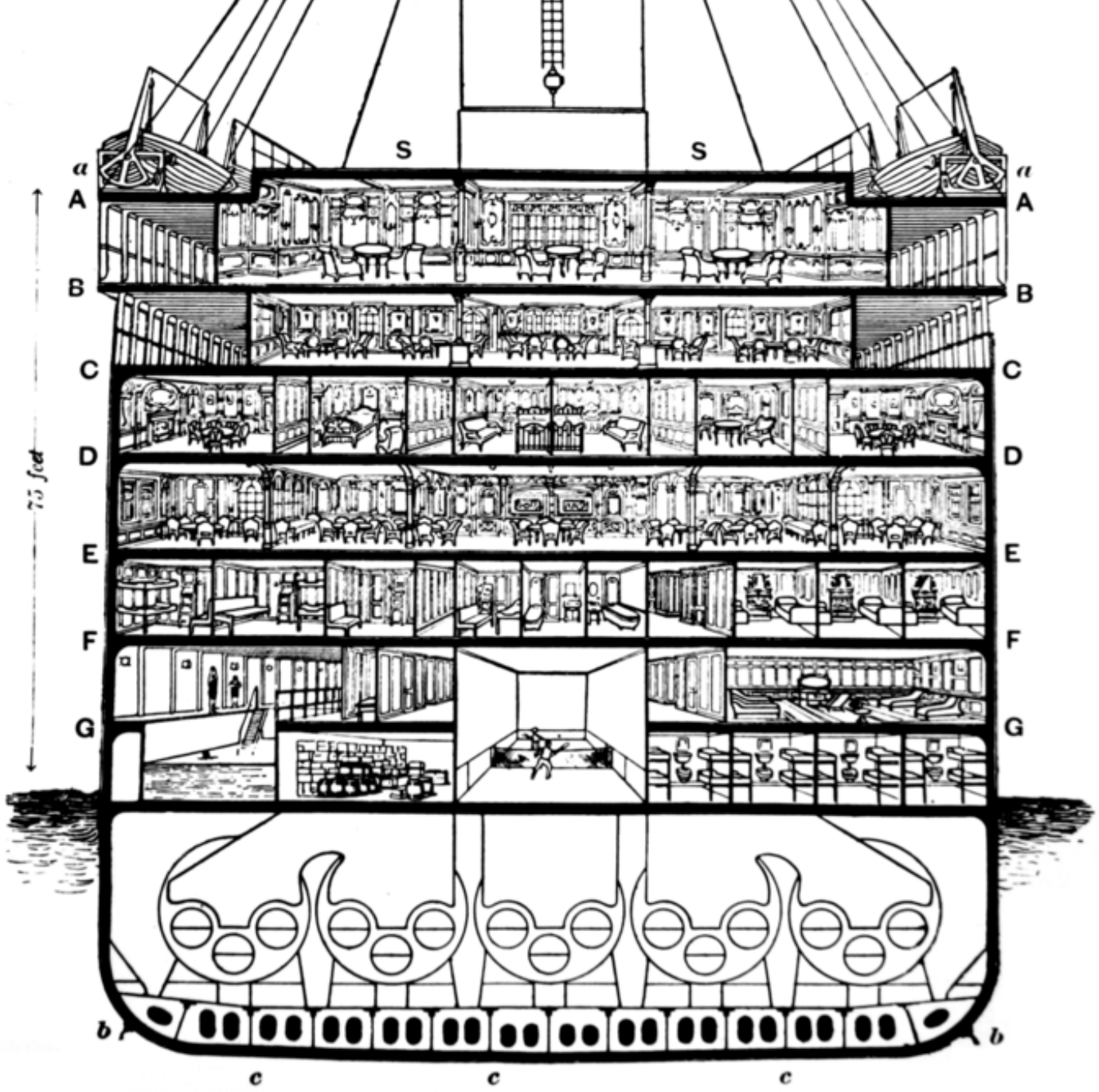
Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
① sibsp	# of siblings / spouses aboard the Titanic	
② parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

