

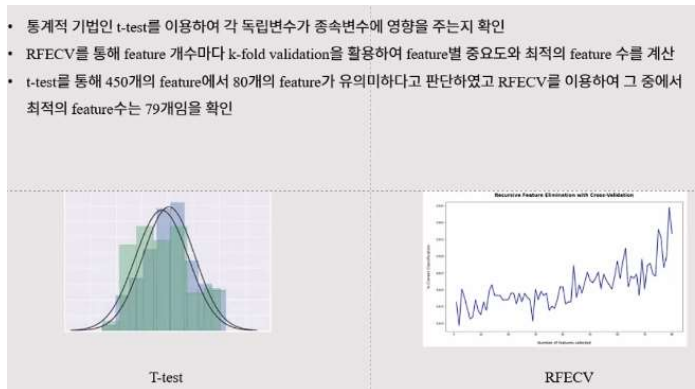
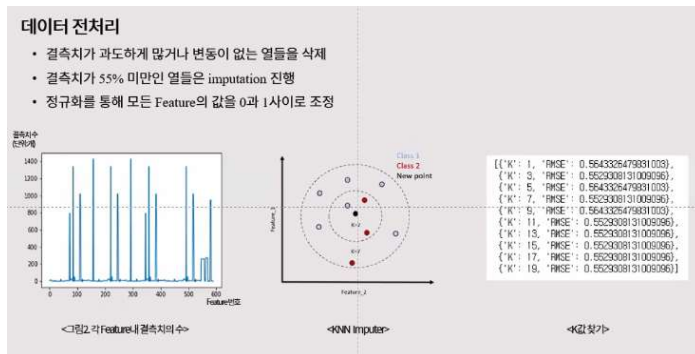
# 빅데이터 애널리틱스

## 발표 3 - 분석

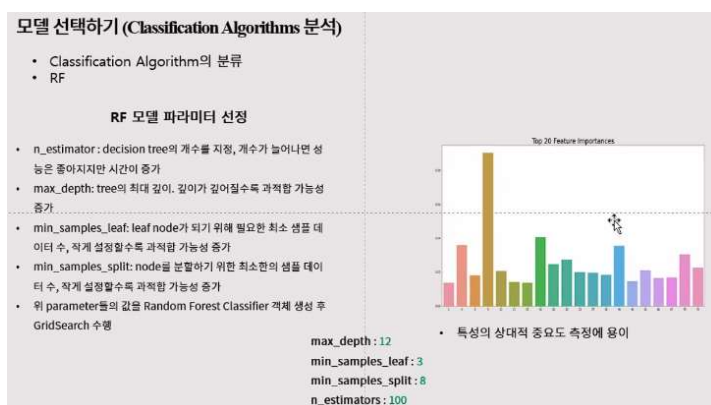
제 주관으로 선정한 상위 4개팀을 비교 분석했습니다.

# 1조

## 장점



이와 같이 적절한 시각화 자료를 첨부함으로써 청자의 이해를 쉽게 하였습니다.

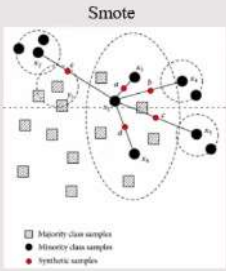


모델링 과정에 있어 모델의 파라미터 정보를 전달하고 있다는 점, 모델 분류와 각 모델에 대한 성능 지표를 시각화를 통해 효과적으로 전달한 점을 장점으로 생각 합니다.

## 추가 의견

### Over Sampling

- 불균형 문제를 해결하기 위하여 smote 알고리즘을 활용하여 oversampling 진행
- Smote 결과 fail 데이터가 1463개 증가하여 총 2926개 행 데이터로 모델링 수행 예정



- smote 기법은 소수 클래스에서 임의의 데이터를 선택하여 시작한 다음 데이터에서 knn을 설정한 후 무작위 데이터와 선택된 knn 사이에 합성 데이터를 만들어 overfitting가능성의 감소

여기서 SMOTE 기법에 대해 간략하게 (윗 문단)으로 바꾸고,

SMOTE 알고리즘을 진행한 결과를 메인 문단으로 가져 왔으면 합니다.

또한 PPT의 디테일이라 할 수 있는 가독성과 텍스트의 직관적 전달이 부족한 것 같습니다.

## 총평

전반적으로 잘 만들어진 발표.

