

LS 빅데이터 스쿨 데이터 레디니스(readiness) 분석 보고서

1. 비즈니스 요구사항
2. 데이터 레디니스 과제 개요
3. 타이타닉 생존예측 모델 개발 -
 - 1) 데이터를 연결하고 이해하였습니다.
 - 2) 데이터에 귀를 기울여 보았습니다.
 - 3) 정제된 데이터로 기계학습을 진행하였습니다.
 - 4) 모델의 예측결과를 해석했습니다.
4. 과제 요약 및 앞으로의 방향성 제안

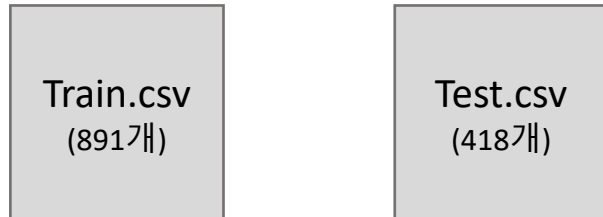
2023. 10. 01 (일)

○○○ 과제수행팀

1. 비즈니스 요구사항

타이타닉 탑승객 생존 예측모델 생성

- 캐글(Kaggle)의 타이타닉 생존/사망자 메타데이터를 분석하여 사망 원인 분석



- pclass: 사회/경제적 지위 값 (1: 1등실, 2: 2등실, 3: 3등실)
- sibSp: 타이타닉호에 탑승한 형제자매/배우자 수
- parch: 타이타닉호에 탑승한 부모/자녀 수
- ticket: 티켓번호
- fare: 요금
- cabin : 객실번호
- embarked: 승선항 (c = 세르부르, Q = 퀸스타운, s = 사우샘프턴)
- 비교적 적은 데이터로 예측모델 구축 → 주어진 데이터를 최대한 사용해야함

보유 데이터 점검하여,
부족한 부분 점검

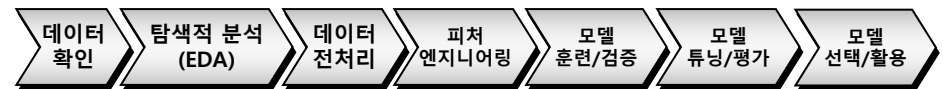
각 데이터 간의 상관성 분석 및
새로운 변수 생성

타이타닉 탑승객 생존 예측
모델 생성

타이타닉 탑승객 생존 예측모델 생성

- Train(학습)데이터와 Test(검증)데이터를 분석하여 높은 성능의 생존예측모델 생성

학습데이터 검증데이터



- 모델 성능 평가 : 정확도(Accuracy)*로 평가
- 학습 데이터: 성별, 요금, 이름, 객실, 생존여부 등 변수 존재
- 검증 데이터: 생존여부 변수를 제외한 동일한 변수 존재

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

where TP = True positive; FP = False positive; TN = True negative; FN = False negative

* 정확도 = $\frac{\text{예측 결과가 동일한 데이터 건수}}{\text{전체 예측 데이터 건수}}$

- 적은 데이터로 인한 과적합을 최대한 방지하며 예측모델 생성
- 탐색적 데이터 분석 시도 → 피처 엔지니어링 → 예측모델 생성

2. 데이터 레디니스 과제 개요

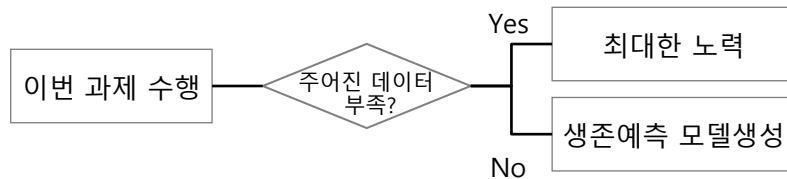
과제 목표 및 일정

▪ 학습 데이터로 분석 가설 검증

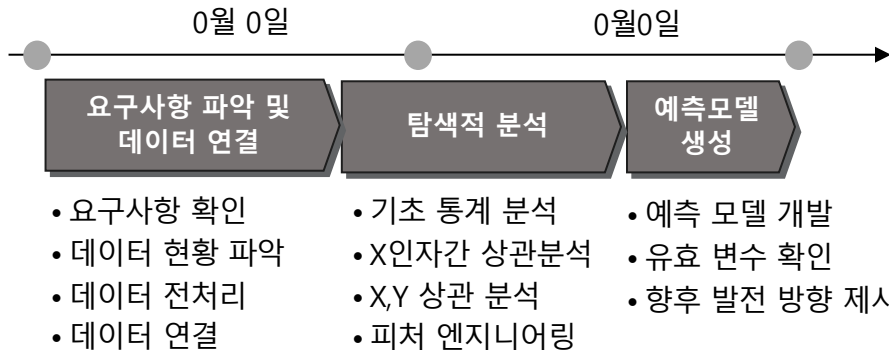
독립변수 - 생존과 밀접한 독립변수 조사

- 정의**
- 생존을 비교를 통한 유의한 의미 분석
 - 생존 변수와 독립변수 간 상관성 분석
 - 피처 엔지니어링을 통한 새로운 변수 생성

▪ 독립변수 정의, 피처 엔지니어링을 통한 모델 성능 향상

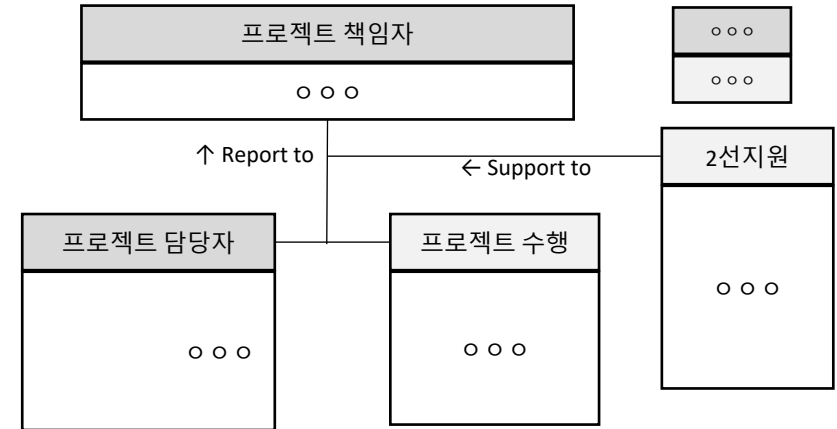


▪ OO일 동안 데이터 탐색 및 생존 예측모델 생성



과제 일정 및 수행 인원

▪ 팀별 내 분석 경험 풍부한 전문 분석 인력 수행



- ○○○ 프로젝트 수행 인력과 ○○프로젝트 담당자 협업

- ○○○○

3. 타이타닉 생존예측 모델 개발 – 1) 데이터를 연결하고 이해하였습니다.

결측치 보정 및 변수 명세 파악

- 학습데이터와 검증데이터에 나타난 결측치는 'Age', 'Cabin', 'Embarked', 'Fare' 4개 변수가 있음

학습데이터(Train set)	검증데이터(Test set)
Age 변수 결측치: 177	Age 변수 결측치: 86
Fare 변수 결측치: 0	Fare 변수 결측치: 1
Cabin 변수 결측치: 687	Cabin 변수 결측치: 327
Embarked 변수 결측치: 2	Embarked 변수 결측치: 0

- 학습데이터(Train) 총 891행 12열
- 검증데이터(Test) 총 418행 11열

구분	독립변수		종속변수	
Train	Pclass	891	생존	342
		
	Fare	891	사망	549
Test	Pclass	418	-	
		
	Fare	417		


학습 데이터 전처리

- Age : 학습데이터에 177개의 결측을 보완하기 위해 상관 계수로 데이터 관계성 파악

변수1	변수2	상관계수	비고
Age	Pclass	0.369226	
Age	SibSp	0.308247	
Age	Parch	0.189119	
Age	Fare	0.096067	미비
Age	Survived	0.077221	미비

- 사회적 지위를 나타내는 Pclass와 Age와 높은 상관성이 높음
- Pclass별, 성별로 그룹의 중앙값을 결측치 대신에 넣음

Pclass	Sex	Age
1	female	NaN
1	male	NaN
2	female	NaN
2	male	NaN
3	female	NaN
3	male	NaN



Pclass	Sex	Age
1	female	35
1	male	40
2	female	28
2	male	30
3	female	21.5
3	male	25

3. 타이타닉 생존예측 모델 개발 – 1) 데이터를 연결하고 이해하였습니다.

학습 데이터 전처리

- **Embarked** : 학습데이터에 2개의 결측을 보완하기 위해 데이터를 확인

Pclass	Name	Sex	Age	Ticket	Fare	Cabin	Embarked
1	Icard, Miss. Amelie	female	38.0	113572	80.0	B28	NaN
1	Stone, Mrs. George Nelson (Martha Evelyn)	female	62.0	113572	80.0	B28	NaN

- 두명의 승객은 같은 Ticket번호와 같은 객실을 사용했으므로, 함께 승선한 것을 알 수 있음
- 실제 인물의 이름*을 가지고 검색해보면, Southampton에서 승선한 것으로 나타남.