약혼자, 유모는 미포함

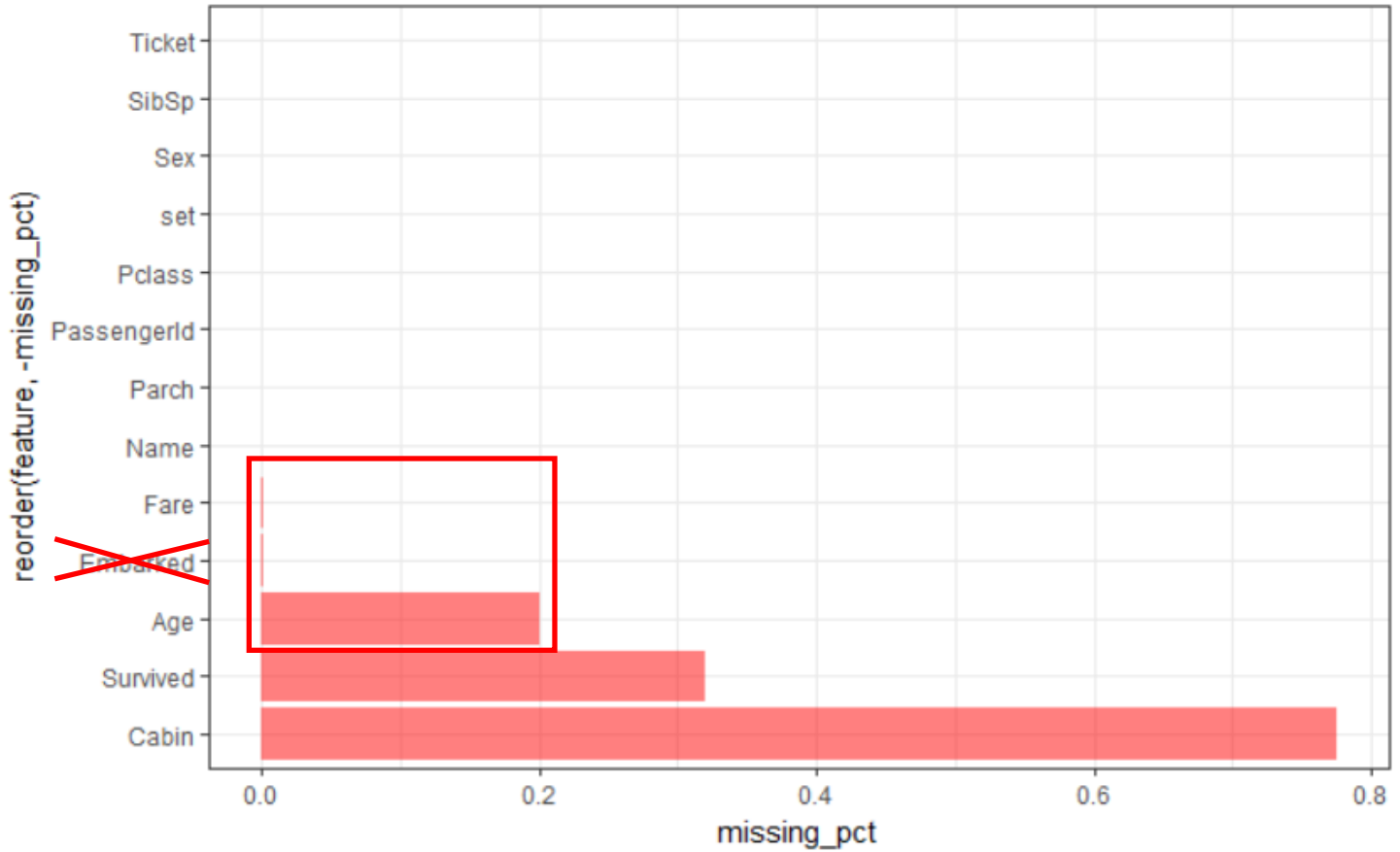
Data Dictionary



pclass: A proxy for socio-economic status
1st = Upper
2nd = Middle
3rd = Lower



Data 전처리

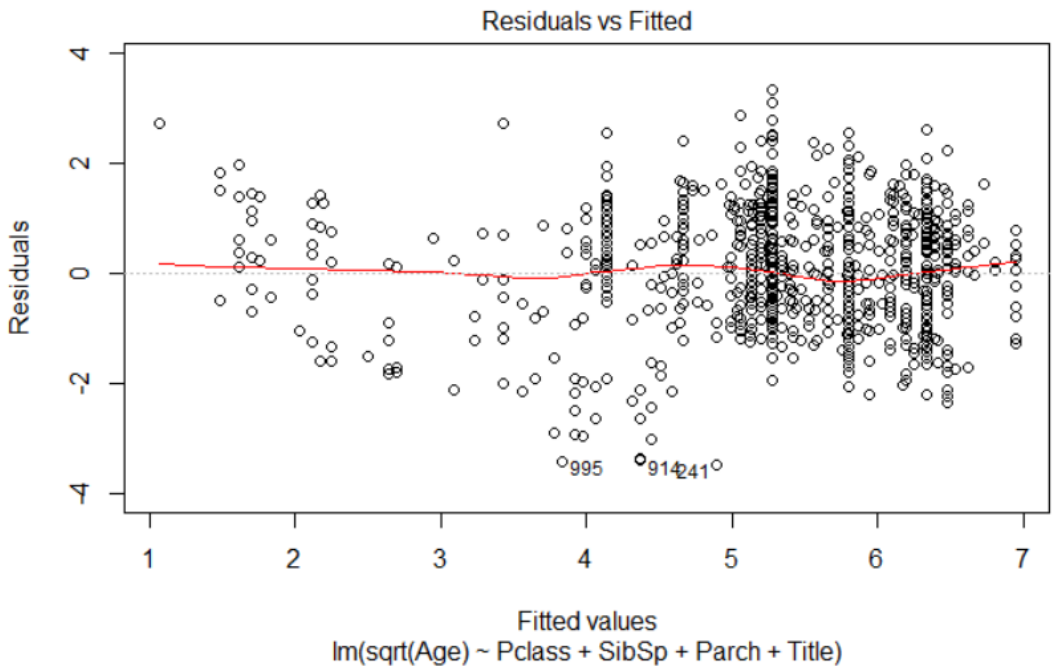


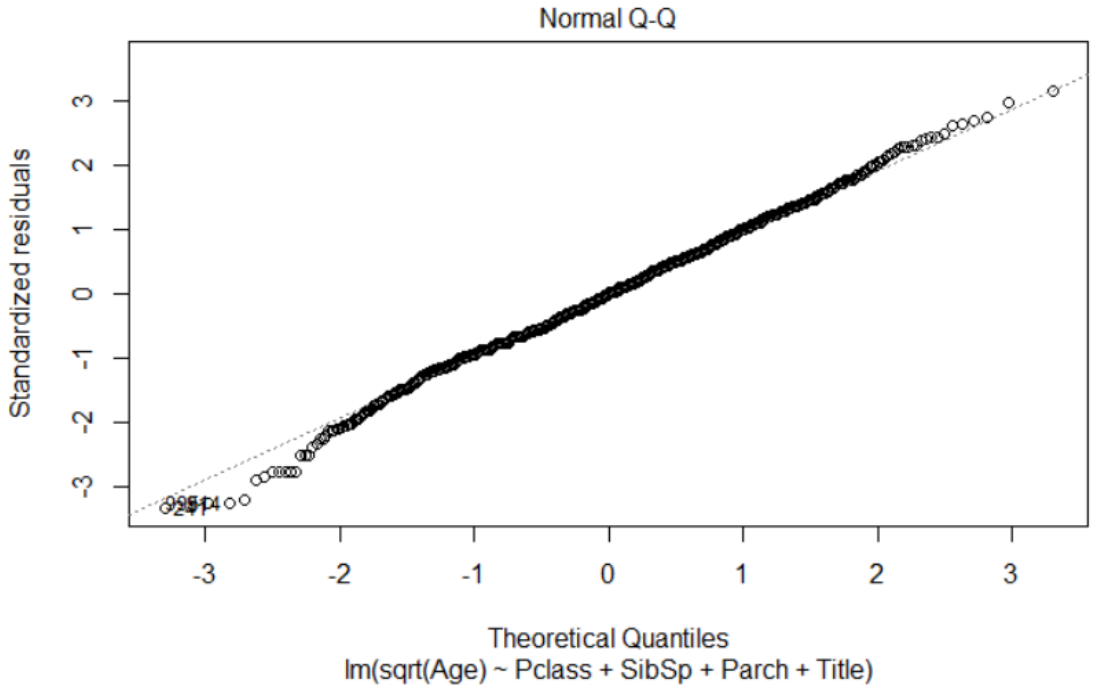
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
0	418	0	0	0	263	0	0	0	1	1014