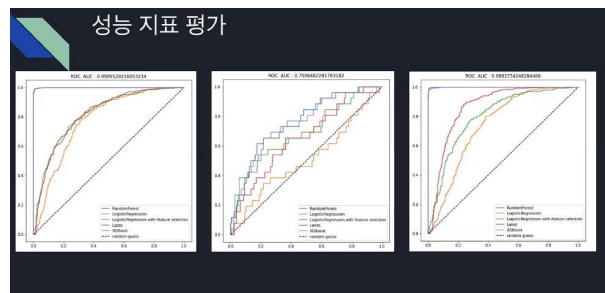
## 2조

### 장점

#### 계획

- 전처리
  - 결측치 처리 - KNN
  - 오버 샘플링 - SMOTE
- 모델링
  - 모델 선정 - GBA, Logistic Regression, Random Forest 중 택1
  - 모델의 성능 지표 비교 - F1 score, AUC

발표 초반에 계획을 언급함으로써 구현 과정에 있던 과정을 표현함.



성능 지표에 있어 그래프를 활용한 점은 장점이나, 약간의 보정을 통해 비교우위를 표현했으면 좋겠음.

### 추가 의견

#### 전처리

- Feature Selection
  - Feature수가 아주 많기 때문에 filter method를 사용함.  
그 중 ANOVA (Analysis of Variance)를 선택.
  - Random Forest, Logistic Regression, Lasso, XGBoost 모델 중 Logistic을 제외한 나머지 모델은 따로 규제가 있기 때문에 Logistic만 Feature Selection을 진행함

이와 같이 데이터 시각화 이미지의 부족으로 인해 전달력이 매우 떨어짐.

또한 Feature Selection 이라는 중요한 파트를 전처리에 묶어 간단히 발표한 점.

### 총평

표현력이 부족하지만 결과 (성능) 에 대한 지표만큼은 표현함.

## 장점

결측치 50% 이상  
feature 삭제

[illegible]

```
[ ] data.isnull().sum(),sum()

2605

[ ] dfnull=data.isnull().sum()/data.shape[0]
dfnull

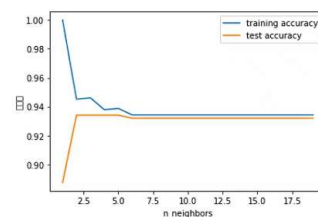
0      0.00931
1      0.00470
2      0.00940
3      0.00940
4      0.00940
...
585      0.000639
586      0.000639
587      0.000639
588      0.000639
589      0.000639
Length: 446, dtype: float64
```

The figure consists of two vertically stacked density plots. The top plot shows the density of the number of iterations  $T$ . The x-axis is labeled  $T$  and ranges from 2000 to 2800 with major ticks every 200 units. The y-axis is labeled 'Density' and ranges from 0.000 to 0.007 with major ticks every 0.001 units. The plot shows a bell-shaped curve centered at approximately 2500, with a peak density of about 0.0065. The bottom plot shows the density of the number of iterations  $M$ . The x-axis is labeled  $M$  and ranges from 0 to 120 with major ticks every 20 units. The y-axis is labeled 'Density' and ranges from 0.000 to 0.040 with major ticks every 0.010 units. The plot shows a very sharp peak at  $M=0$  with a density of approximately 0.045, and a much smaller peak at  $M=10$  with a density of about 0.010. The rest of the distribution is near zero.

SMOTE \_ ADASYN  
(Adaptive Synthetic Sampling Approach)

데이터 전 처리 과정에서 처리 과정을 한 단계씩 시각화 자료와 함께 제공함.

