Titanic

WITH FLEXINK

김동현

IDNEX

01 분석주제

주제 목적

타이타닉에 탑승한 사람들의 신상정보를 활용하여, 승선한 사람들의 생존여부를 예측하는 모델을 생성할

02 데이터 전처리

데이터 구성 및 모델

	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

	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

02 데이터 전처리

데이터 정제

Name 변수에서 이름을 정규표현식으로 추출

이름별로 그룹화

그룹별 나이 중간값 추출

그룹별 중간 값으로 Age 변수의 결측치 처리

02 데이터 전처리
데이터 정제

FamSize 변수 생성

SibSp + Parch 변수 결합

나 자신포함으로 1 더함

02 데이터 전처리
데이터 정제

변수 Embarked 결측치 처리

- Embarked 변수에 최빈값 확인
- 최빈값으로 결측치 처리

변수 Fare 결측치 처리

- Fare 변수에 중간값 확인
- 중간 값으로 결측치 처리

02 데이터 전처리
데이터 정제

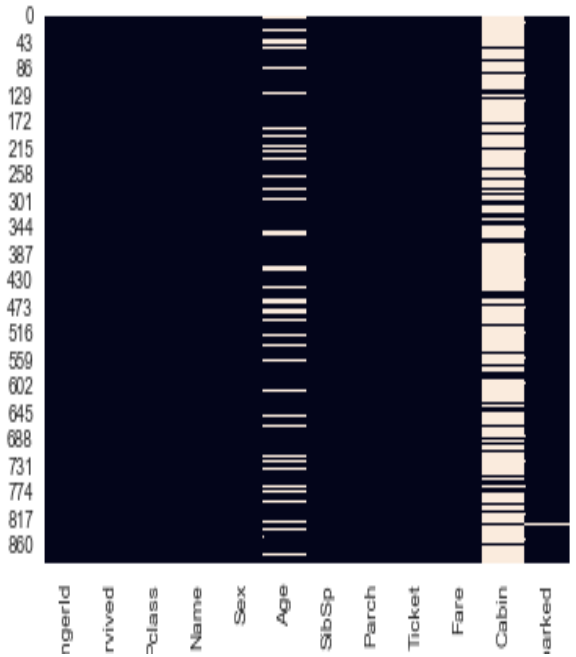
1. Cabin 변수 제거

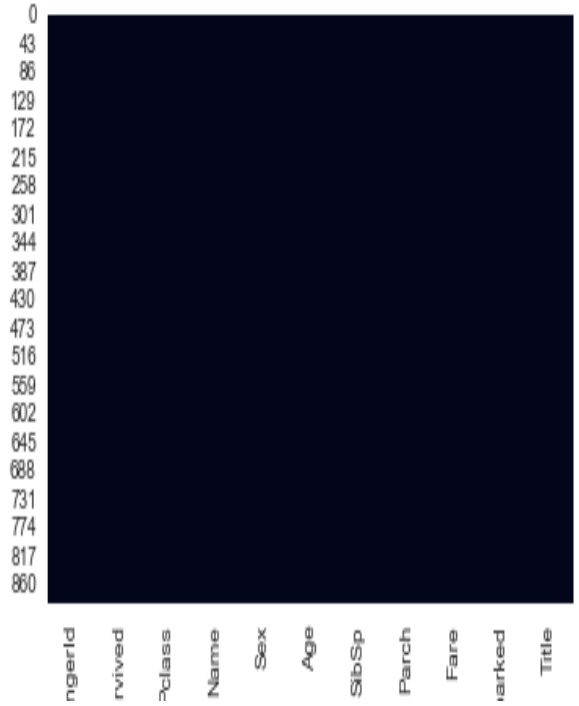
2. Ticket 변수 제거

02 데이터 전처리
데이터 정제

```
Train Data Frame
PassengerId       0
Survived          0
Pclass            0
Name              0
Sex               0
Age             177
SibSp             0
Parch             0
Ticket            0
Fare              0
Cabin           687
Embarked          2
dtype: int64

Test Data Fram
PassengerId       0
Pclass            0
Name              0
Sex               0
Age              86
SibSp             0
Parch             0
Ticket            0
Fare              1
Cabin           327
Embarked          0
dtype: int64
```