Embarked	set	Title
2	0	0

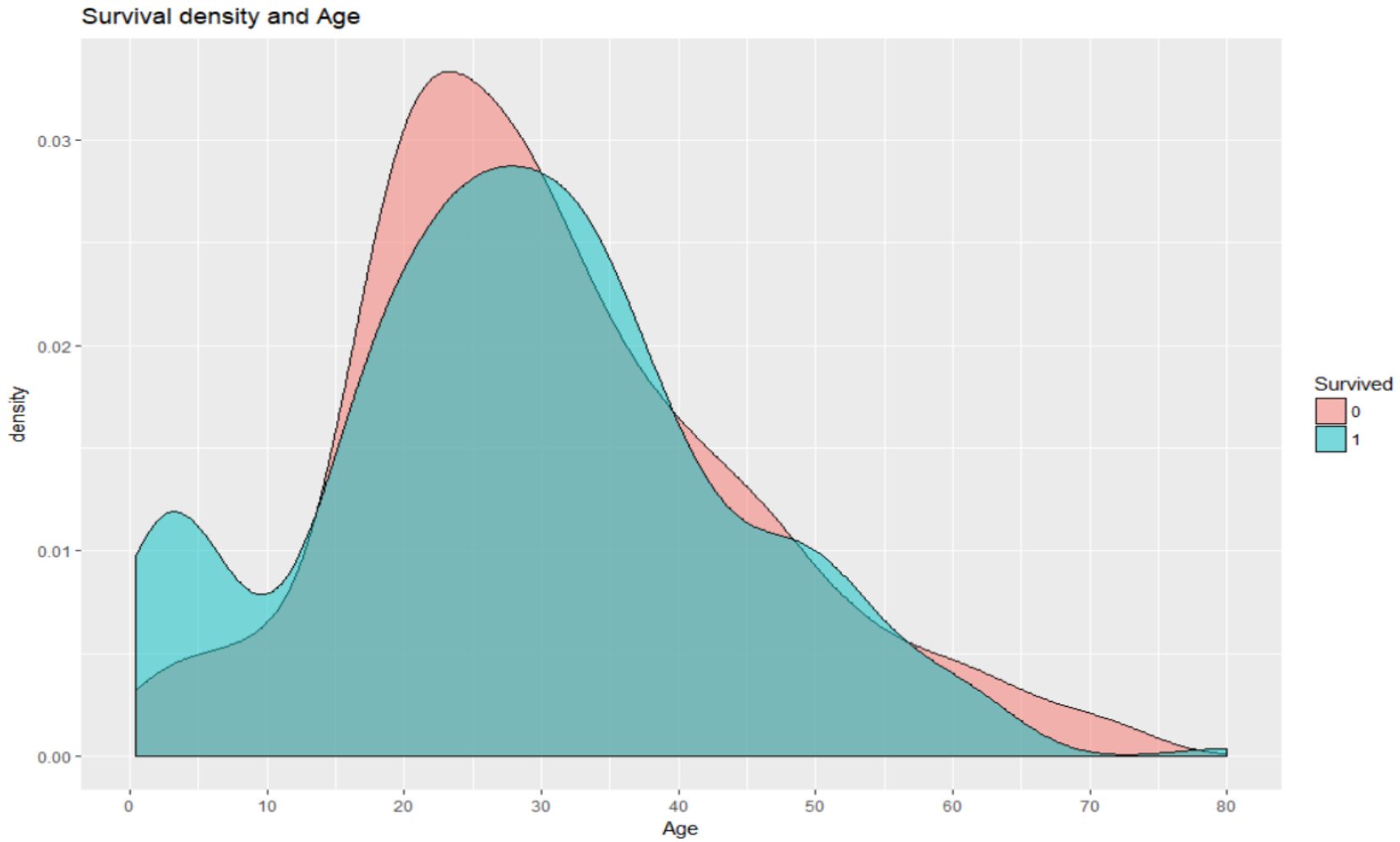
Data 전처리

```r
# 나이 예측(선형회귀)
AgeLM2 <- lm(sqrt(Age) ~ Pclass + SibSp + Parch + Title, data=full[!is.na(full$Age),])
summary(AgeLM2)
plot(AgeLM2)
vif(AgeLM2) # Sex, Title에서 vif가 각각 71.34, 98.62
```