Pclass	Name	Sex	Age	Ticket	Fare	Cabin	Embarked
1	Icard, Miss. Amelie	female	38.0	113572	80.0	B28	S
1	Stone, Mrs. George Nelson (Martha Evelyn)	female	62.0	113572	80.0	B28	S

* Martha Evelyn: <https://www.encyclopedia-titanica.org/titanic-survivor/martha-evelyn-stone.html>

학습 데이터 전처리

- **Fare** : 검증데이터에 1개의 결측을 보완하기 위해 데이터를 확인

Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
3	male	60.5	0	0	3701	NaN	NaN

- SibSp, Parch의 값이 0인 것으로 보아 가족 없이 혼자 탑승함
- Pclass가 3인 것으로 보아 3등석에 탑승함
- 동승자가 없는 3등석 남자들의 중앙값을 결측치 대신에 넣음

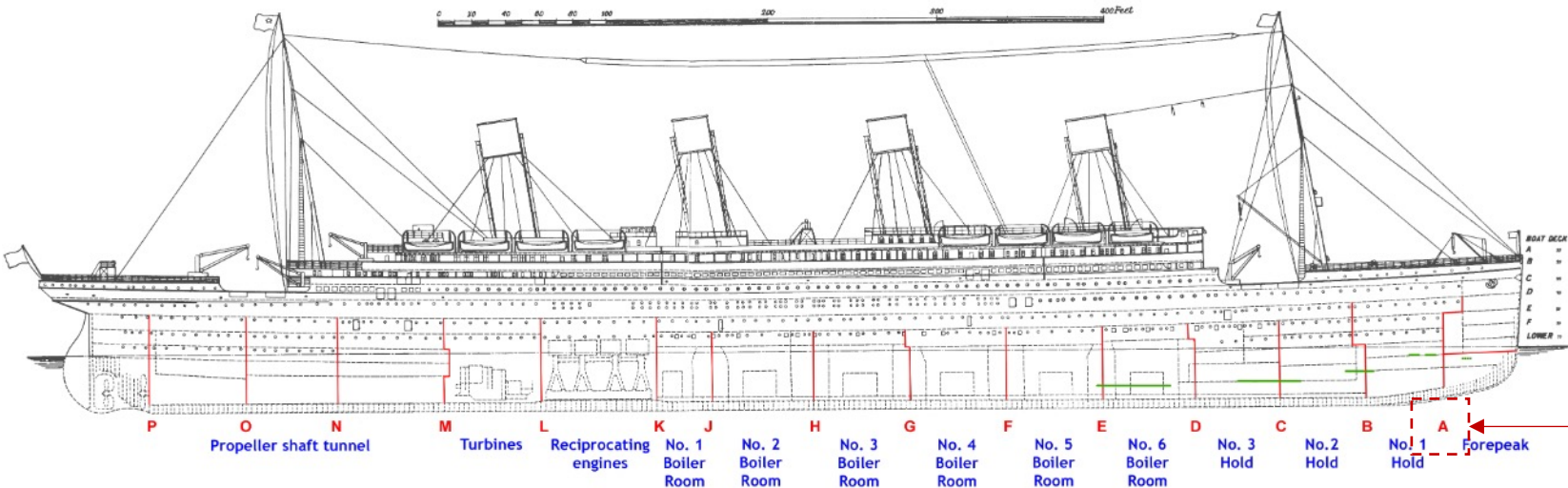
Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
3	male	60.5	0	0	3701	7.8958	NaN

3. 타이타닉 생존예측 모델 개발 – 1) 데이터를 연결하고 이해하였습니다.

학습 데이터 전처리

- Cabin : 학습데이터에 687개, 검증데이터에 327개의 결측치가 존재함.
 - 객실을 의미하는 Cabin은 결측값이 많으며, 앞 글자는 갑판인 "Deck"을 의미함
 - Pclass와 요금에 따라 Cabin은 달라지므로 확인이 필요함.

Pclass	Sex	Age	SibSp	Parch	Ticket	Cabin
1	male	28	0	0	113788	A6
1	female	35	1	0	PC 17569	B78
1	female	35	1	0	113803	C123
1	male	54	0	0	17463	E46
3	female	4	1	1	PP 9549	G6



3. 타이타닉 생존예측 모델 개발 – 2) 데이터에 귀를 기울여 보았습니다. (1/9)

피쳐 엔지니어링

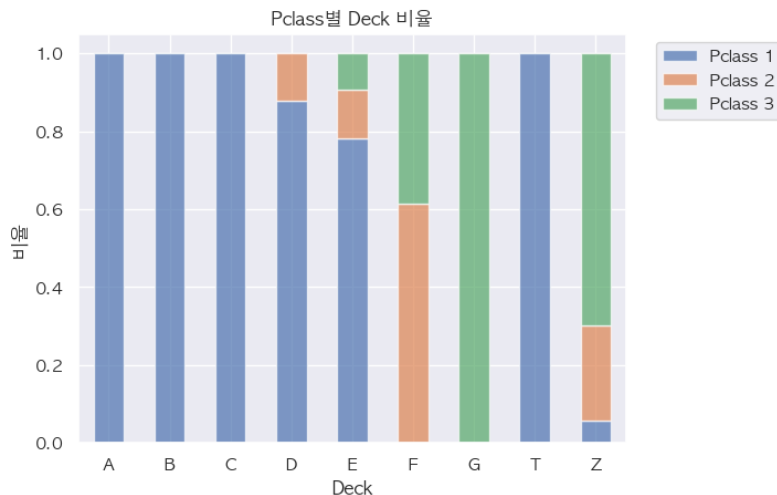
▪ Cabin에서 Deck (갑판) 변수 추출 → 신규 변수 추가

- 결측값에는 'Z'를 추가하였음

Cabin	Deck
A6	A
B78	B
C123	C
NaN	Z

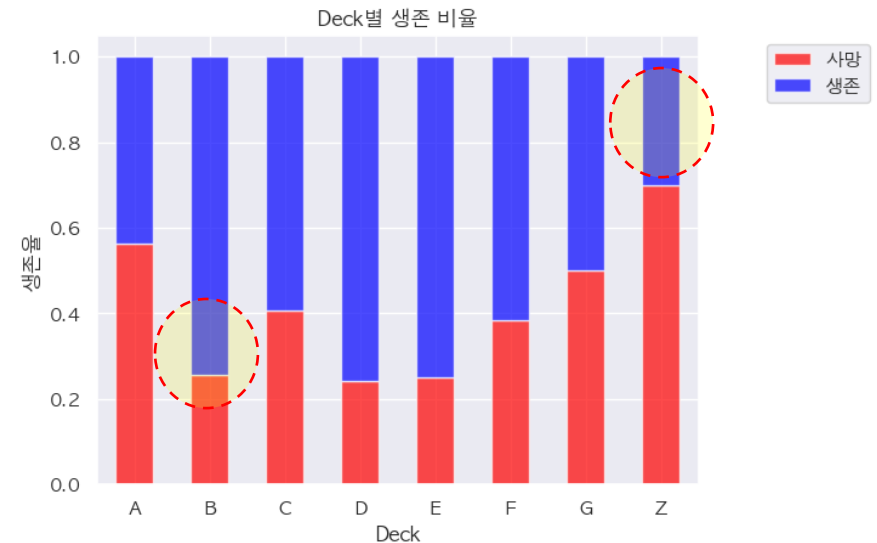
▪ 각 Pclass별 승선한 Deck의 비율 비교

- A, B, C 갑판은 1등석 전용칸, F, G는 2등, 3등석 전용칸
- D, E는 모든 클래스를 포함한 칸
- T는 데이터가 1개 뿐이고, 1등석이라 'T' → 'A' 로 변환 가능



Deck별 생존율 비교

▪ Deck별 Pclass가 다르므로 생존/사망의 비율 확인



▪ Deck별 생존/사망율이 다르므로, 독립변수로 추가 가능

- B, D, E 갑판의 생존율이 가장 높음 (Pclass가 대부분 1등석)
- 결측치였던 Z 갑판의 생존율이 가장 낮음 (Pclass가 대부분 3등석)
- Pclass 비율과 생존율이 비슷한 패턴끼리 그룹화

Deck	Deck
A, B, C	ABC
D, E	DE
F, G	FG
Z	Z

3. 타이타닉 생존예측 모델 개발 – 2) 데이터에 귀를 기울여 보았습니다. (2/9)

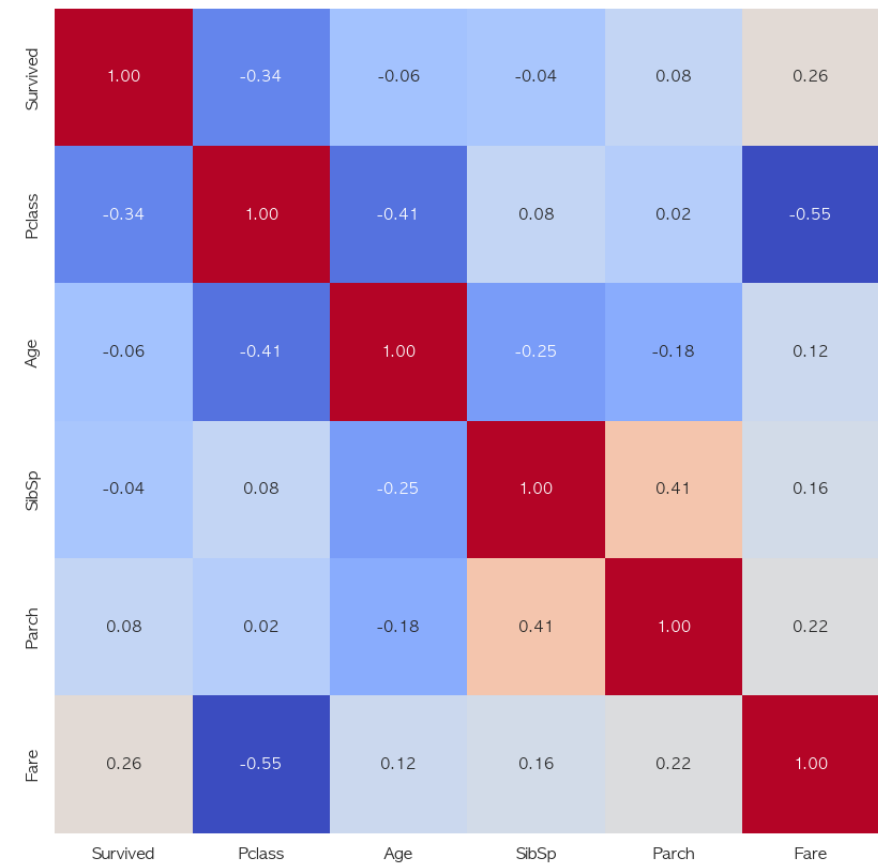
변수간 상관성 분석

Q. 생존과 독립 변수들 간의 상관성은?

A. 상관성이 큰 순서대로 나열하면 다음과 같음

X1	X2	상관계수(절대값)	비고
Fare	Pclass	0.550	높은 상관관계
Parch	SibSp	0.415	
Age	Pclass	0.414	
Survived	Pclass	0.338	
Fare	Survived	0.257	
SibSp	Age	0.250	
Fare	Parch	0.216	
Age	Parch	0.176	
SibSp	Fare	0.160	비교적 약한 상관관계
Age	Fare	0.123	

- 독립변수들 간의 상관성이 있는 것으로 나타남.
- 변수들 간 관계성이 있다고 판단되므로 각 변수별 분포를 확인하고 피처 엔지니어링을 통해 신규 변수를 생성할 수 있음.