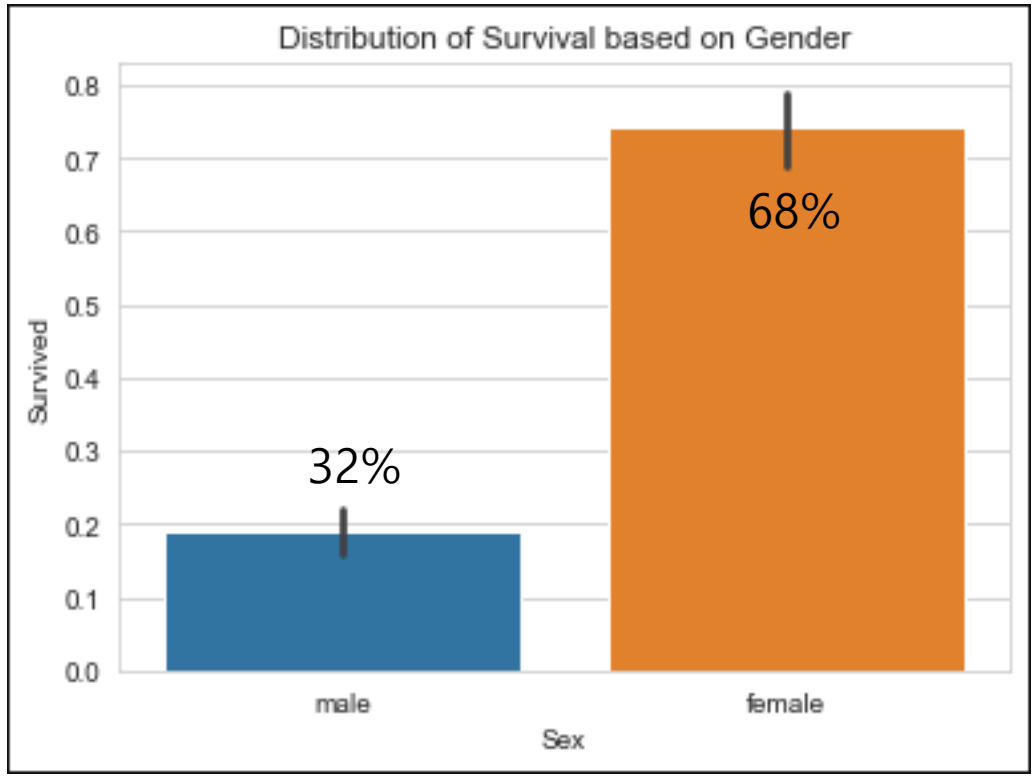
```
Train Data Frame
PassengerId     0
Survived        0
Pclass          0
Name            0
Sex             0
Age             0
SibSp           0
Parch           0
Fare            0
Embarked        0
Title           0
dtype: int64

Test Data Fram
PassengerId     0
Pclass          0
Name            0
Sex             0
Age             0
SibSp           0
Parch           0
Fare            0
Embarked        0
Title           0
dtype: int64
```