Data 전처리



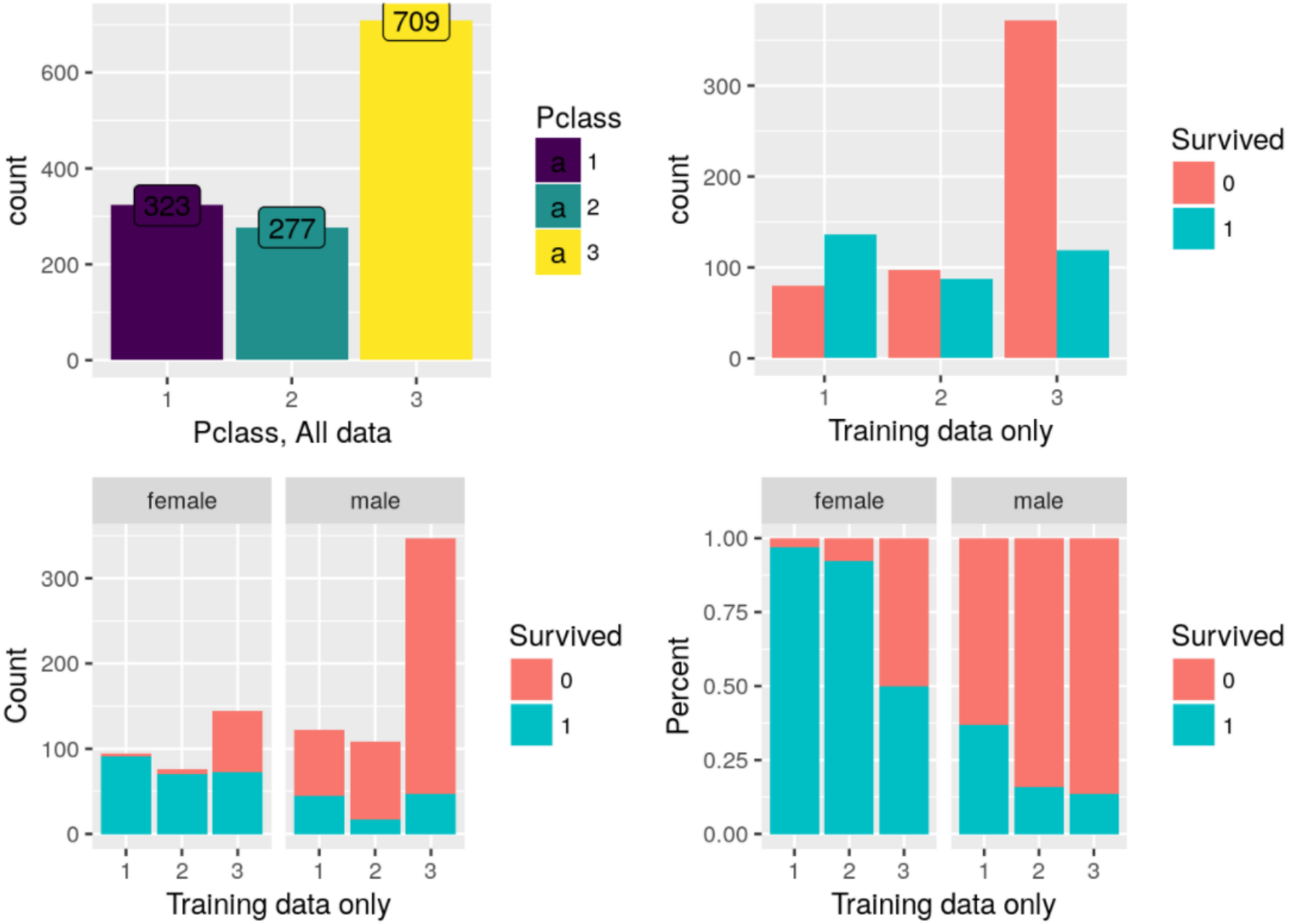

```
class(full$Age)
full$Agetype<- full %>% with(ifelse(Age >= 0 & 17> Age,"child",
                                    ifelse(Age >= 17 & 47> Age, "adult","older")))
```

Data 전처리

Moran, Mr. James	male
McCarthy, Mr. Timothy J	male
Palsson, Master. Gosta Leonard	male
Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female
Nasser, Mrs. Nicholas (Adele Achem)	female
Sandstrom, Miss. Marguerite Rut	female
Bonnell, Miss. Elizabeth	female

```
###MISS, Mrs, Master and Mr are taking more numbers

###Better to group Other Titles into bigger basket by ch

full$Title[full$Title == 'Mlle']        <- 'Miss'
full$Title[full$Title == 'Ms']          <- 'Miss'
full$Title[full$Title == 'Mme']         <- 'Mrs'
full$Title[full$Title == 'Lady']          <- 'Miss'
full$Title[full$Title == 'Dona']          <- 'Miss'

## I am afraid creating a new varible with small data ca
## However, My thinking is that combining below feauter
doctor and nobel peoples

full$Title[full$Title == 'Capt']        <- 'Officer'
full$Title[full$Title == 'Col']        <- 'Officer'
full$Title[full$Title == 'Major']   <- 'Officer'
full$Title[full$Title == 'Dr']   <- 'Officer'
full$Title[full$Title == 'Rev']   <- 'Officer'
full$Title[full$Title == 'Don']   <- 'Officer'
full$Title[full$Title == 'Sir']   <- 'Officer'
full$Title[full$Title == 'the Countess']   <- 'Officer'
full$Title[full$Title == 'Jonkheer']   <- 'Officer'

table(full$Title)
```

Data 전처리


709
323
277
600
400
200
0
count
1
2
3
Pclass, All data
Pclass
a 1
a 2
a 3
300
200
100
0
count
Training data only
Survived
0
1
female
male
300
200
100
0
Count
Training data only
Survived
0
1
female
male
1.00
0.75
0.50
0.25
0.00
Percent
Training data only
Survived
0
1

파생변수 생성

함께 티켓을 구매했다고 가족은 아니다.

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	set	Title
1	1	Cherry, Miss. Gladys	f	30.00000	0	0	110152	86.5000	B77	S	train	Miss
1	1	Maioni, Miss. Roberta	f	16.00000	0	0	110152	86.5000	B79	S	train	Miss
1	1	Rothes, the Countess. of (Lucy Noel Martha Dyer-Edwards)	f	33.00000	0	0	110152	86.5000	B77	S	train	Officer

FamilySize = > Sibsp(배우자와 형제) + Parch(부모와 자식) + 1(본인)

Ticketcnt = > 티켓을 함께 구매한 그룹의 수 (약혼자 + 유모 + 친구 + 직장동료 등등 관계가 있는 사람)

Groupsize = > MAX(FamilySize, Ticketcnt)

Namelength= > 이름의 길이?? => 부연설명

Vestrom, Miss. Hulda Amanda Adolfina	f
Hewlett, Mrs. (Mary D Kingcome)	f
Rice, Master. Eugene	m
Williams, Mr. Charles Eugene	m
Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	f
Masselmani, Mrs. Fatima	f
Fynney, Mr. Joseph J	m
Beesley, Mr. Lawrence	m
McGowan, Miss. Anna "Annie"	f