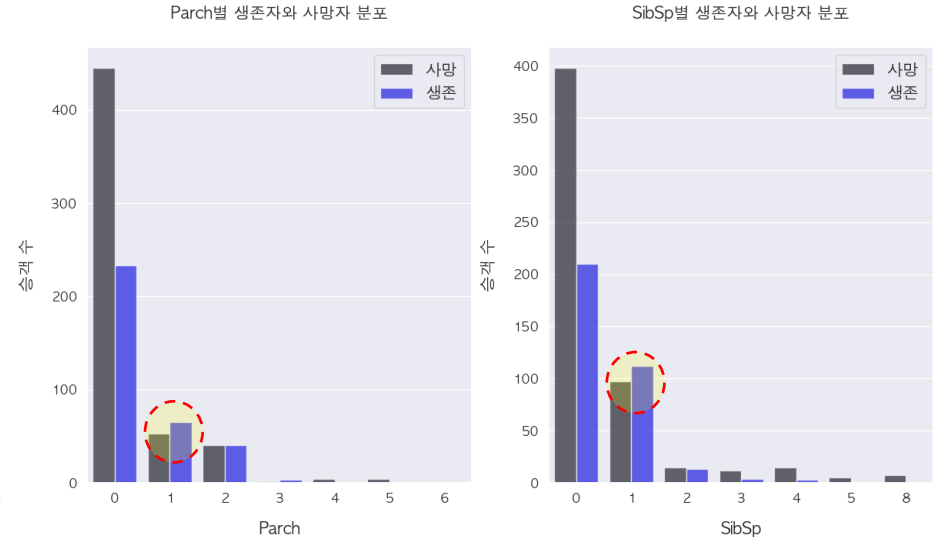
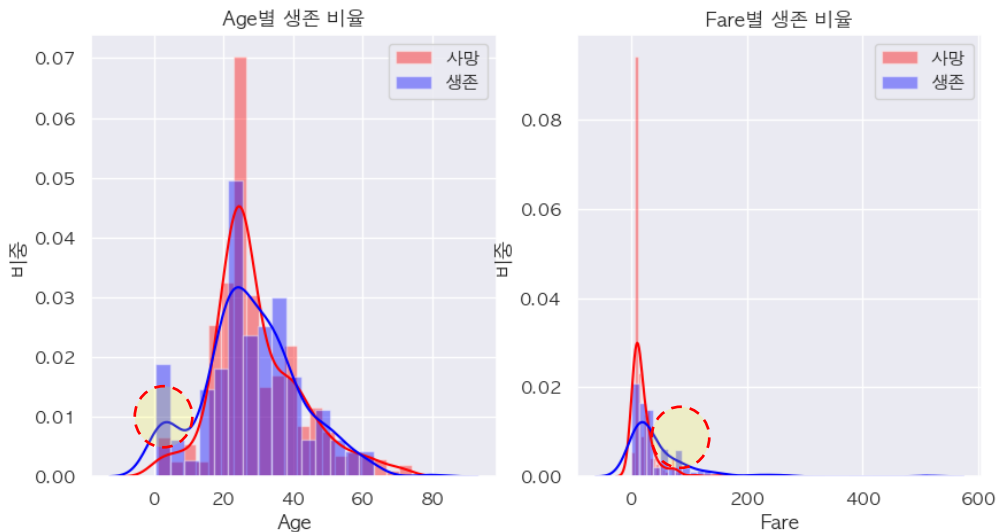


3. 타이타닉 생존예측 모델 개발 – 2) 데이터에 귀를 기울여 보았습니다. (3/9)

연속형 변수간 생존율 비교

■ 나이, 요금, 가족 탑승 수에 따른 사망률 비교 → 특정 구간(범위) 마다 생존율이 다름

- Age를 보면 20세 전으로 생존율이 높은 구간이 확인됨.
- Fare를 보면 타이타닉 탑승비용이 비쌀 수록 생존율이 높음.
- SibSp, Parch를 보면 혼자 탄 승객보다 가족이 1명 동승한 승객의 생존율이 높음.

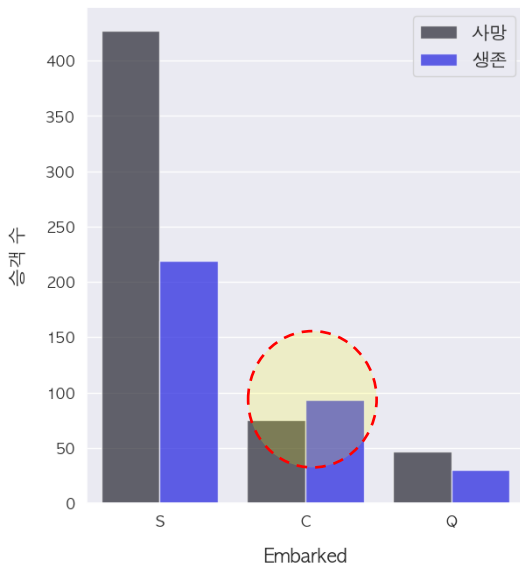


3. 타이타닉 생존예측 모델 개발 – 2) 데이터에 귀를 기울여 보았습니다. (4/9)

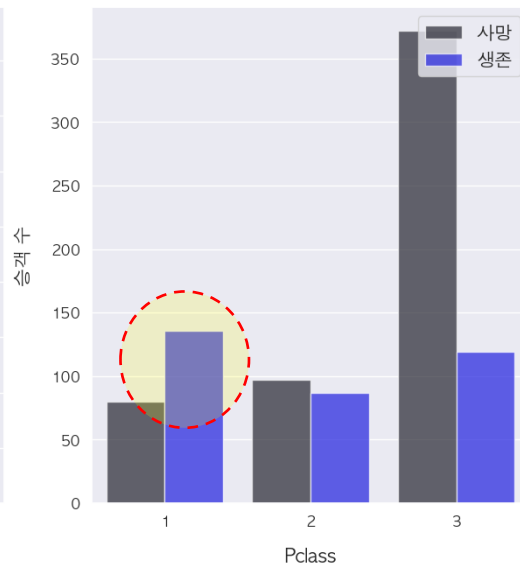
범주형 변수간 생존율 비교

- Embarked, Pclass, Sex, Deck별로 생존율이 차이가 있음 → 해당 변수들이 독립변수로 적합하다는 신호
 - Embarked를 S(사우스햄튼)의 탑승객이 다른 곳에 비해 사망자가 많고, C(세르부르)에서 탑승객이 가장 많이 생존함.
 - Pclass를 보면, 1등실 승객의 생존율이 높음.
 - Deck를 보면, ABC, DE인 승객의 생존율이 높음. (Pclass가 1등실이 다수인 그룹)
 - Sex를 보면, 여성인 승객의 생존율이 높음.

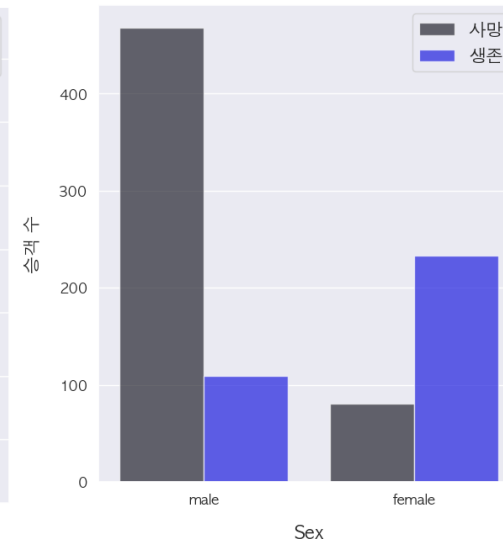
Embarked별 생존자와 사망자 분포



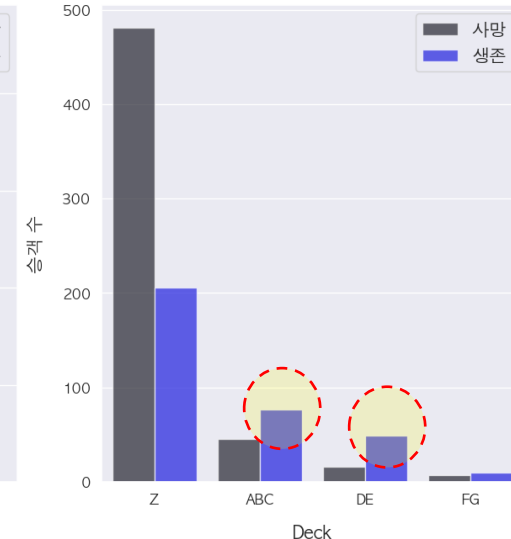
Pclass별 생존자와 사망자 분포



Sex별 생존자와 사망자 분포



Deck별 생존자와 사망자 분포

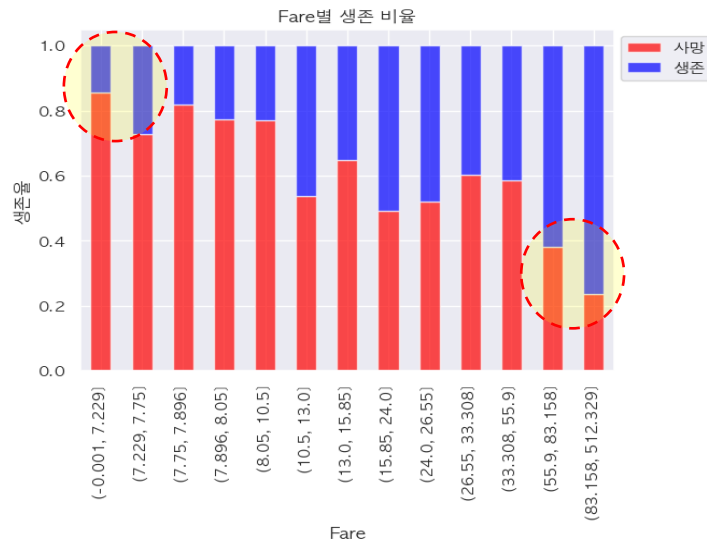


3. 타이타닉 생존예측 모델 개발 - 2) 데이터에 귀를 기울여 보았습니다. (5/9)

피쳐 엔지니어링 - 요금

- 요금의 구간별로 생존율에 차이가 있는 것을 발견함.
 - 연속형인 요금을 범위로 구분해 생존율에 대한 기댓값을 부여함.

