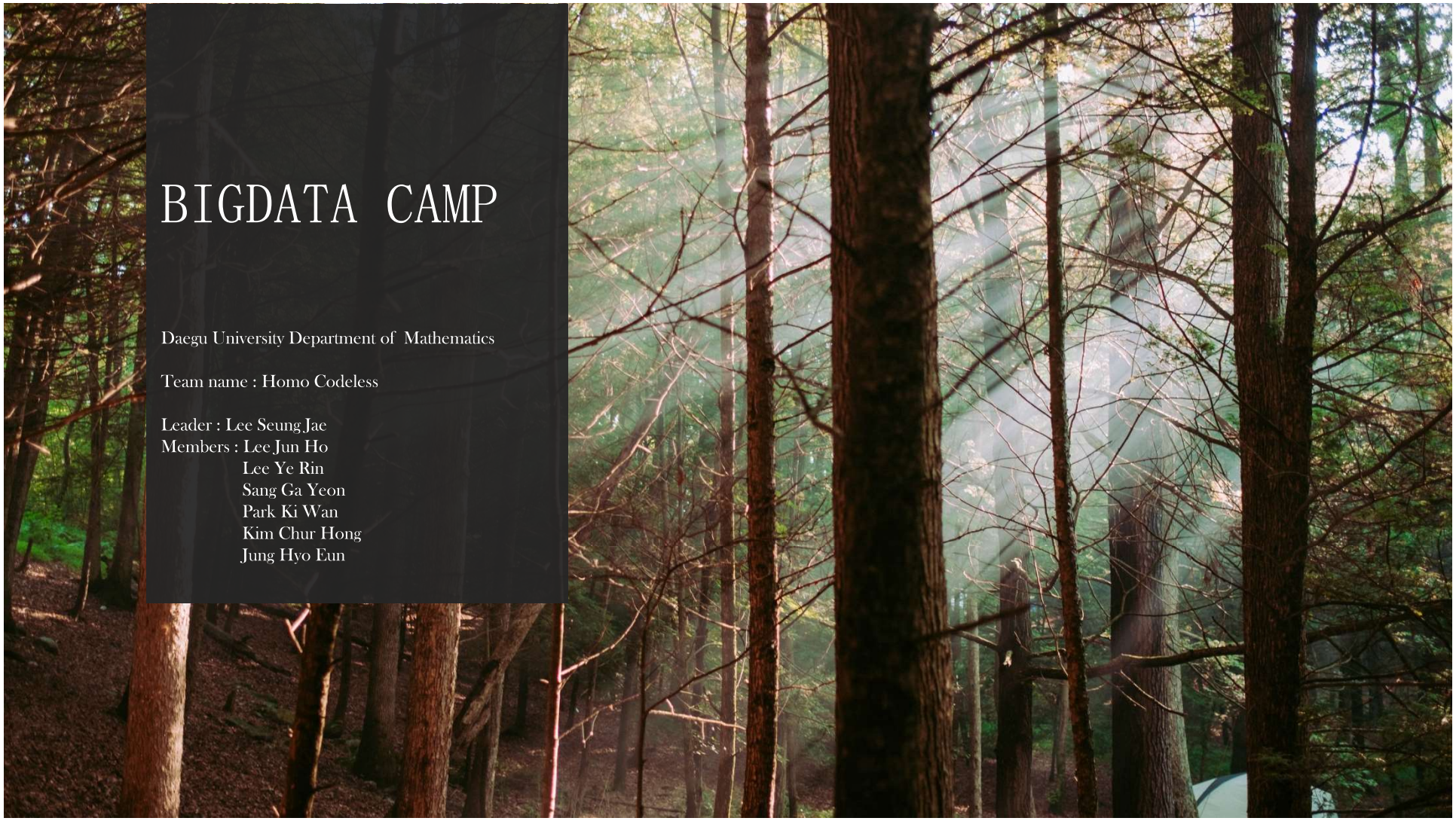