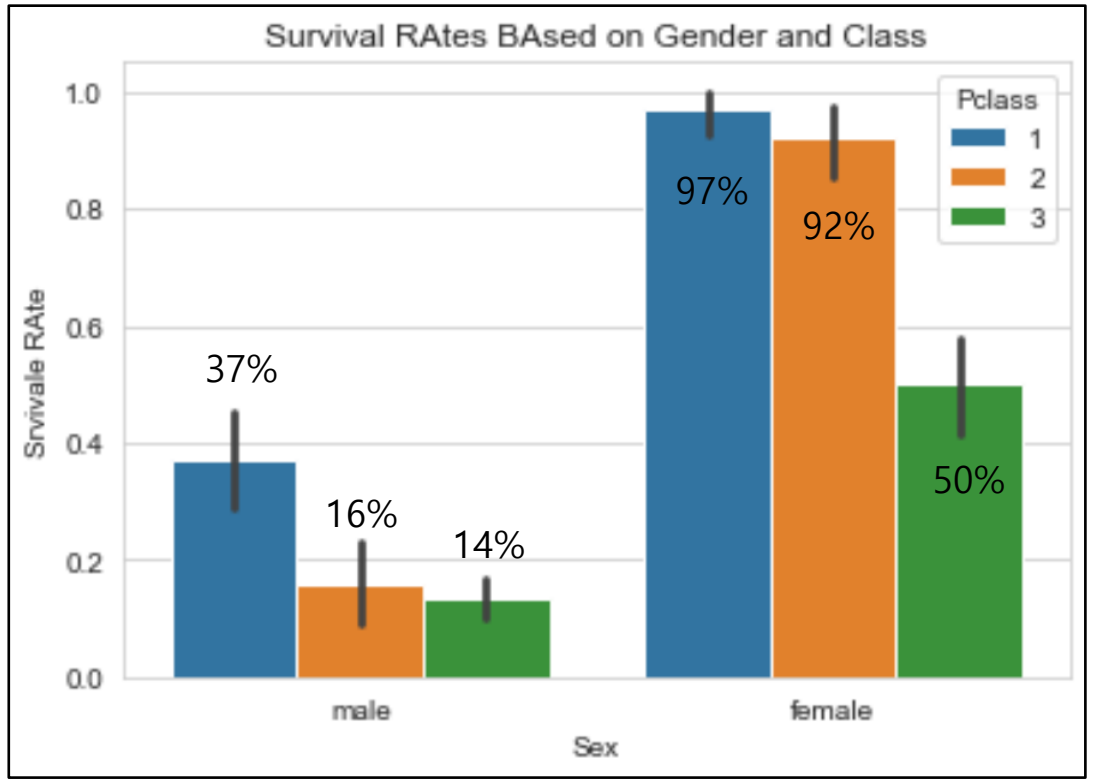




03 분석 및 결과

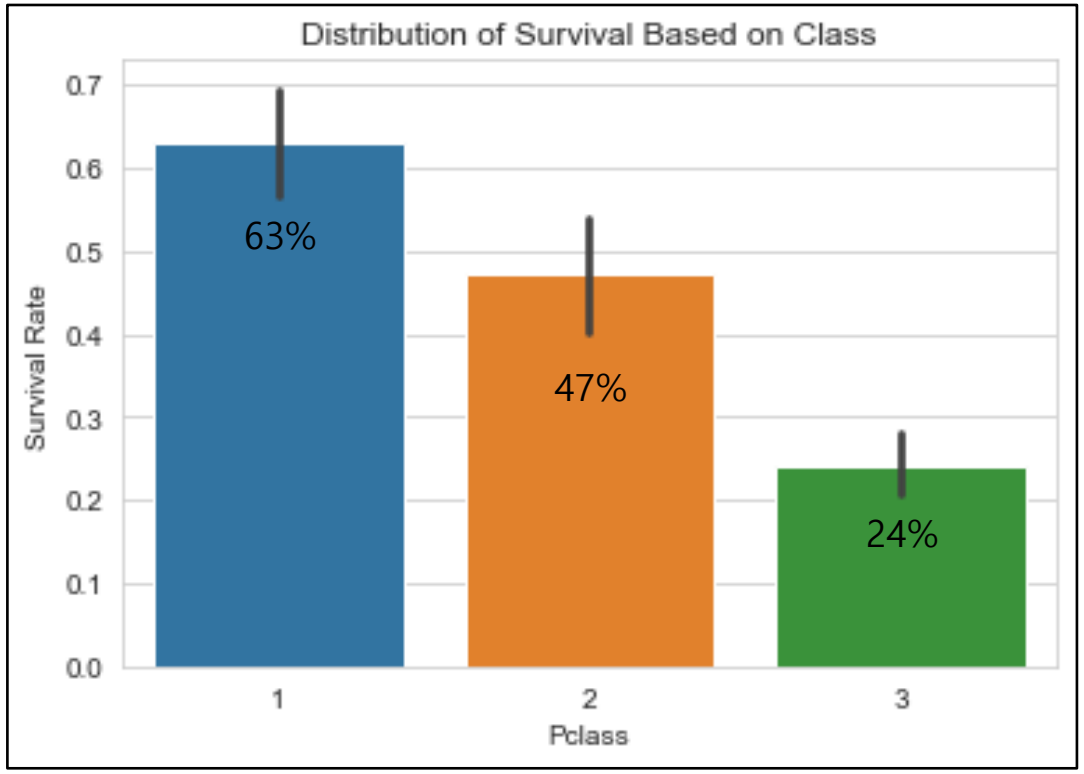
분석 모델링 및 성능 비교

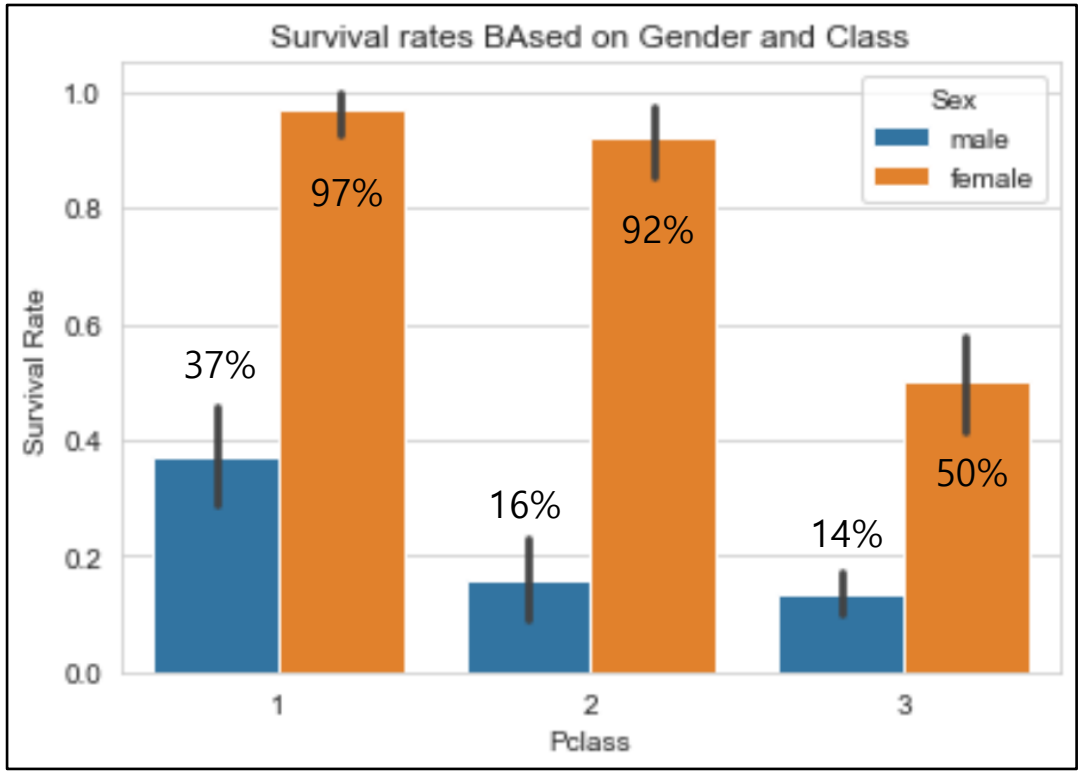




03 분석 및 결과

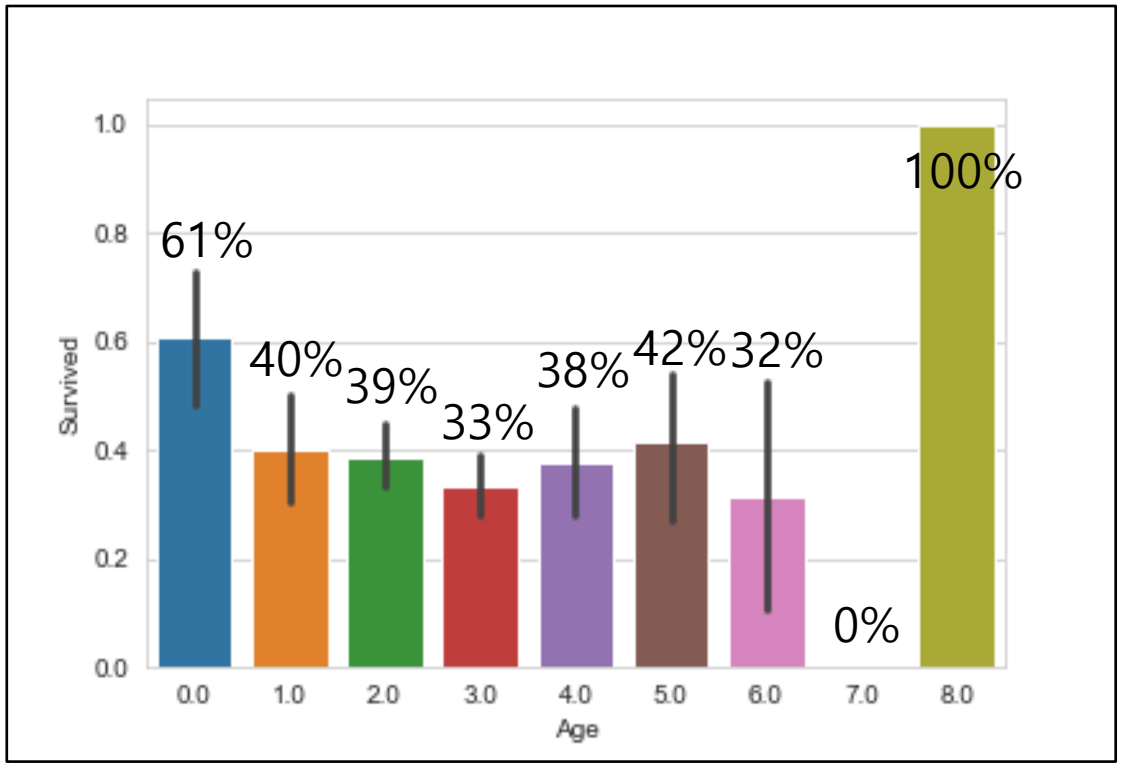
분석 모델링 및 성능 비교

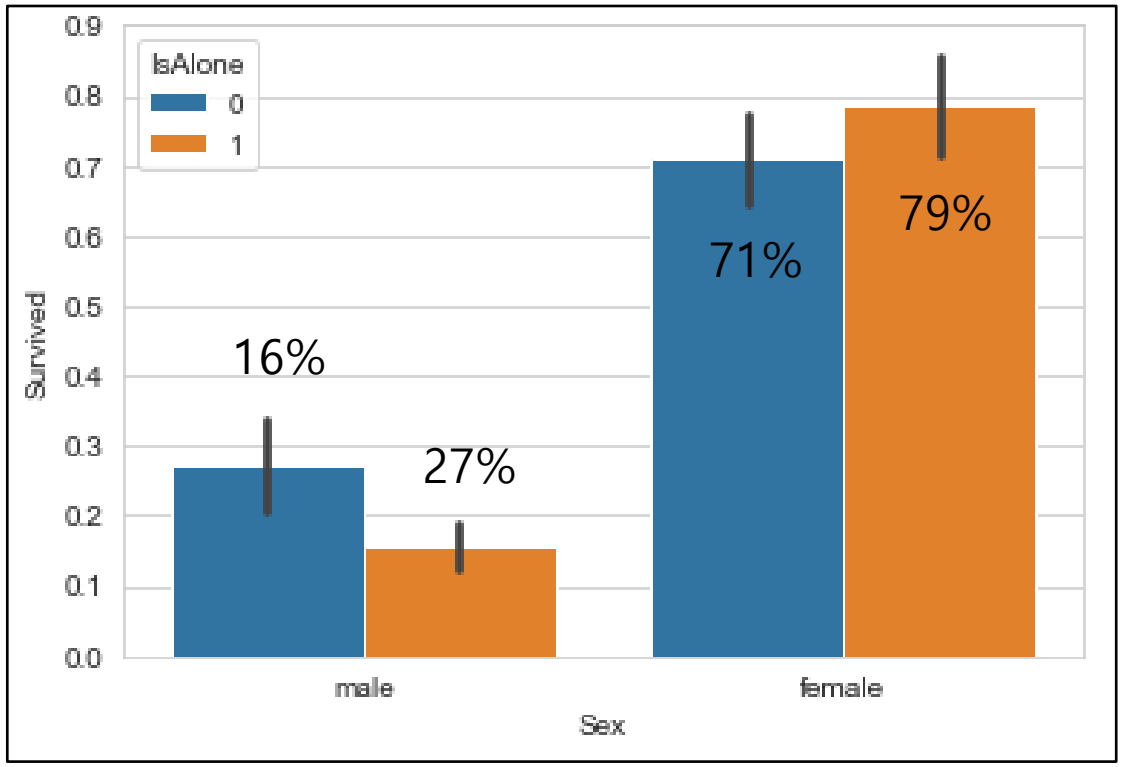




03 분석 및 결과

분석 모델링 및
성능 비교





03 분석 및 결과

분석 모델링 및 성능 비교

Model	Accuracy
Random Forest	0.832402
SVC	0.815642