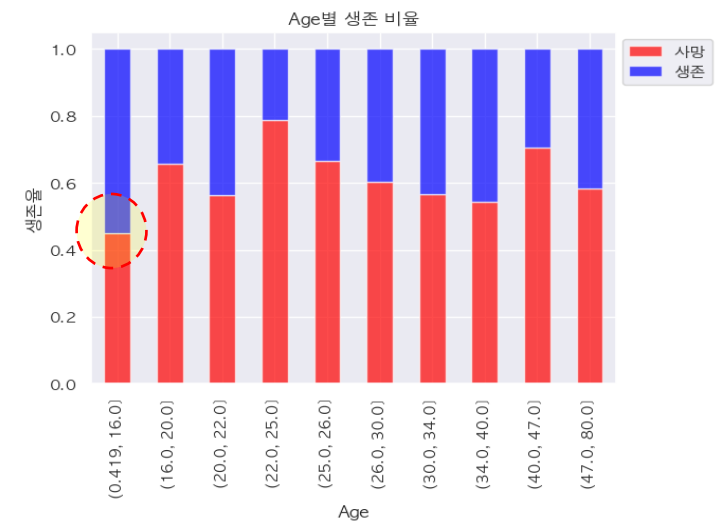
Survived Fare	사망	생존
(-0.001, 7.229]	0.857	0.143
(7.229, 7.75]	0.729	0.271
(7.75, 7.896]	0.819	0.181
...
(26.55, 33.308]	0.603	0.397
(33.308, 55.9]	0.586	0.414
(55.9, 83.158]	0.382	0.618
(83.158, 512.329]	0.235	0.765



피쳐 엔지니어링 - 나이

- 나이의 구간별로 생존율에 차이가 있는 것을 발견함.
 - 요금 변수처럼 구간별로 생존율에 대한 기댓값이 차이나지는 않지만, 16세 이하의 연령의 생존율이 가장 높음.

Survived Age	사망	생존
(0.419, 16.0]	0.45	0.55
(16.0, 20.0]	0.658	0.341
(20.0, 22.0]	0.563	0.436
...
(30.0, 34.0]	0.565	0.434
(34.0, 40.0]	0.543	0.456
(40.0, 47.0]	0.704	0.295
(47.0, 80.0]	0.584	0.415



3. 타이타닉 생존예측 모델 개발 – 2) 데이터에 귀를 기울여 보았습니다. (6/9)

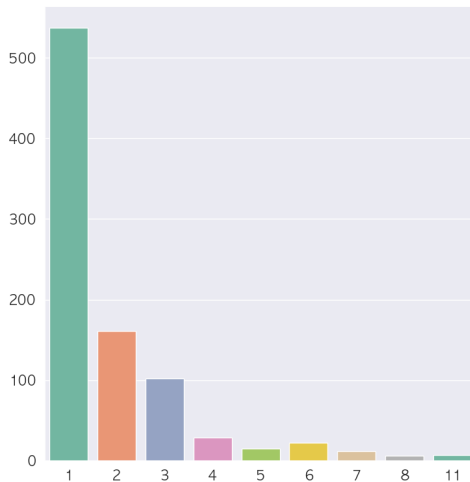
피처 엔지니어링 – 가족 수

- SibSp(형제/자매 수), Parch(부모/자녀 수) 변수를 이용해, Family_size라는 '가족 수' 변수로 합침 → 차원축소

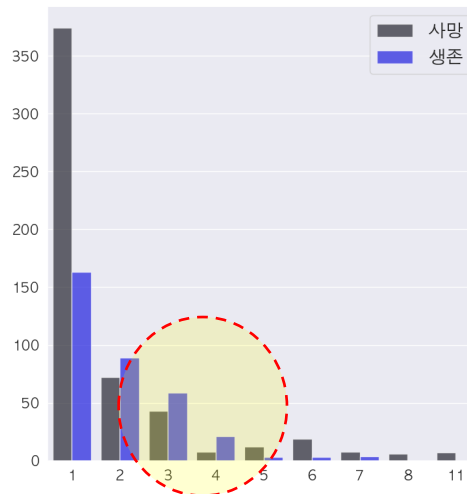
$$Family_Size = SibSp + Parch + 1$$

- Family_Size가 1이면, 혼자 탑승한 승객을 의미함.
- 혼자 탑승한 승객보다 가족 1~3명과 함께 탑승한 승객의 생존율이 높았음.

Family_Size 분포



Family_Size별 생존자 수 비교

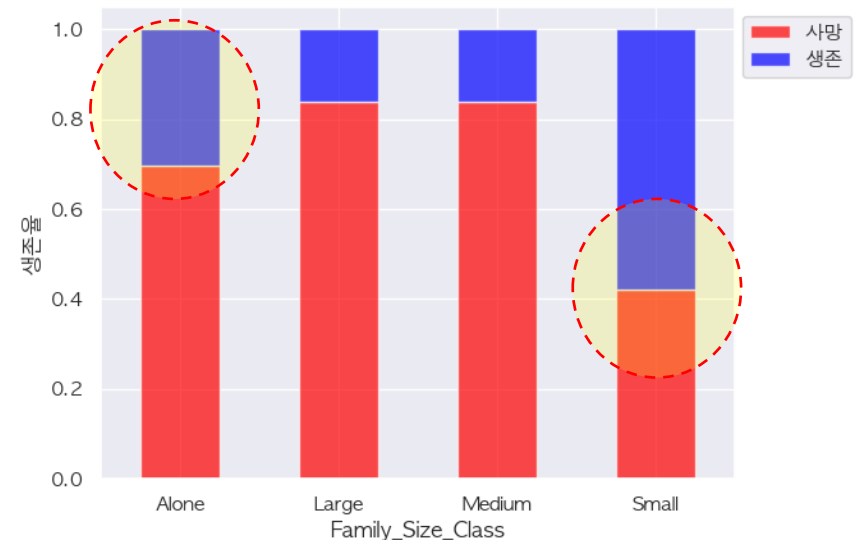


피처 엔지니어링 – Family_Size_Class

- 동등한 가족 수의 구간별로 생존율에 차이가 있는 것을 발견함.
- Family_Size가 1인 'Alone'의 경우 생존율이 가장 낮았고, 'Small'인 그룹의 생존율이 가장 높았으므로, 새로 정의한 Family_Size_Class는 독립변수로 사용해도 무방함.

Family_Size	Label	사망률	생존률
X = 1	Alone	0.696	0.304
X < 5	Small	0.421	0.579
X < 7	Medium	0.838	0.162
X >= 7	Large	0.840	0.160

Family_Size_Class별 생존 비율



3. 타이타닉 생존예측 모델 개발 – 2) 데이터에 귀를 기울여 보았습니다. (7/9)

피쳐 엔지니어링 – Ticket

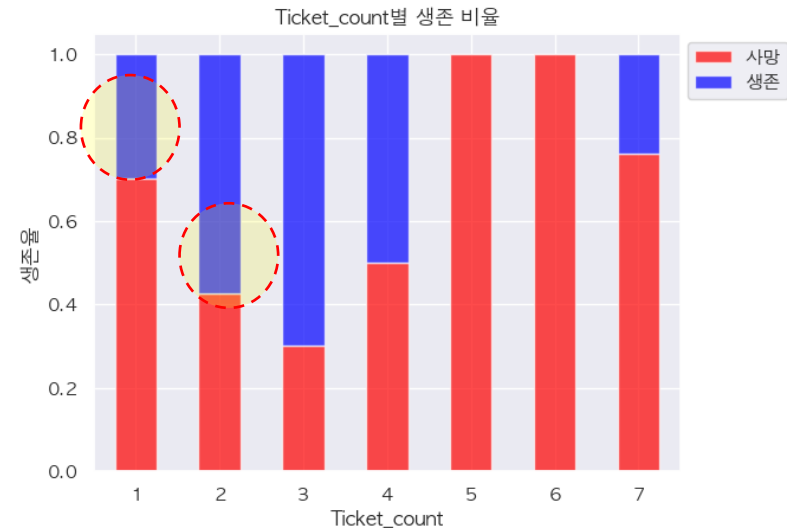
- Ticket은 동일한 번호는 함께 탑승한 승객을 의미함 → 변수 생성
 - Ticket은 고유한 번호를 공유하고 있는 데이터가 많음.
 - 앞서 생성한 Family_Size는 다른 점은, 함께 탑승한 승객 중 가족관계가 아닌, 친구, 부모, 가정부 등이 포함됨.
 - 같은 Ticket 번호를 가지고 있는 승객들끼리 합하여 새로운 변수 'Ticket_count' 변수를 생성

Pclass	Name	Sex	Ticket	Fare	Cabin
1	Maioni, Miss. Roberta	female	110152	12	B79
1	Cherry, Miss. Gladys	female	110152	12	B77
1	Roths, the Countess. of (Lucy Noel Martha Dye...	female	110152	12	B77
1	Taussig, Mr. Emil	male	110413	11	E67
1	Taussig, Mrs. Emil (Tillie Mandelbaum)	female	110413	11	E67
1	Taussig, Miss. Ruth	female	110413	11	E68

Cabin		Ticket_count
110152	→	3
110152		3
110152		3
W/C 14208		1

피쳐 엔지니어링 – Ticket_count

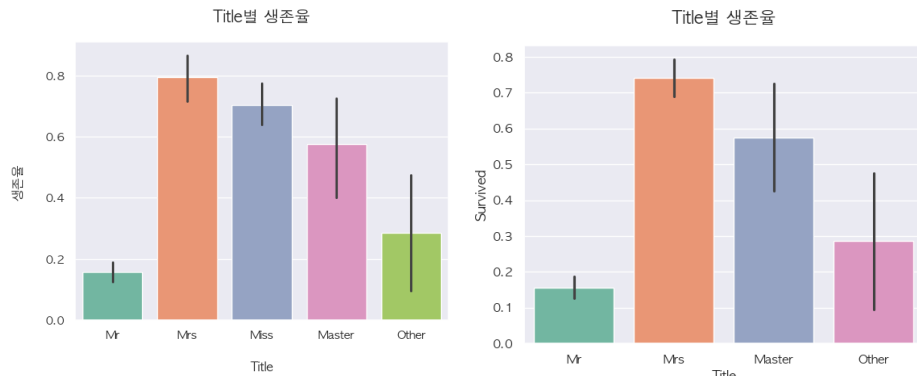
- Ticket_count별 생존율에 차이가 있는 것을 발견함.
 - 혼자 탑승한 그룹인 '1'인 승객들의 생존율은 낮았고, 1~3명씩 함께 탑승한 그룹의 승객들의 생존율은 비교적 높았음.
 - Ticket_count 변수의 패턴은 Family_Size 변수와 유사하지만, 차이가 있음. (동승자 4명 그룹 사망율 100%)
 - 그룹화를 하지 않고 연속형 타입으로 독립변수로 사용