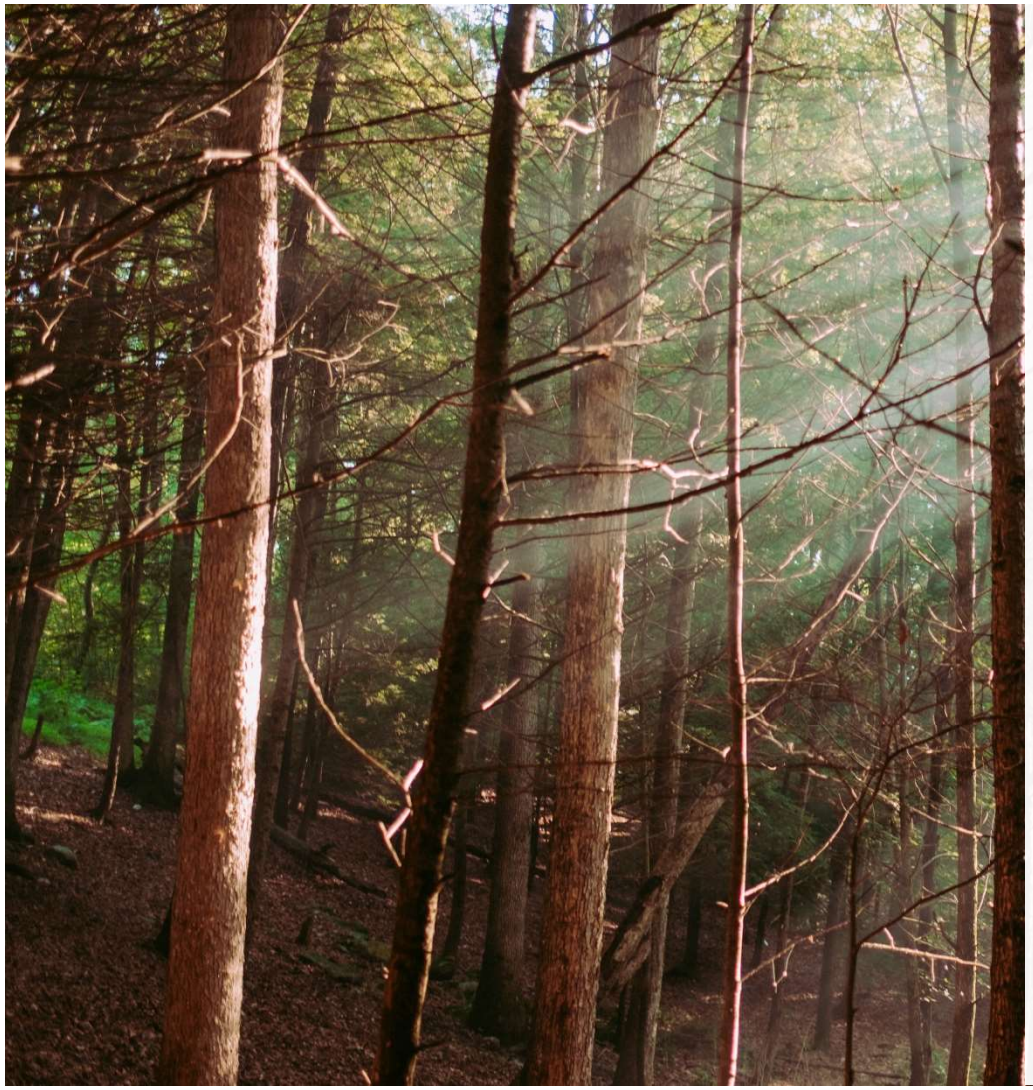