	model	time	
0	Ridge(alpha=0.0001)	223.707709	
1	Ridge(alpha=0.001)	39.433713	
2	Ridge(alpha=0.01)	38.130774	
3	Ridge(alpha=0.05)	37.760295	
4	Ridge(alpha=0.1)	38.61124	
5	Ridge(alpha=0.5)	37.680874	
6	Ridge(alpha=100)	1.525314	
7	Elastic(alpha=0.0001)	0.684719	k: 2, score is 0.945842
8	Elastic(alpha=0.001)	0.684010	k: 3, score is 0.947171
9	Elastic(alpha=0.01)	0.684719	k: 4, score is 0.938075
10	Elastic(alpha=0.05)	0.670799	k: 5, score is 0.944013
11	Elastic(alpha=0.1)	0.673380	k: 6, score is 0.938114
12	Elastic(alpha=0.5)	0.673380	k: 7, score is 0.938114
13	Elastic(alpha=100)	0.673380	k: 8, score is 0.938114
14	Elastic(alpha=0.0001)	0.693952	k: 9, score is 0.938114
15	Elastic(alpha=0.001)	0.693952	k: 10, score is 0.938114
16	Elastic(alpha=0.01)	0.693952	k: 11, score is 0.938114
			k: 12, score is 0.938114
			k: 13, score is 0.938114
			k: 14, score is 0.938114
			k: 15, score is 0.938114
			k: 16, score is 0.938114
			k: 17, score is 0.938114
			k: 18, score is 0.938114
			k: 19, score is 0.938114
			k: 20, score is 0.938114
			k: 21, score is 0.938114
			k: 22, score is 0.938114
			k: 23, score is 0.938114
			k: 24, score is 0.938114
			k: 25, score is 0.938114
			k: 26, score is 0.938114
			k: 27, score is 0.938114
			k: 28, score is 0.938114
			k: 29, score is 0.938114
			k: 30, score is 0.938114
			k: 31, score is 0.938114
			k: 32, score is 0.938114
			k: 33, score is 0.938114
			k: 34, score is 0.938114
			k: 35, score is 0.938114
			k: 36, score is 0.938114
			k: 37, score is 0.938114
			k: 38, score is 0.938114
			k: 39, score is 0.938114
			k: 40, score is 0.938114
			k: 41, score is 0.938114
			k: 42, score is 0.938114
			k: 43, score is 0.938114
			k: 44, score is 0.938114
			k: 45, score is 0.938114
			k: 46, score is 0.938114
			k: 47, score is 0.938114
			k: 48, score is 0.938114
			k: 49, score is 0.938114
			k: 50, score is 0.938114
			k: 51, score is 0.938114
			k: 52, score is 0.938114
			k: 53, score is 0.938114
			k: 54, score is 0.938114
			k: 55, score is 0.938114
			k: 56, score is 0.938114
			k: 57, score is 0.938114
			k: 58, score is 0.938114
			k: 59, score is 0.938114
			k: 60, score is 0.938114
			k: 61, score is 0.938114
			k: 62, score is 0.938114
			k: 63, score is 0.938114
			k: 64, score is 0.938114
			k: 65, score is 0.938114
			k: 66, score is 0.938114
			k: 67, score is 0.938114
			k: 68, score is 0.938114
			k: 69, score is 0.938114
			k: 70, score is 0.938114
			k: 71, score is 0.938114
			k: 72, score is 0.938114
			k: 73, score is 0.938114
			k: 74, score is 0.938114
			k: 75, score is 0.938114
			k: 76, score is 0.938114
			k: 77, score is 0.938114
			k: 78, score is 0.938114
			k: 79, score is 0.938114
			k: 80, score is 0.938114
			k: 81, score is 0.938114



종합적인 지표를 제공함. 그러나 직관성이 떨어지는 점은 아쉬움.

## 추가 의견

스케일링을 Robust Scaler로 한 근거가 부족함.

상관계수 VIF 에 대한 근거가 부족함

## 총평

전반적으로 좋았고, 적절한 시각화 자료, 구현한 결과물은 좋으나,

구현 과정의 기술적 근거가 부족한 점이 매우 아쉬움.

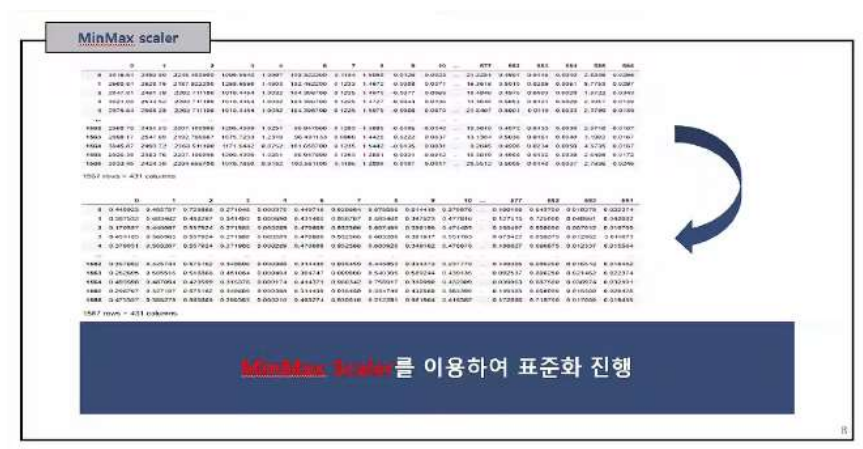
*If you would like an English or Chinese translation of this document, please leave an email in the comments.*