3. 타이타닉 생존예측 모델 개발 – 2) 데이터에 귀를 기울여 보았습니다. (8/9)

피쳐 엔지니어링 – Name

- Name 변수 중 영어 경칭을 의미하는 'Mr.', 'Mrs.', 'Miss' 등 Title를 추출할 수 있음.
 - *Title은 결혼상태 유무, 사회적 지위 등 그 사람에 대한 사회적 상태를 의미함.
 - Title로 결혼 유무에 대한 변수 생성 가능 → **Is_Married** 생성



- 기혼 여성을 의미하는 Mrs. 의 생존율이 가장 높은 것으로 나타남
- 미혼 여성과 기혼 여성을 하나로 합쳐 하나의 클래스(Mrs)로 만들

Mr. : 남자
Mrs. : 기혼 여성
Miss. : 미혼 여성 또는 어린 여성
Master. : 어린 남성
Other : 그 외 기타

독립변수 정의

- 피쳐 엔지니어링을 통해 기존 변수를 활용하여 독립변수를 생성하고 재정의함
 - Name 에서 추출한 'Title', 'Is_Married' 변수 사용
 - SibSp, Parch 에서 새로 정의한 'Family_Size_Class' 변수 사용
 - Cabin에서 새로 정의한 'Deck' 변수 사용

독립변수명	데이터	사용여부	비고
Survived	0	0	종속변수
Pclass	3	O	
Sex	male	O	
Age	(20.0, 22.0]	O	
Fare	(7.229, 7.75]	O	
Embarked	S	O	
Deck	Z	O	
Family_Size_Class	Small	O	
Ticket_count	1	O	
Title	Mr	O	
Is_Married	0	O	
PassengerId	1	X	
Name	Braund, Mr. Owen Harris	X	
SibSp	1	X	
Parch	0	X	
Ticket	A/5 21171	X	
Cabin	Z	X	
Family_Size	2	X	

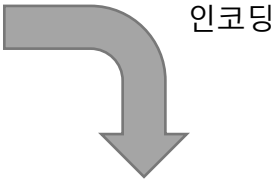
3. 타이타닉 생존예측 모델 개발 – 2) 데이터에 귀를 기울여 보았습니다. (9/9)

범주형 변수 인코딩

- 숫자가 아닌 Class나 범위를 나타내는 범주형 변수를 **One-Hot-Encoding(원핫인코딩)**을 진행해야함.
- OHE(원핫인코딩) : 변수마다 0 또는 1을 만들어, 해당 범주에 속하면 1 아니면 0을 할당함.

독립변수명	데이터	의미	비고
Embarked	S	승선지를 나타내는 범주형 변수	
Sex	male	성별을 나타내는 범주형 변수	
Title	Mr	사회적 위치를 나타내는 범주형 변수	
Pclass	3	등급을 나타내는 범주형 변수	
Deck	Z	갑판을 나타내는 범주형 변수	
Family_Size_Class	Small	가족규모를 나타내는 범주형 변수	
Is_Married	0	0 or 1로만 인코딩 되어있는 상태	인코딩 불필요

Embarked	Sex	Family_Size_Class
S	male	Small
C	female	Large
S	female	Small
S	female	Small



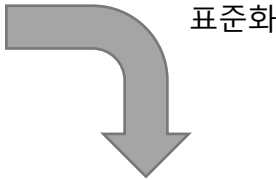
Embarked_C	Embarked_Q	Embarked_S	Sex_female	Sex_male	...	Family_Size_Class_Small
0	0	1	0	1	...	1
1	0	0	1	0	...	1
0	0	1	1	0	...	0
0	0	1	1	0	...	1

연속형 변수 스케일링(Scaling)

- 연속형 변수의 경우 서로 다른 변수의 값 범위를 일정한 수준으로 맞추는 **스케일링(Scaling)**이 필요함.
- 표준화(StandardScaler) : 개별 변수를 평균이 0이고 분산이 1인 값으로 변환하여 모델의 성능을 증가시킬 수 있음.

독립변수명	데이터	의미	비고
Ticket_count	1	동승자 수를 나타내는 연속형 변수	
Age	2	범위를 의미하지만, 양이 증가할 수록 증가하므로 연속형 변수로 선택 가능	
Fare	1	범위를 의미하지만, 양이 증가할 수록 증가하므로 연속형 변수로 선택 가능	

Age	Fare	Ticket_count
2	1	1
7	11	1
4	3	1
7	10	2



Age	Fare	Ticket_count
-0.806821	-1.299855	-0.579162
0.930437	1.346788	-0.579162
-0.111918	-0.770527	-0.579162
0.930437	1.082124	0.155928

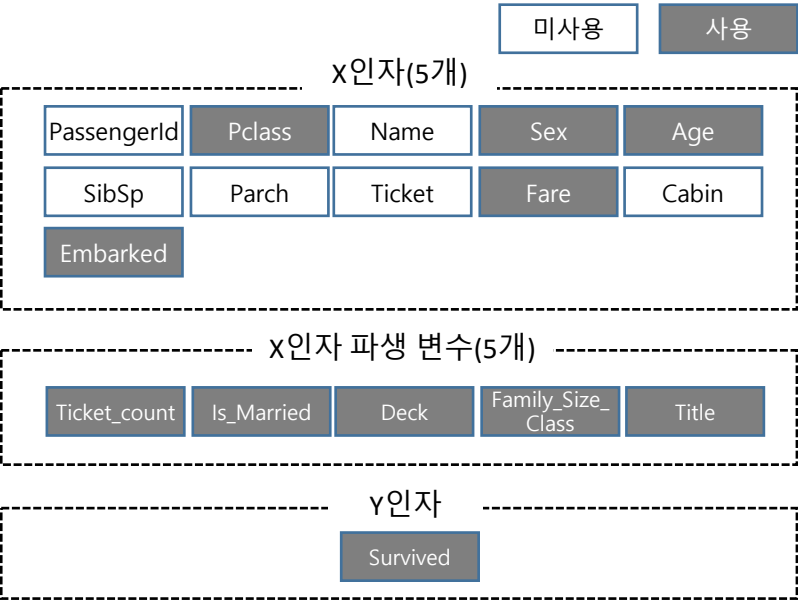
3. 타이타닉 생존예측 모델 개발 – 3) 정제된 데이터로 기계학습을 진행하였습니다. (1/2)

타이타닉 생존예측 모델 개발

- 891개 데이터를 8:2 비율로 나누어 학습하고 검증함

데이터 구분	생존	사망	합계	비율(%)
학습	281	431	712	80
검증	61	118	179	20

- 탐색적 분석과 피쳐 엔지니어링을 통해 독립변수와 새롭게 생성한 파생변수로 기계학습 실행



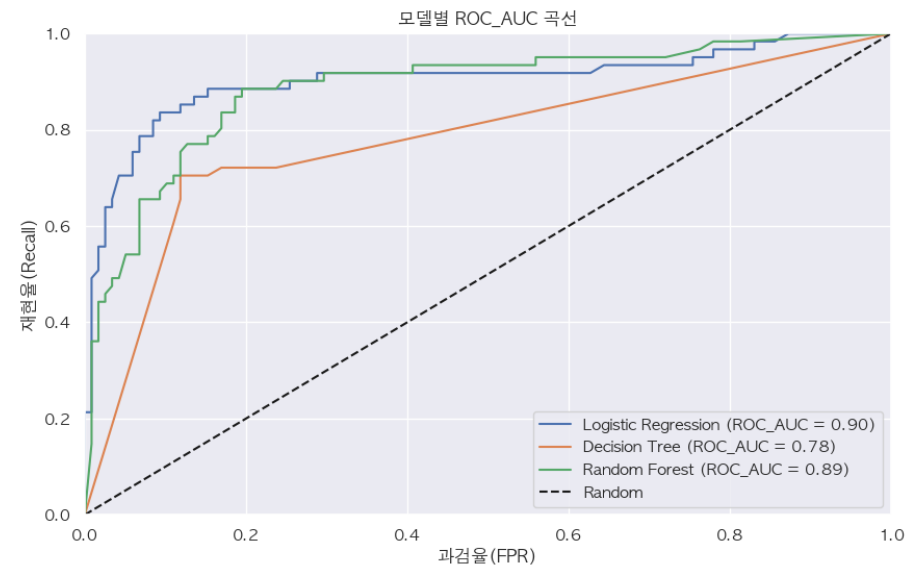
타이타닉 생존예측 모델 학습 및 검증

- 3가지 머신러닝 알고리즘(Logistic Regression, Decision Tree, Random Forest)으로 학습 및 검증

구분		Logistic Regress		Decision Tree		Random Forest	
		사망	생존	사망	생존	사망	생존
실제	사망	108	10	104	14	105	13
	생존	13	48	20	41	18	41
합계		121	58	124	55	123	54

- Logistic Regression의 성능이 가장 우수함 (AUROC ≈ 1 일수록 성능이 좋음)

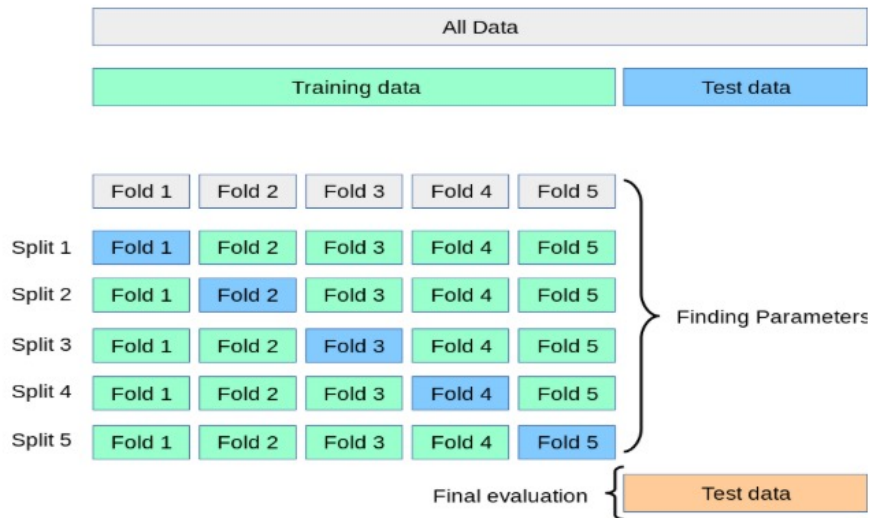
알고리즘	오답수	정확도(%)	미검율(%)	재현율(%)	과검율	AUROC
Logistic Regression	23	0.8715	0.9153	0.7869	0.2131	0.90
Decision Tree	34	0.8101	0.8814	0.6721	0.3279	0.78
Random Forest	31	0.8268	0.8898	0.7049	0.2951	0.89



3. 타이타닉 생존예측 모델 개발 – 3) 정제된 데이터로 기계학습을 진행하였습니다. (2/2)

파라미터 튜닝

- 모델의 파라미터를 튜닝하지 않았으므로 GridSearchCV를 통해 모델의 최적의 파라미터를 찾기.
- GridSearchCV는 교차 검증을 기반으로 모델의 최고 성능을 낼 수 있는 최적의 파라미터 값을 찾아줌.