# BIGDATA CAMP


Daegu University Department of Mathematics

Team name : Homo Codeless

Leader : Lee Seung Jae
Members : Lee Jun Ho
Lee Ye Rin
Sang Ga Yeon
Park Ki Wan
Kim Chur Hong
Jung Hyo Eun



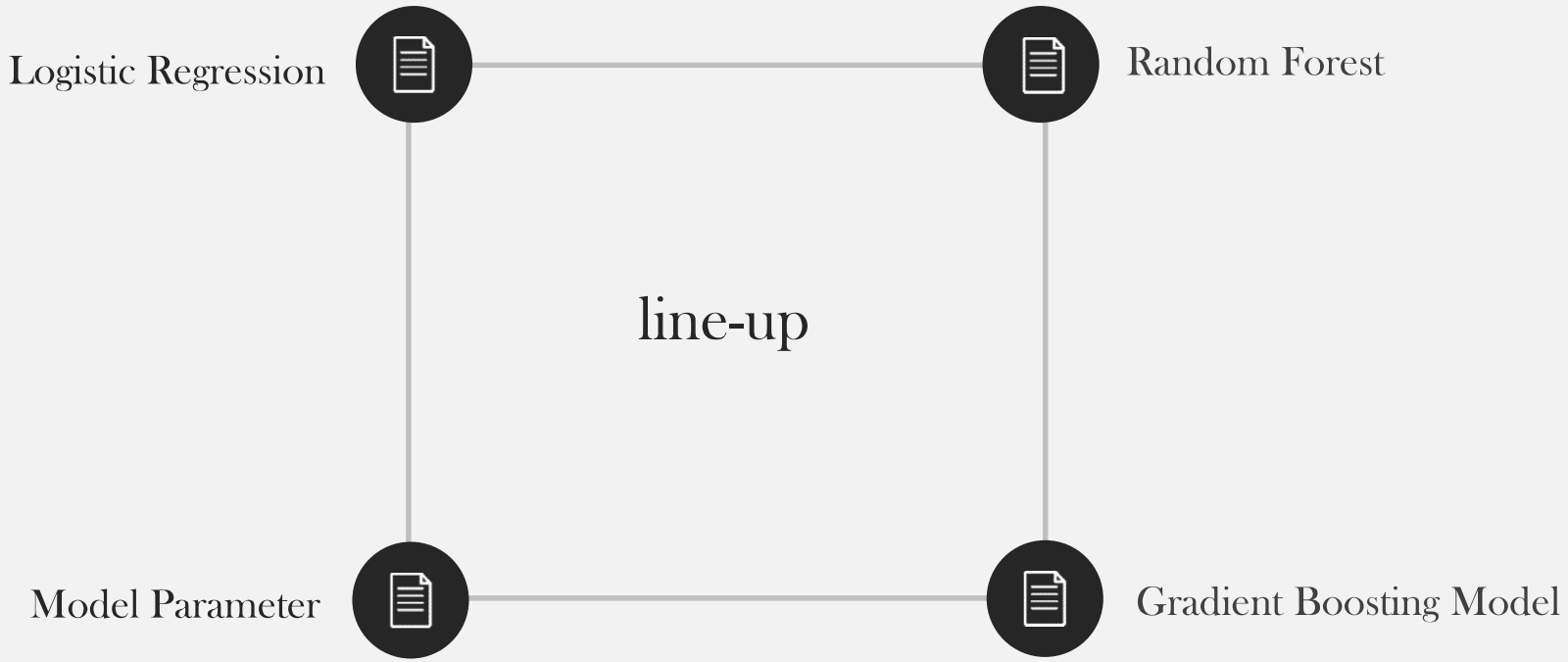
# MODEL

01 Logistic Regression

02 Random Forest

03 Gradient Boosting Model

# 01 INDEX


Logistic Regression
Random Forest
line-up
Model Parameter
Gradient Boosting Model

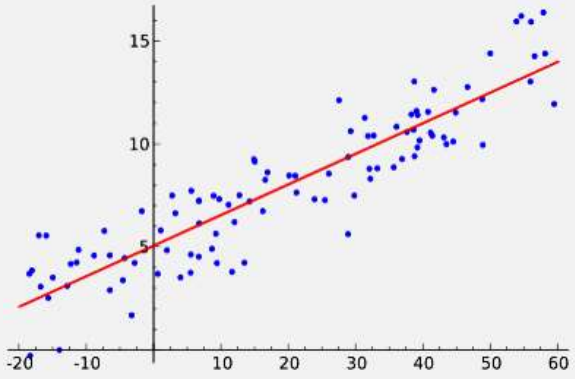
# 02 Logistic Definition

▶ Logistic 회귀분석의 정의

: 분석하고자 하는 대상들이 종속변수가 범주형 데이터를 대상으로 하고 있을 때,
관측치들이 어느 집단에 분류 되는가를 분석하고 이를 예측하는 모형을 개발하는데 사용되는 사용기법