## 5조

## 장점

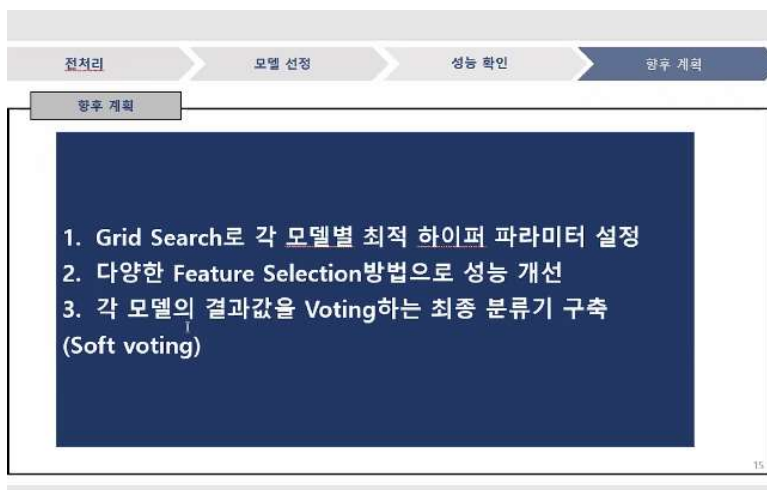


이러한 방향성 제시는 좋으나, 계속 차지하고 있으면 불편할 수 있음.



각 과정에 대해 결과물을 명시하고, bulk한 내용이 아니면 시각화 자료를 제공하는 점은 장점으로 생각함.

## 추가 의견



발표 자료의 퀄리티가...

총평

이분들은 공대가 분명하다.

## 6조

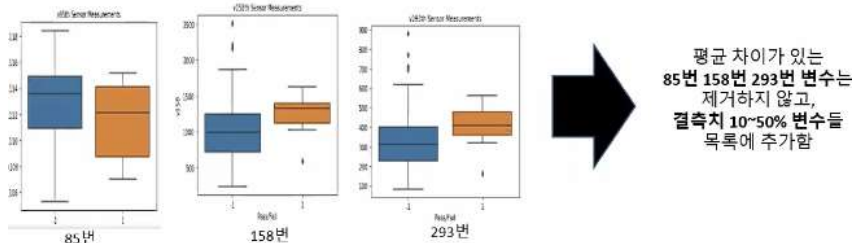
### 장점

#### 결측치 처리

##### 1. 결측치 처리 방향

-> 관측 데이터의 분포를 최대한 유지하는 방향으로 결측치를 처리해야 함

1) 결측치가 50% 이상인 변수 : **box-plot** 활용하여 평균차이가 거의 없는 변수 제거

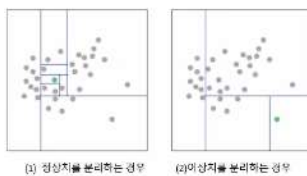
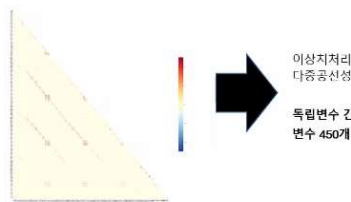


한 페이지에 정보 전달을 과하게 하지 않고, 데이터 시각화 자료를 토대로 결과를 제시함.

#### Feature Selection : Filter

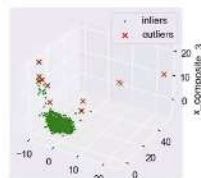
[Correlation Filter Methods]

다중공선성 방지를 위해 독립변수 간 상관관계가 높은 feature를 지



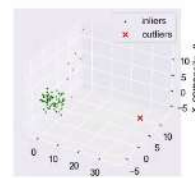
(1) 정상치를 분리하는 경우

(2) 이상치를 분리하는 경우



Pass data 이상치 탐지결과

→ 이상치 row 14개 제거



Fail data 이상치 탐지결과

→ 이상치 row 1개 제거

[Isolation Forest]

각 관측치를 고립(=분리)시키는 것은  
이상치가 정상 데이터보다 쉽다.

데이터 시각화를 효과적으로 제시함. 4산경의 힘인가 이게.

### 추가 의견

시각화 자료가 과한 부분이 있으나, 조금만 적당히 분배하면 나을 듯.

### 총평

산경공 \* 4 = 이야...

끝.