Gaussian Naïve Bayes	0.805692
Linear SVC	0.798507
Logistic Regresion	0.798507
Decision Tree	0.779851
K Nearest Neighbors	0.764925

03 분석 및 결과

분석 모델링 및 성능 비교

Grid SearchCV RF	Best_estimator_
criterion	gini, entropy
Max_depth	2, 3, 5, 10
Max_features	auto, sqrt, log2
N_estimators	4, 5, 6, 7, 8, 9, 10, 15
Min_samples_split	2, 3, 5, 10
Min_samples_leaf	1, 5, 8, 10

Grid SearchCV SVC	Best_estimator_
C	0.01, 0.1, 1, 10, 100, 200, 1000
Gamma	0.01, 0.1, 1, 10, 100, 200, 1000

03 분석 및 결과

분석 결과

Model	Accuracy
Random Forest	0.838365
SVC	0.832753

Kaggle Submission	Public Score
SVC	0.79904
Random Forest	0.78468

04 결과

- Random Forest 와 SVC 중 Kaggle 점수로는 SVC가 조금 더 높다.
- 예측점수는 Random Forest가 조금 더 높다.
- 티켓 가격과 나이를 범주화해주었더니 더 점수가 잘 나왔다.
- 엑셀로 했을 때와 다른 점 변수 추가.

END