▶ 선형회귀분석

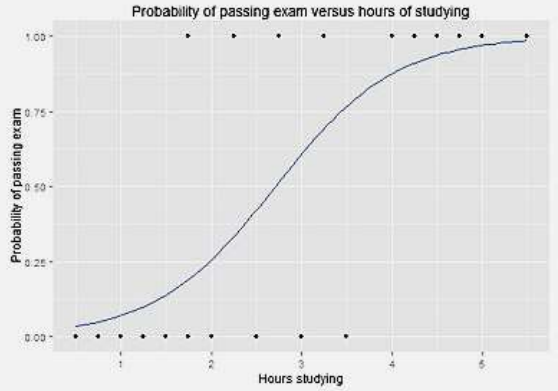
$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_q x_q$$




▶ Logistic 회귀분석

$$y = \frac{1}{1+e^{-z}} \ ,$$

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_q x_q$$



# 03 Logistic

## 01 데이터 탐색

-입력변수를 확인하고 구별
-예측 할 목표를 설정


## 03 Logistic 회귀모델 적용

- 테스트 집합 결과 예측 및 컨퓨젼 매트릭스
- 생성
- 정확도 추출

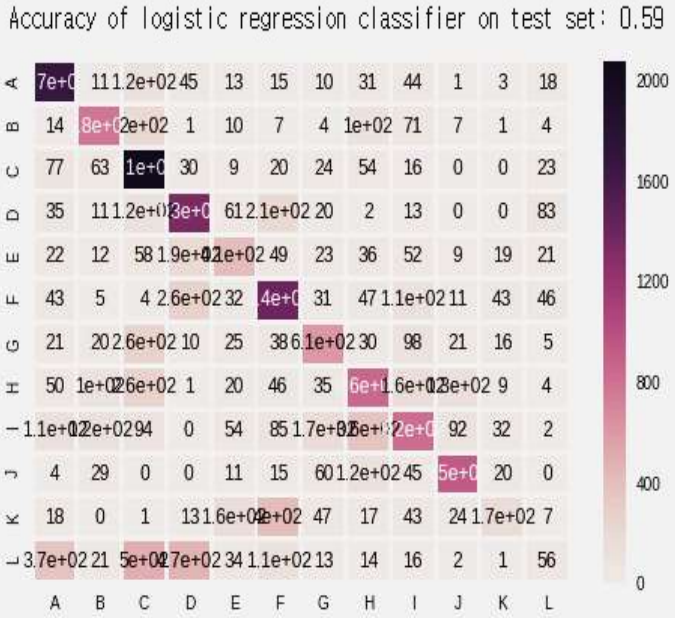
## 02 데이터 전처리

- 알 수 없는 열 삭제
- 데이터를 class 및 테스트 데이터로 나눔

## 04 시각화

# 04 Logistic Regression

IDEA 각각의 변수가 A~L까지의 Class에 들어갈 확률이 어떻게 될까?




1) Class와 그 외 변수를 비교하여 둘의 관계를 행렬로 나타냄.
2) 행렬 성분을 확률로 변환시킨 후 주대각 성분들의 합이 최종 확률.

RESULT 59%

# 05 IDEA

어떤 요인을 고려해서 모형을 설계할까?