Data 전처리

최고점수 : 0.76076

[초기 데이터 예측 GBM 결과]

Submission and Description	Public Score	Use for Final Score
answer.csv 4 days ago by kim sun been add submission details	0.73205	☐
answer.csv 4 days ago by kim sun been add submission details	0.72248	☐
answer.csv 4 days ago by kim sun been add submission details	0.76076	☐
answer.csv 4 days ago by kim sun been add submission details	0.69377	☐

목표 => 80을 넘어보자

Data 전처리

최고점수 : 0.76076

[초기 데이터 예측 GBM 결과]

Submission and Description	Public Score	Use for Final Score
answer.csv 4 days ago by kim sun been add submission details	0.73205	☐
answer.csv 4 days ago by kim sun been add submission details	0.72248	☐
answer.csv 4 days ago by kim sun been add submission details	0.76076	☐
answer.csv 4 days ago by kim sun been add submission details	0.69377	☐

목표 => 80을 넘어보자

모델링

최고점수 : 0.76076

```
x<-c("PassengerId",        # 고객번호
     "Survived",           # 생존여부
     "Pclass",             # 객실등급
     "Name",               # 이름
     "Sex",                # 성별
     "Age",                # 나이
     "SibSp",              # 배우자 + 형제
     "Parch",              # 자식 + 부모
     "Ticket" ,            # 티켓
     "Fare" ,              # 요금
     "Cabin",              # 객실번호
     "Embarked",           # 탑승지
     "set",                # 훈련, 테스트
     "Title",              # 미스터, 미스
     "FamilySize" ,        # 가족크ㄱ
     "Pc_Se_Em",          # 성별 + 객실등급
     "ageGroup",           # 나이그룹
     "AgeBylm",            # 회귀예측 나이
     "AgetypeBylm",        # 회귀예측 나이그룹
     "Ticketcnt",          # 티켓이 같은 그룹 수
     "Groupsize",          # 가족수과 티켓수 중 MAX
     "FamilyPer",          # 그룹 중 가족의 비율
     "NonFamilyPer",       # 그룹 중 가족이 아닌 사람의 비유
     "adult",              # 그룹 내 어른의 수
     "child",              # 그룹 내 아이의 수
     "older",              # 그룹 내 노인의 수
     "AdultPer",           # 그룹 내 어른의 비율
     "ChildPer",           # 그룹 내 아이의 비율
     "OlderPer",           # 그룹 내 노인의 비율
     "Alone",              # 혼자인지 아닌지
     "subNick",
     "firstname"
     )
```

model1

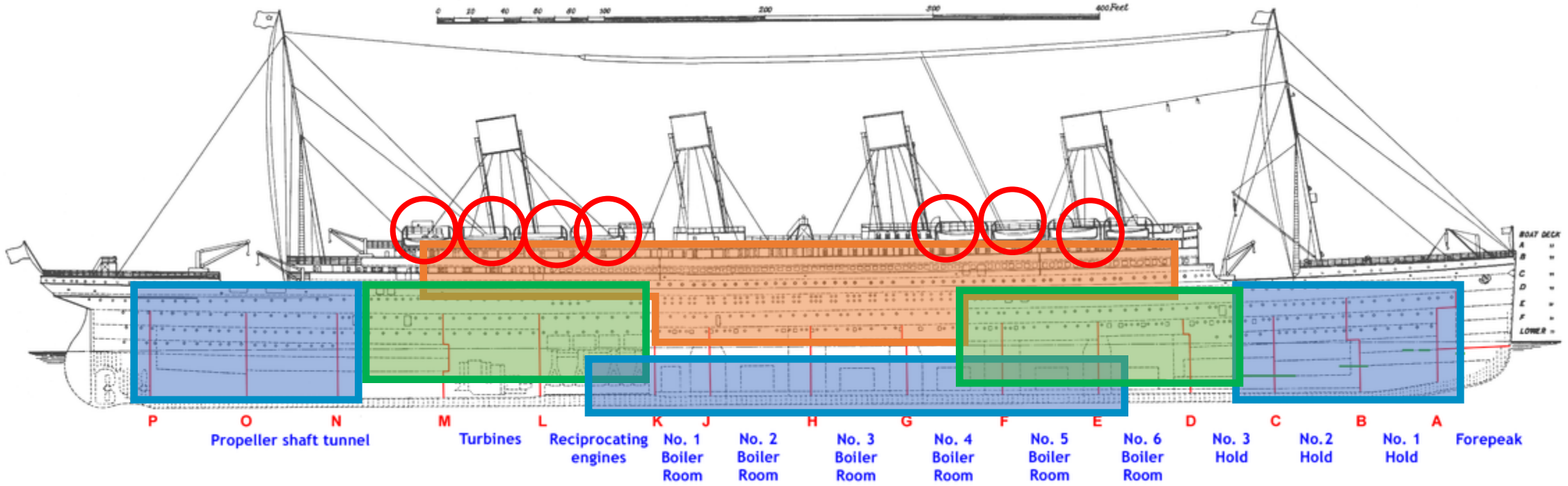
혼자 배를 탄 경우

model2

가족 또는 그룹이 같이 탄 경우

모델링

최고점수 : 0.76076



모델링

최고점수 : 0.76076

혼자 배를 탄 경우

Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:

	0	1	Error	Rate
0	298	29	0.088685	=29/327
1	42	79	0.347107	=42/121
Totals	340	108	0.158482	=71/448

Maximum Metrics: Maximum metrics at their respective thresholds

	metric	threshold	value	idx
1	max f1	0.447225	0.689956	103
2	max f2	0.089836	0.703226	275
3	max f0point5	0.447225	0.714286	103
4	max accuracy	0.447225	0.841518	103
5	max precision	0.988863	1.000000	0
6	max recall	0.018699	1.000000	397
7	max specificity	0.988863	1.000000	0
8	max absolute_mcc	0.447225	0.585672	103
9	max min_per_class_accuracy	0.206742	0.733945	170
10	max mean_per_class_accuracy	0.447225	0.782104	103

가족 또는 그룹이 같이 탄 경우

Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:

	0	1	Error	Rate
0	165	57	0.256757	=57/222
1	16	205	0.072398	=16/221
Totals	181	262	0.164786	=73/443

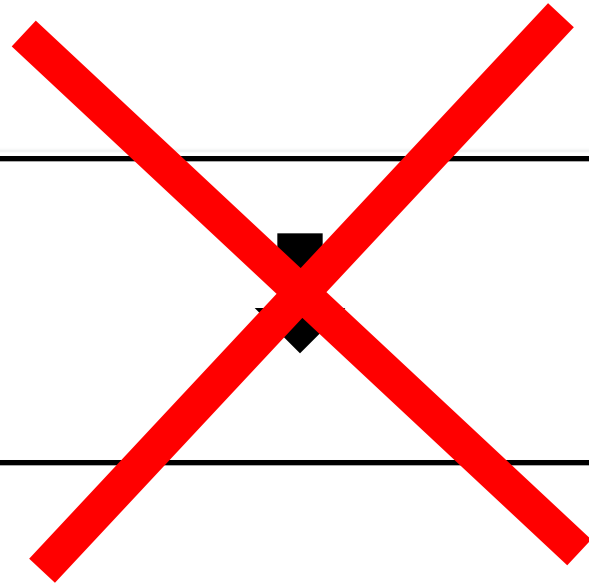
Maximum Metrics: Maximum metrics at their respective thresholds

	metric	threshold	value	idx
1	max f1	0.289958	0.848861	232
2	max f2	0.258964	0.894555	243
3	max f0point5	0.653313	0.891192	159
4	max accuracy	0.653313	0.857788	159
5	max precision	1.000000	1.000000	0
6	max recall	0.001313	1.000000	399
7	max specificity	1.000000	1.000000	0
8	max absolute_mcc	0.653313	0.724582	159
9	max min_per_class_accuracy	0.495616	0.833333	194
10	max mean_per_class_accuracy	0.650131	0.857619	161

모델링

최고점수 : 0.76076

ans_gbm.csv
2 hours ago by BEENSUN
분리모델 CV score : 0.80이상
0.69377



$$ 조정값 조정

ans_rf.csv
an hour ago by BEENSUN
분리모델 조정값 0.6이상
0.75598

모델링

최고점수 : 0.76076

	model_id	auc	logloss
1	GBM_grid_0_AutoML_20180512_054256_model_1	0.8700772	0.4167367
2	StackedEnsemble_BestOfFamily_0_AutoML_20180512_054256	0.8675474	0.4097578
3	GBM_grid_0_AutoML_20180512_054256_model_3	0.8650757	0.4264711
4	GLM_grid_0_AutoML_20180512_054256_model_0	0.8648438	4117769
5	GBM_grid_0_AutoML_20180512_054256_model_0	0.8647320	0.4 649
6	StackedEnsemble_AllModels_0_AutoML_20180512_054256	0.8641938	0.41053
7	GBM_grid_0_AutoML_20180512_054256_model_2	0.8638129	0.4240346
8	DeepLearning_grid_0_AutoML_20180512_054256_model_1	0.8620284	0.4961315
9	DRF_0_AutoML_20180512_054256	0.8603102	0.7032464
10	GBM_grid_0_AutoML_20180512_054256_model_4	0.8582028	0.418766
11	DeepLearning_grid_0_AutoML_20180512_054256_model_0	0.8581945	0.49 0
12	DeepLearning_grid_0_AutoML_20180512_054256_model_2	0.8562941	44867
13	DeepLearning_0_AutoML_20180512_054256	0.8558221	0.4510615
14	XRT_0_AutoML_20180512_054256	0.8550644	0.5745181
15	DeepLearning_grid_0_AutoML_20180512_054256_model_4	0.8478976	0.4765189
16	GBM_grid_0_AutoML_20180512_054256_model_5	0.8477154	0.4432105
17	DeepLearning_grid_0_AutoML_20180512_054256_model_3	0.8363295	0.5830514
18	GBM_grid_0_AutoML_20180512_054256_model_6	0.7804892	0.6469794

[다시 통합모델로 넘어와서~]

ans_ml.csv 0.75119

a few seconds ago by BEENSUN

오토엠엘 CV : 0.87

모델링

최고점수 : 0.79425

Variable Importances:

	variable	relative_importance	scaled_importance	percentage
1	Pc_Se_Em	468.547607	1.000000	[illegible]
2	AgeBylm	134.618301	0.287310	0.150447
3	Fare	119.734665	0.255541	0.133813
4	Title	51.424274	0.109753	0.057[illegible]
5	Groupsize	42.953259	0.091673	0.048004
6	ChildPer	25.983786	0.055456	0.029039
7	FamilyPer	18.013582	0.038446	0.020132
8	AdultPer	15.048985	0.032118	0.016818

개꿀!