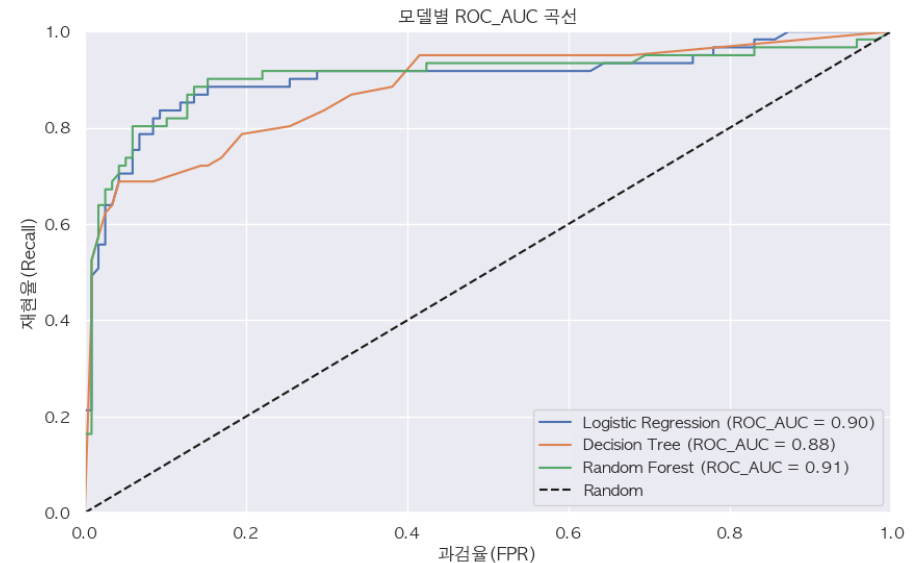
향상된 모델 성능

- 3가지 머신러닝 알고리즘(Logistic Regression, Decision Tree, Random Forest)으로 학습 및 검증

구분		Logistic Regress		Decision Tree		Random Forest	
		사망	생존	사망	생존	사망	생존
실제	사망	108	10	108	10	111	7
	생존	13	48	19	42	14	47
합계		121	58	127	52	125	54

- Random Forest의 성능이 가장 우수함

알고리즘	오답수	정확도(%)	미검율(%)	재현율(%)	과검율	AUROC
Logistic Regression	23	0.8715	0.9153	0.7869	0.2131	0.90
Decision Tree	29	0.8380	0.9153	0.6885	0.3115	0.88
Random Forest	21	0.8827	0.9407	0.7705	0.2295	0.91



3. 타이타닉 생존예측 모델 개발 - 4) 모델의 예측결과를 해석했습니다.

랜덤포레스트의 변수 중요도

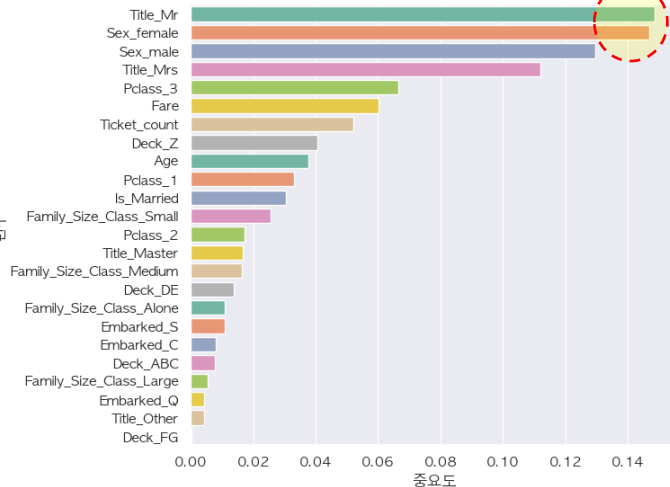
■ 중요도가 클수록 생존과 밀접한 변수를 의미함

- 파생변수에 속한 Title의 Mr. , 성별이 여성이면 생존과 밀접한 연관이 있다고 분석됨.
- 이번 만들어진 모델에선 생존에 중요한 변수는 성별임

Q. 왜 변수 중요도가 전부다 양수인지?

- A. 랜덤포레스트는 여러 개의 의사결정나무들을 조합한 앙상블 모델이므로, 각 트리가 담당하는 역할이 다르고 변수 중요도는 각 변수가 전체 모델의 얼마나 기여를 하는지 나타내고 있음.

Random Forest 변수 중요도



변수	중요도
Title_Mr	0.1486
Sex_female	0.1469
Sex_male	0.1297
Title_Mrs	0.1120
Pclass_3	0.0666
Fare	0.0601
...	...

영화 타이타닉 주인공 데이터로 생존 예측하기

Q. 영화 타이타닉 주인공을 가상의 데이터로 예측해보기

Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
3	Master. Jack Dawson	male	20	0	0	Jack	8	None	S
1	Miss. Rose DeWitt Bukater	female	17	1	0	Rose	247	B58	C



랜덤 포레스트
생존
예측모델

- A. 남자 주인공인 Jack의 경우 생존할 확률이 33.06%, 사망할 확률이 66.94%로 예측되었고,
여자 주인공인 Rose의 경우 생존할 확률이 95.57%, 사망할 확률이 4.43%로 예측됨.

4. 과제 요약 및 앞으로의 방향성 제안

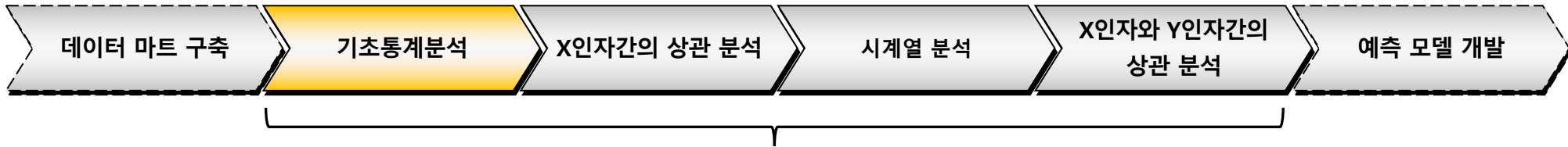
분석 과제 결과 요약

- 데이터 결측치에 대해 상관계수가 높은 변수를 찾고 그룹의 평균값으로 결측치를 보완했습니다.
- 각 변수별 생존율을 비교하며 피처 엔지니어링을 통해 파생변수들을 생성하여 모델에 추가하였습니다.
- 성별이 가장 중요한 변수로 나타났고, 사회적 지위와 나이, 요금 등도 생존에 영향을 끼치는 것으로 나타났습니다.
- 영화 속 주인공의 데이터를 실제 모델에 적용시켜 생존여부를 예측했습니다.

방향성 제안

- 기계학습 입문과정에 필수인 타이타닉 생존 문제로 많은 부분을 학습할 수 있었음.
- 학습 데이터가 891개는 적은 데이터지만, 적은 데이터로 파생변수를 만들어내어 모델의 성능을 향상 시키는 경험을 해볼 수 있었음.
- 모델의 파라미터를 수정하거나, 피처 엔지니어링을 통해 다른 파생변수를 더 생성하여 모델의 성능을 개선시킬 수 있음.
- 개별적으로 학습하고 싶은 다른 모델을 본 과제에 적용시켜 핸들링해볼 수 있는 기회를 얻음.

End Of Documents



데이터 탐색적 분석(Exploratory Data Analysis)

■ 기술통계분석

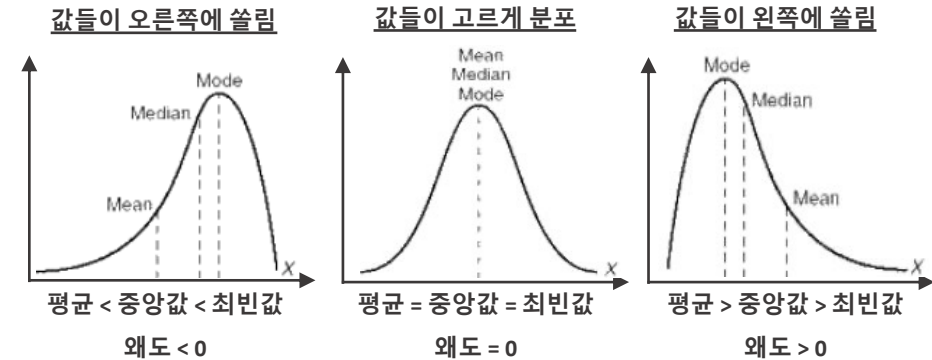
- 방대한 자료를 분석하기에 앞서, 평균 등 전체 데이터 특성을 몇 개의 숫자나 그래프로 정리 및 요약
- 통계적 수치

- ① **개수** : 데이터의 개수는 어떠한가?
- ② **집중화 경향** : 데이터가 어느 위치에 집중되어 있는가?
- ③ **산포도** : 데이터는 어떻게 퍼져 있는가?
- ④ **분포의 형태** : 데이터 분포의 형태와 대칭성은 어떠한가?