## Model Parameter

최적의 Parameter 찾기


## Feature Engineering

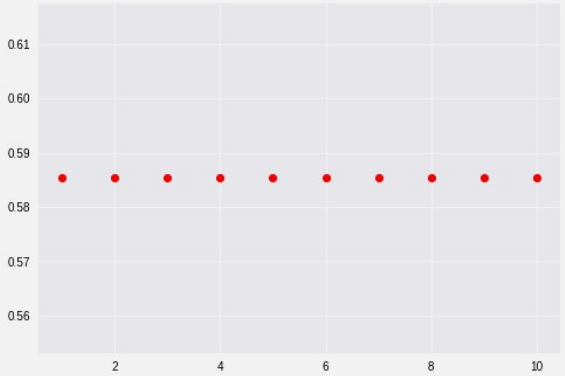
1) 상품 분류 군집화
2) 유의미한 feature 생성

## PCA

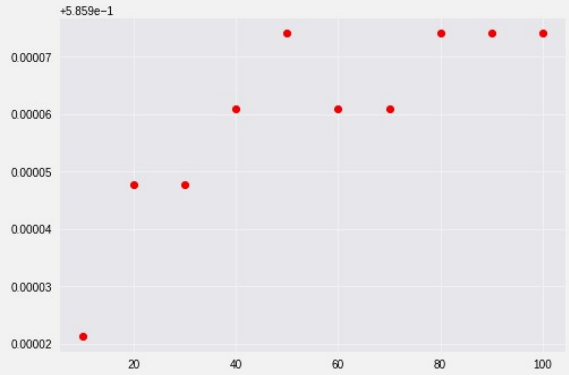
(Principal Component Analysis)
주성분 분석

# 06 Model Parameter

IDEA c값을 고정 or 시행횟수 값을 고정




▶ c값을 100으로 잡고
시행착오를 1~10번까지의 값




▶ c값을 10~100으로 잡은 값

**Definition**
랜덤 실험 또는 환경의 결과를 나타내는 수

1) c값 고정 (c : 100)
시행횟수 값을 1~10까지(1단위)
→ 0.58531

2) 시행횟수 값 고정 (시행횟수 : 5)
c값을 10~100까지 (10단위)
→ 58%

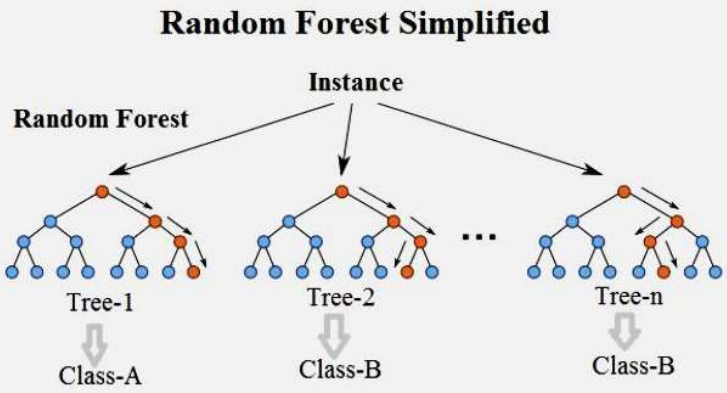
**RESULT** 58%

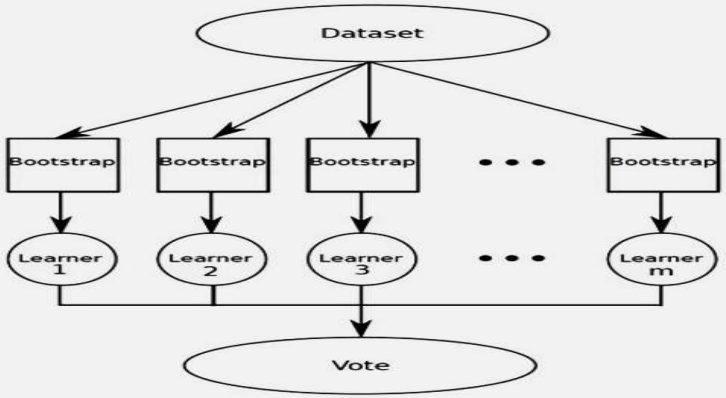
# 07 Random Forest Definition

▶ Random Forest정의

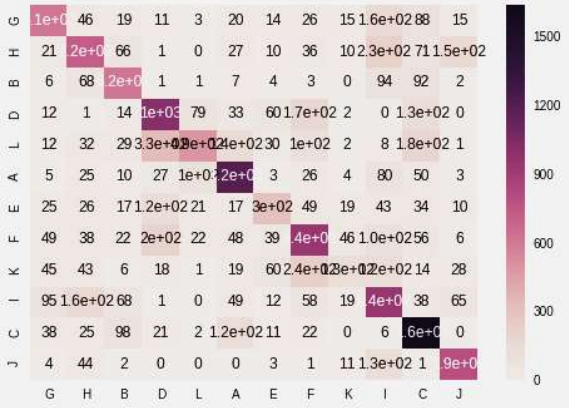
: 여러 개의 조금씩 다른 decision tree를 만든 뒤 다수결 투표 결과로 데이터 분류