ans_rf(0.79).csv
2 hours ago by BEENSUN
add submission details

0.79425

모델링

최고점수 : 0.81818

ans_rf(0.79).csv
2 hours ago by BEENSUN
add submission details
0.79425

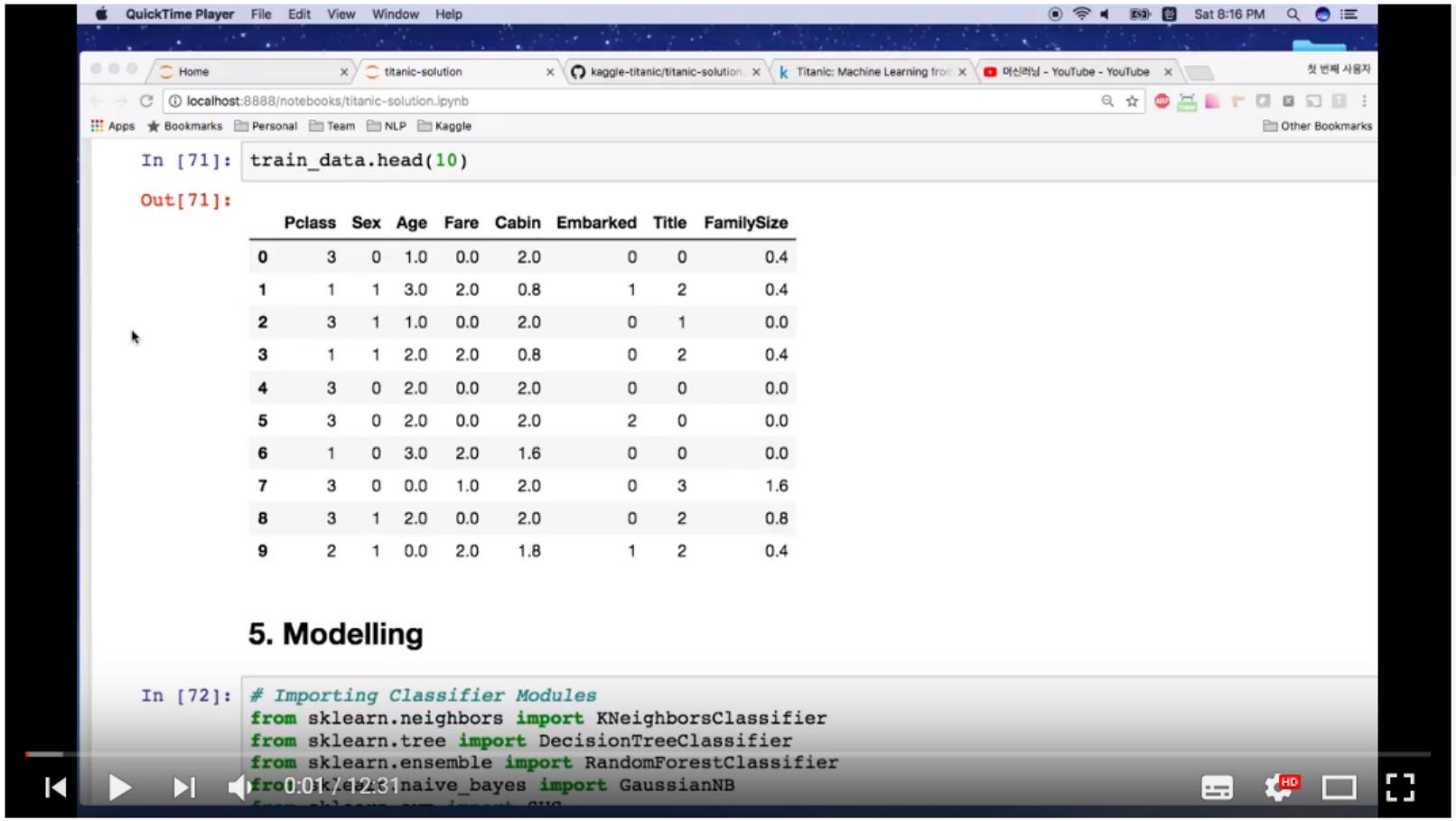
$$ 조정값 조정

앙상블4.csv
24 minutes ago by BEENSUN
앙상블4
0.81818

앙상블3.csv
28 minutes ago by BEENSUN
앙상블 3
0.77990

앙상블2.csv
32 minutes ago by BEENSUN
통합 RF + GBM
0.76076

또 우연히 걸림



캐글 (Kaggle) 실전 데이터분석 배우기

Minsuk Heo 허민석 - 4 / 4

1 캐글 (Kaggle) 소개 - 데이터 과학 (머신러닝) 실전 예제 다루기 — Minsuk Heo 허민석 — 2:57

2 캐글 - 타이타닉 생존자 예측하기 [1/3] - 데이터 분석 — Minsuk Heo 허민석 — 11:38

3 캐글 - 타이타닉 생존자 예측하기 [2/3] - Feature Engineering — Minsuk Heo 허민석 — 17:26

▶ 캐글 - 타이타닉 생존자 예측하기 [3/3] - modeling, validation, testing — Minsuk Heo 허민석 — 12:32

캐글 - 타이타닉 생존자 예측하기 [3/3] - modeling, validation, testing

조회수 1,142회

16 0 공유

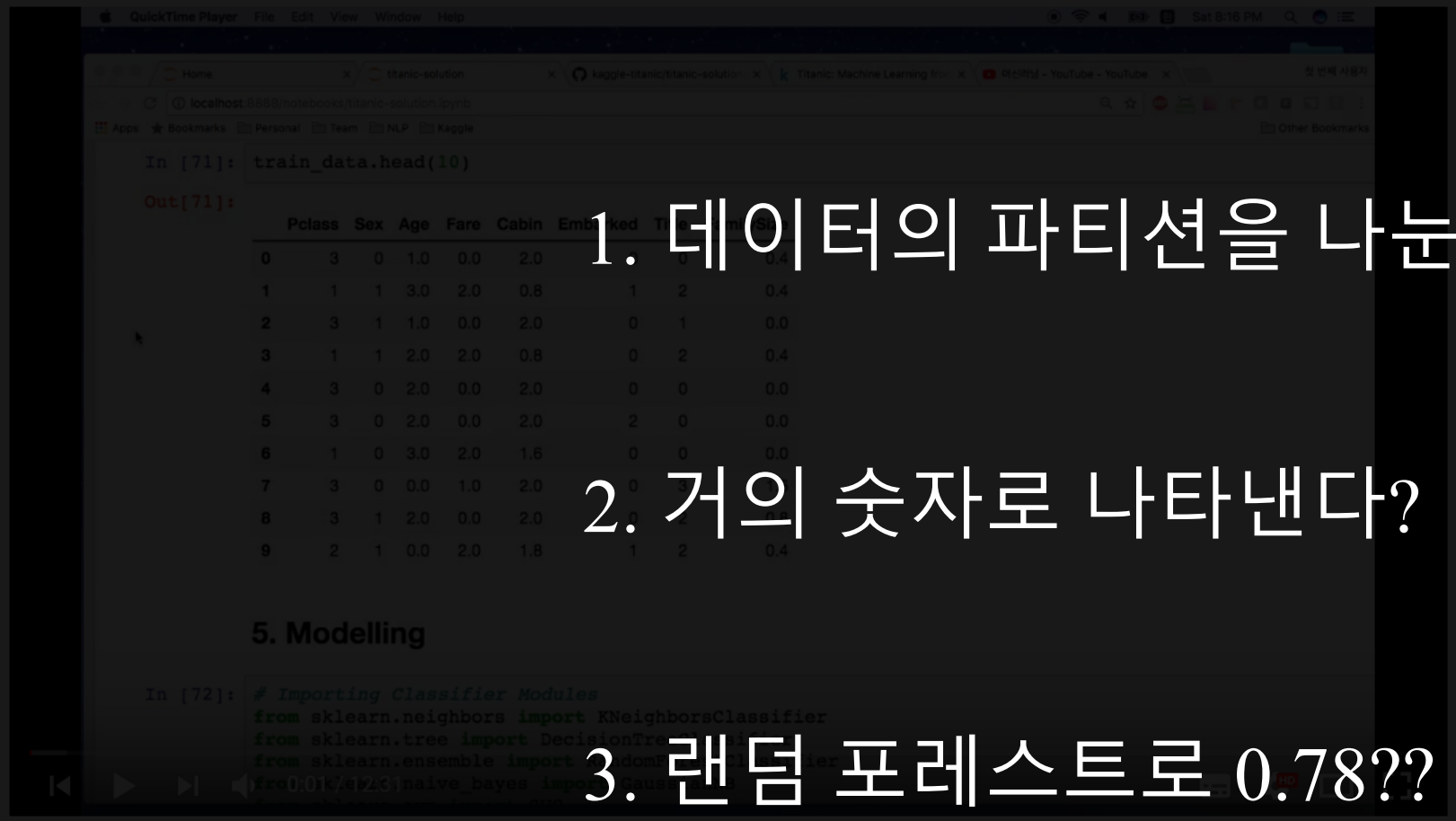
Minsuk Heo 허민석
게시일: 2017. 10. 28.