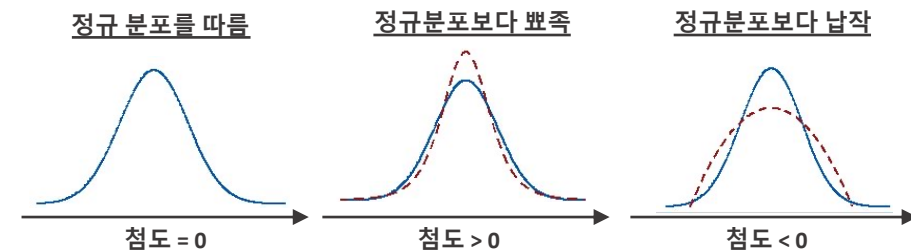
개수	데이터 개수, 유일한 값 개수, 결측치 개수
집중화 경향	평균, 중앙값, 최빈값
산포도	최댓값, 최솟값, 범위, 사분위수, 표준편차
분포의 형태	첨도, 왜도

■ 기초통계분석을 통해 알게 되는 사항

- ① 자료의 개수, 자료에 결측치가 존재하는지 파악 가능
- ② 전체 구간에서 값이 고르게 분포하는지, 쏠려 있는지 파악 가능



- ③ 자료의 분포가 정규 분포와 어떻게 다른지 파악 가능

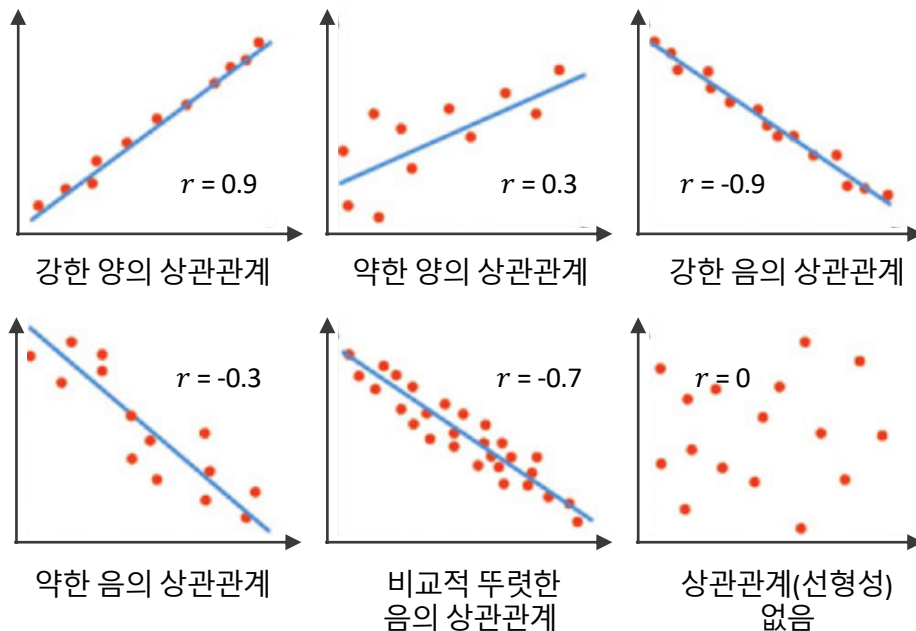




데이터 탐색적 분석(Exploratory Data Analysis)

■ x인자간의 상관 분석

- 양적인 두 변수간의 선형적·비선형적 관계성을 확인함



■ x인자간의 상관 분석을 통해 알게 되는 사항

① 상관계수의 의미와 수치 해석

- 두 변수의 밀접성(선형관계) 강도와 방향을 요약하는 수치
- 대표적인 지표: 피어슨 상관 계수(Pearson's r)

$$r = \frac{X_1 \text{과 } X_2 \text{가 함께 변하는 정도}}{X_1 \text{과 } X_2 \text{가 각각 변하는 정도}}, \quad (-1 \leq r \leq 1)$$

- $|r| = 1$: 완전한 선형관계, $|r| = 0$: 선형관계 없음
- $0.7 < |r| < 1.0$: 강한 상관관계
- $0.3 < |r| \leq 0.7$: 비교적 뚜렷한 상관관계
- $0.1 < |r| \leq 0.3$: 약한 상관관계
- $0 < |r| \leq 0.1$: 거의 무시될 수 있는 상관관계

② 상관분석에 대한 가설 검정

- 귀무가설(H_0): 두 변수간의 상관관계가 없다.
- 대립가설(H_1): 두 변수간의 상관관계가 있다.



데이터 탐색적 분석(Exploratory Data Analysis)

■ 시계열 분석

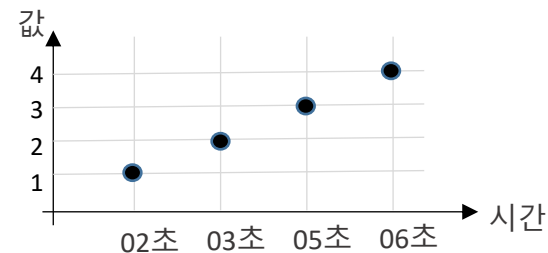
- 각 변수들이 3가지 성질(추세성, 계절성, 순환성)이 있는지 확인

개념	의미	확인 방법
① 추세성	장기적으로 일정하게 변하는 패턴이 존재하는가?	선형 적합 후, 회귀계수가 0에 가까운가?
② 계절성	연도별 반복 패턴이 존재하는가?	육안으로 확인
③ 순환성	주기적으로 발생하는 패턴이 있는가?	육안으로 확인

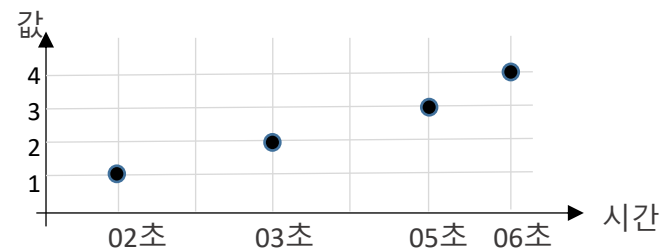
■ 시계열 시각화 방법

시간(시:분:초)	값
00:01:02	1
00:01:03	2
00:01:05	3
00:01:06	4

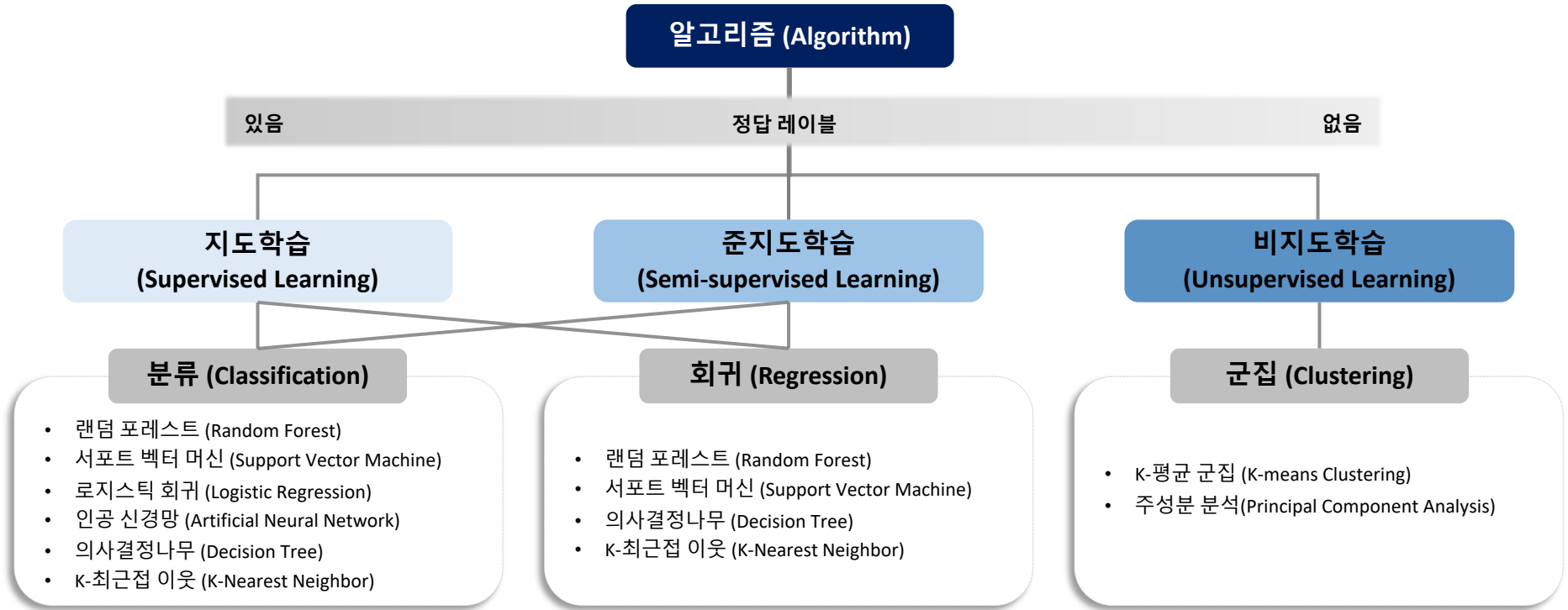
방법1) 시간 간격을 고려하지 않고 순서대로 시각화



방법2) 시간 간격을 고려하여 시각화



■ 알고리즘(Algorithm): 코드로 수행되고 데이터로 학습하는 절차



■ 모델(Model): 데이터로 학습한 머신러닝 알고리즘의 산출물

- 알고리즘에 기반해 실제 연산을 수행한 프로그램
- 양불판정 모델, 잔존수명 예측 모델, 상품 추천 모델 등

→ LS글로벌은 알고리즘을 만들지는 않고, 알고리즘을 이용하여 양불판정 모델을 만듦

■ 룰 기반 vs 기계 학습(Machine Learning)

- 룰 기반: 프로그래밍을 통해 개발자가 룰을 구현
- 기계 학습: 데이터를 기계가 공부(학습)하여 답(프로그램)을 얻는 방식

룰 기반(일반적인 프로그래밍)