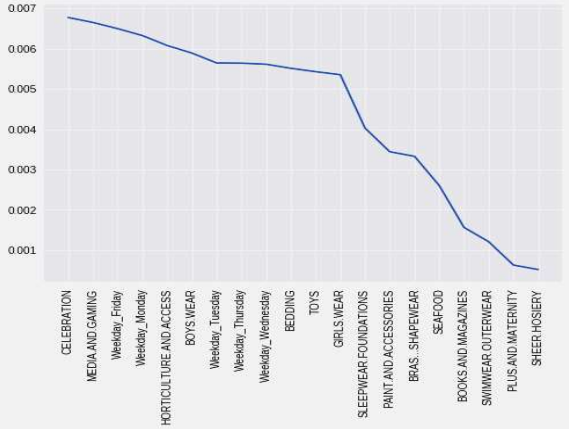
▶ Random Forest



# 08 Random Forest



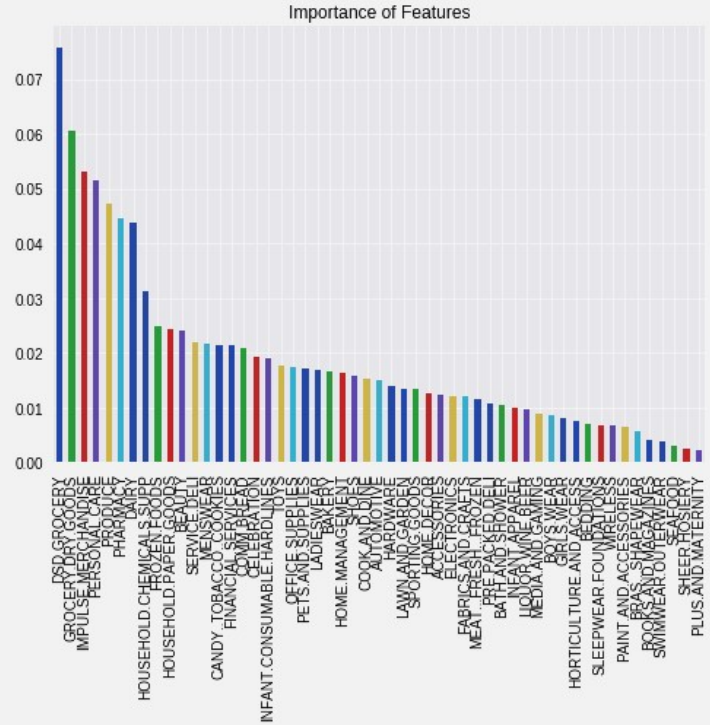

**IDEA** 같은 알고리즘을 사용하는 서로 독립인 예측 모델들을 평균/다수결로 합치면 정확도가 상승 한다?

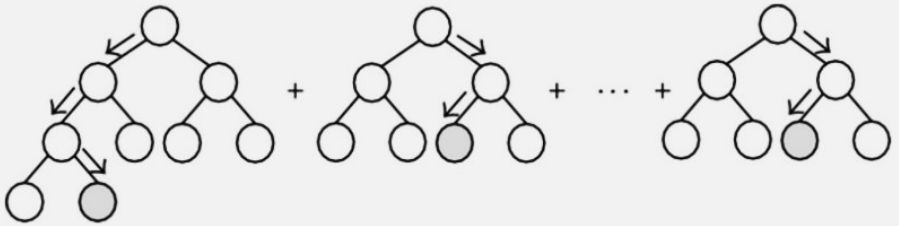
Decision Tree 의 개수와 max_depth를 조정

**RESULT** 61.94%

# 09 Gradient Boosting Model

Accuracy of the GBM on test set: 0.646



## IDEA

max_depth값을 변화시키며 모형의 정확도를 계산

→ max_depth : 10 result : 0.646

# 10 전체 데이터 비교


Logistic Regression


Model Parameter


Random Forest


Gradient Boosting Model


Q&A


thank you
for watching