구독중 4.6천

https://github.com/minsuk-heo/kaggle-...

강의 맛보기
광고 realclass.co.kr
매진임박

YouTube KR
kaggle 타이타닉



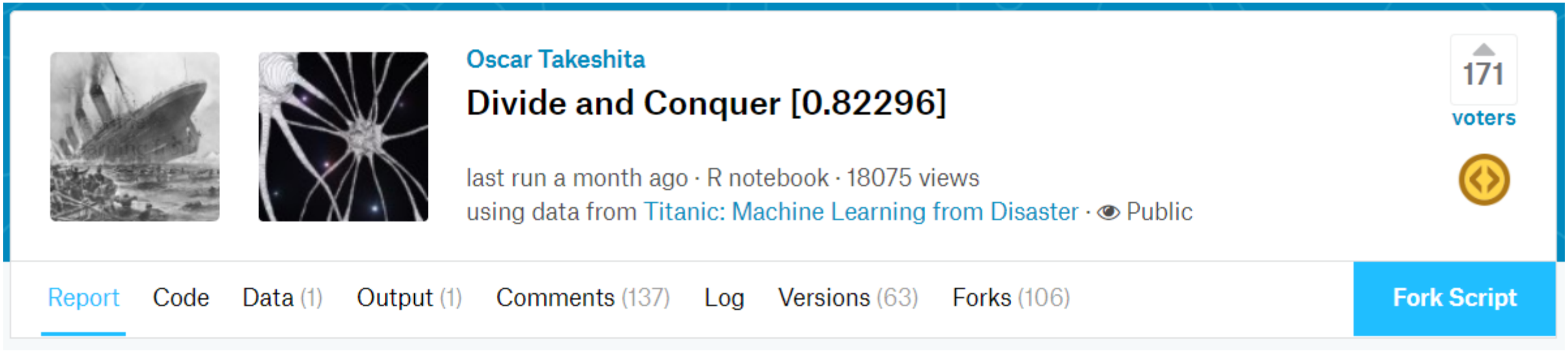
1. 데이터의 파티션을 나눈다.

2. 거의 숫자로 나타낸다?

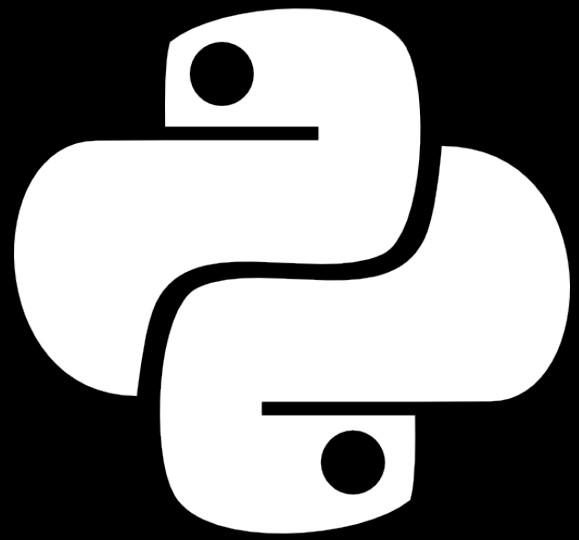
3. 랜덤 포레스트로 0.78??

캐글 - 타이타닉 생존자 예측하기 [3/3] - modeling, validation, testing
조회수 1,142회
16 0 공유
Minsuk Heo 허민석
게시일: 2017. 10. 28.
https://github.com/minsuk-heo/kaggle-...
구독중 4.6천
캐글 (Kaggle) 실전 데이터분석 배우기
Minsuk Heo 허민석 - 4 / 4
캐글 (Kaggle) 소개 - 데이터 과학 (머신러닝) 실전 예제 다루기 2:57 Minsuk Heo 허민석
캐글 - 타이타닉 생존자 예측하기 [1/3] - 데이터 분석 11:38 Minsuk Heo 허민석
캐글 - 타이타닉 생존자 예측하기 [2/3] - Feature Engineering 17:26 Minsuk Heo 허민석
캐글 - 타이타닉 생존자 예측하기 [3/3] - modeling, validation, testing 12:32 Minsuk Heo 허민석
LEARN FROM THE REAL
강의 맛보기
광고 realclass.co.kr
매진임박

시간 날때..



이 커널을 좀 다시 보기로...




R

Q&A