기계학습

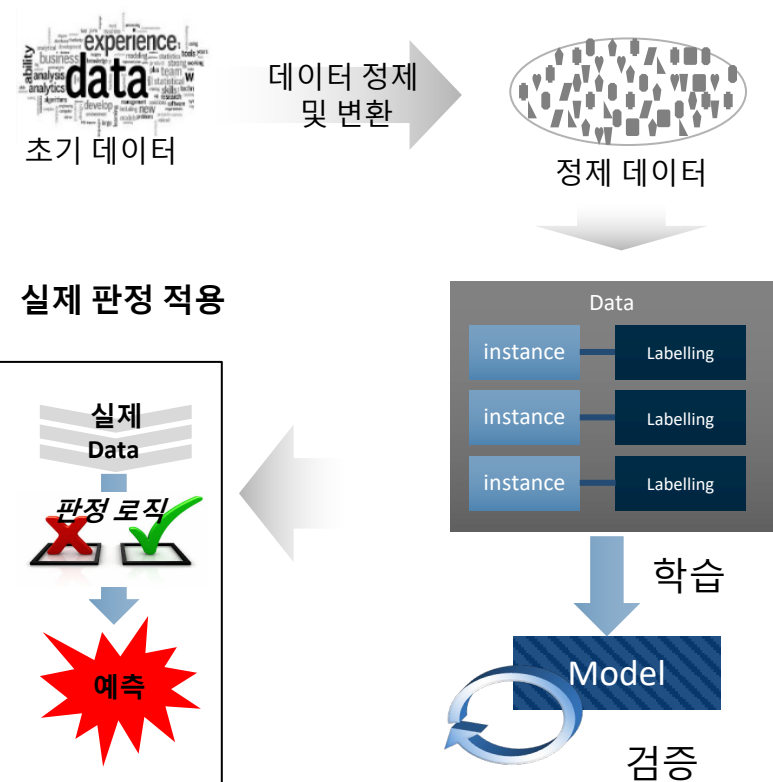


👉 룰, 프로그램으로 정의 내리기 어려운 부분도 구별 가능

ex> 개와 고양이를 구분

■ 기계 학습(Machine Learning) 적용 방법

- 데이터를 정제, 학습 → 판정 로직(모델)이 나옴

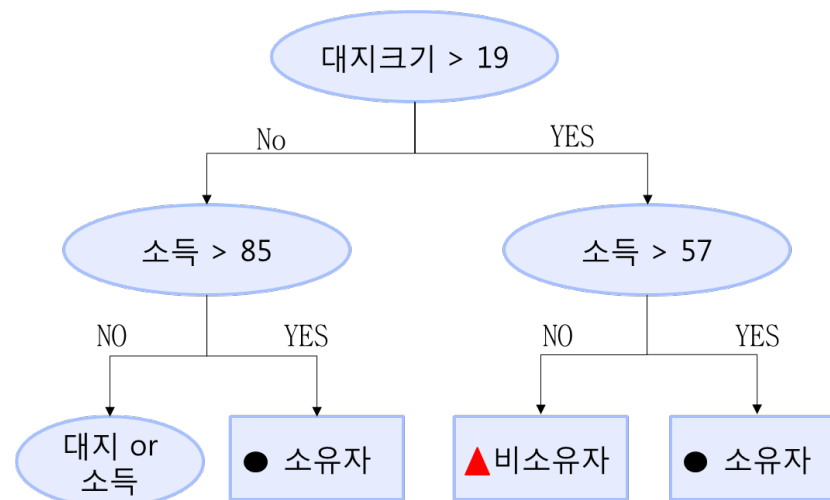
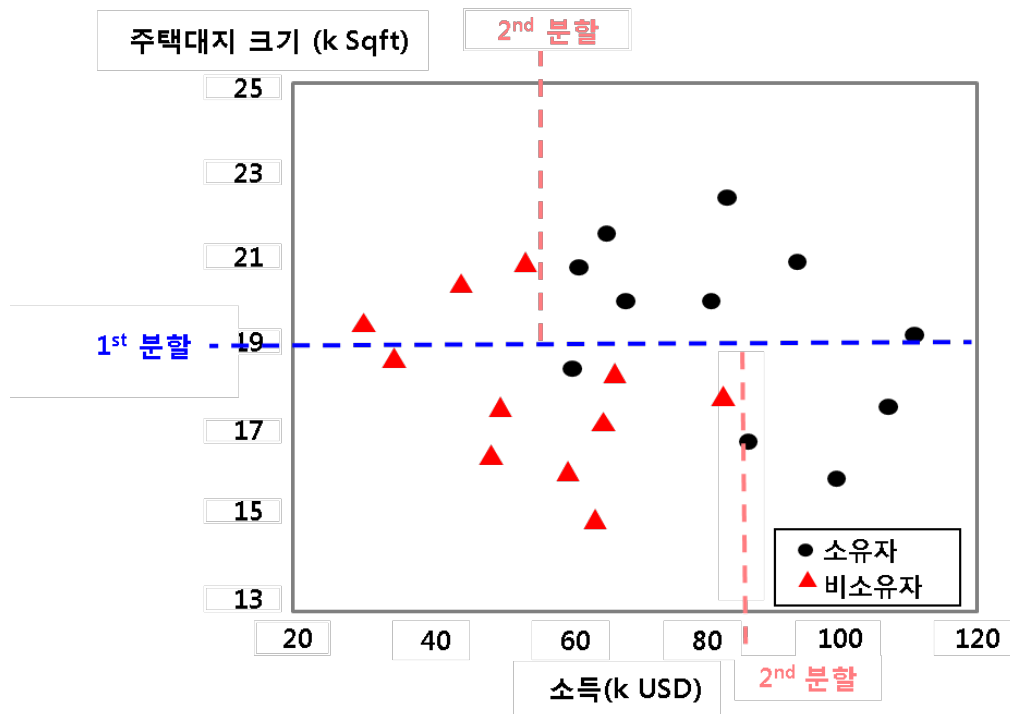


👉 가장 핵심은 대량의 질 좋은 데이터를 어떻게 확보 하느냐가 관건

[참고] 의사결정나무(Decision Tree) 설명을 위한 예시

돌아가기

(교과서에 나오는) 『주택대지크기』와 『소득』 기준 잔디깎기기계 소유여부를 의사결정나무로 분류하는 예시



시작시점 불순도(Impurity)

- 통합불순도 = $1 - \left\{ \left(\frac{11}{22} \right)^2 + \left(\frac{11}{22} \right)^2 \right\} = 50\%$

2nd 분할까지 불순도

- 왼쪽상단 불순도 = $1 - \left(\frac{3}{3} \right)^2 = 0\%$
- 오른쪽상단 불순도 = $1 - \left(\frac{7}{7} \right)^2 = 0\%$
- 왼쪽하단 불순도 = $1 - \left\{ \left(\frac{8}{9} \right)^2 + \left(\frac{1}{9} \right)^2 \right\} = 20\%$
- 오른쪽하단 불순도 = $1 - \left(\frac{3}{3} \right)^2 = 0\%$

☞ 통합 불순도 = $\left(\frac{3}{22} \right) * 0\% + \left(\frac{7}{22} \right) * 0\% + \left(\frac{9}{22} \right) * 20\% + \left(\frac{3}{22} \right) * 0\% = 8.1\%$

(분할前 대비 16%수준)

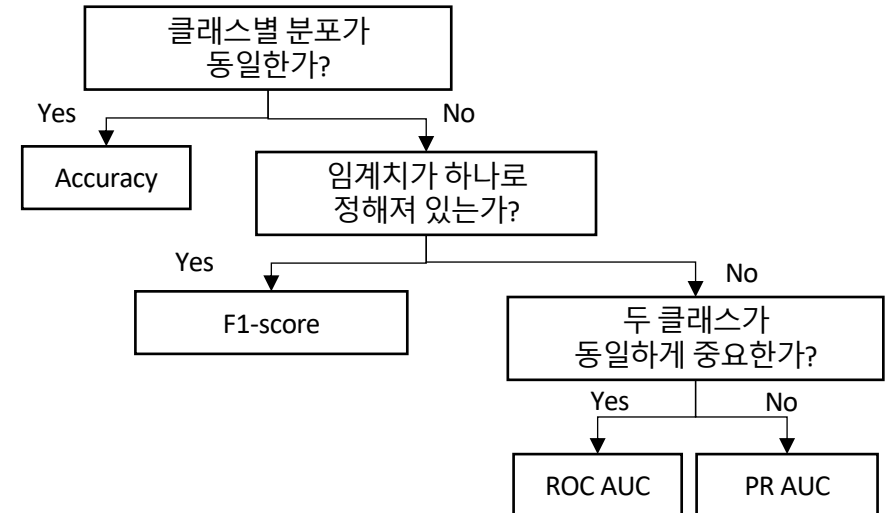
이진분류모델은 오분류표를 작성하고 수치를 계산

		예측		합계
		양품	불량품	
실제	양품	TP	FN	TP + FN
	불량품	FP	TN	FP + TN
합계		TP + FP	FN + TN	TP + TN + FP + FN

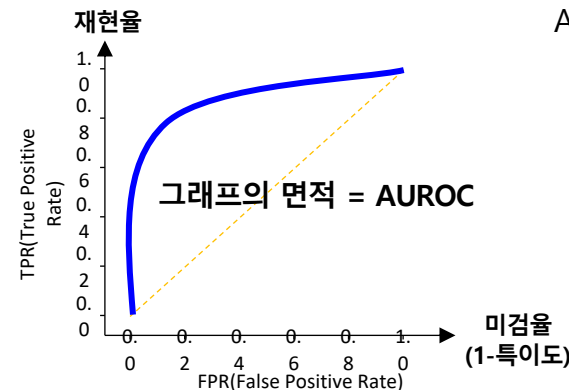
지표	정의	수식
정확도 (Accuracy)	전체 중 양품을 정확히 분류한 비율	$\frac{TP + TN}{TP + TN + FP + FN}$
정밀도 (Precision)	양품이라 예측한 (TP + FP) 개 중 실제 양품 비율	$\frac{TP}{TP + FP}$
재현율 (Recall)	실제 양품 (TP + FN) 개 중 양품으로 예측한 비율	$\frac{TP}{TP + FN}$
미검율 (False Positive Rate)	실제 불량품 (TN + FP) 개 중 불량 FP개를 못 걸러낸 비율 = 1-특이도	$\frac{FP}{TN + FP}$
특이도 (Specificity)	실제 불량품 (TN + FP) 개 중 불량으로 예측한 비율 = 1-미검율	$\frac{TN}{TN + FP}$
과검율	불량이라 예측한 (TN + FN) 개 중 양품 FN개를 잘못 분류한 비율	$\frac{FN}{FN + TN}$
F ₁ - Score	Precision과 Recall의 조화평균	$2 * \frac{Precision * Recall}{Precision + Recall}$

이진분류 모델의 평가지표

- 이진 분류 모델은 어떻게 평가해야 하는가?



- ROC Curve란?



AUROC 수치별 모델 의미

구간	의미
0.5~0.6	Fail
0.6~0.7	Poor
0.7~0.8	Fair
0.8~0.9	Good
0.9~1.0	Excellent

