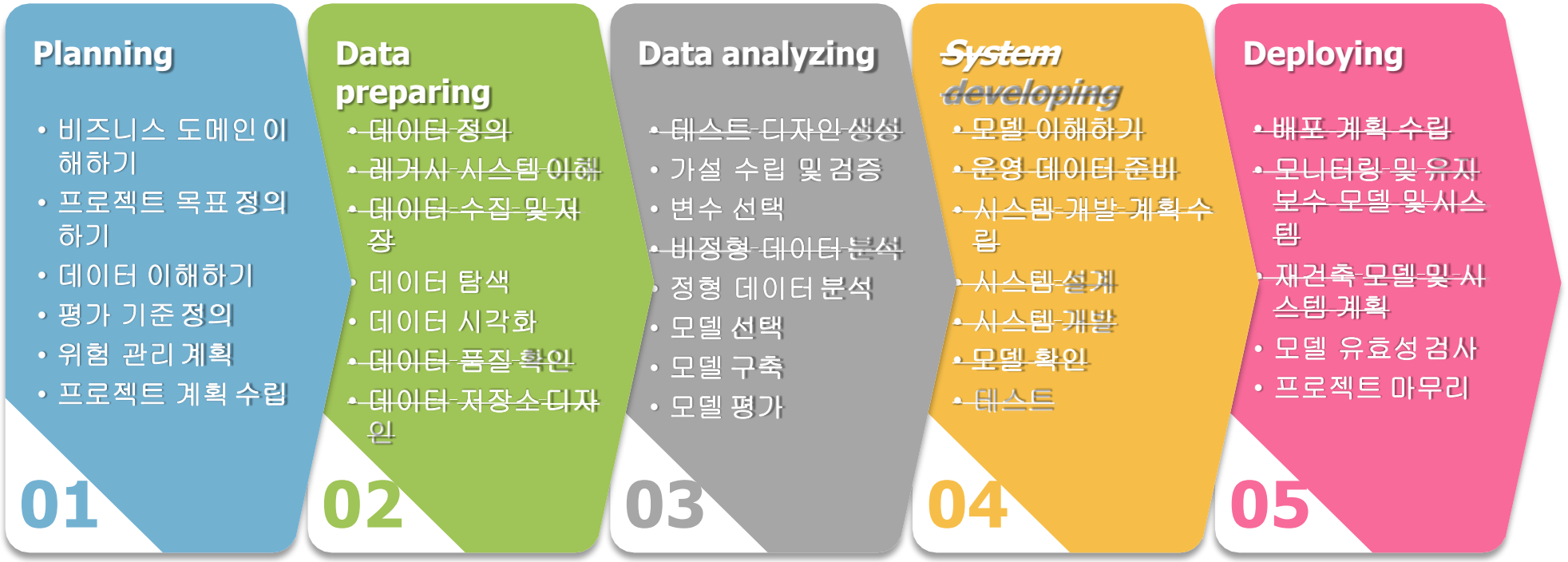
머신러닝 프로젝트

Process Roadmap

ANALYSIS PROJECT > PHASE



System developing

웹 애플리케이션을 구축하지 않는다면 **System developing** 단계에서의 활동은 하지 않아도 됩니다.

Project Planning

Planning Phase

ANALYSIS PROJECT > PHASE > Planning

Planning

- 비즈니스 도메인 이해하기
- 프로젝트 목표 정의하기
- 데이터 이해하기
- 평가 기준 정의
- 위험 관리 계획
- 프로젝트 계획 수립

01

Project Plan

- **Objective**
- **Plan Point**
- **Goal (Quantitative, Qualitative)**
- **Analytic (Classification, Estimation, Prediction, Association, Clustering)**
- **Assess Criteria (Quality)**
- **Required Data**
- **Teaming**
- **Time Box Schedule**
- **Risk Management**

Project Sponsorship

- **Quantitative Measurement(정 양적 측정) -** 목표는 통계로 측정됩니다.
- **Qualitative Measurement(정 성적 측정) -** 목표는 어떤 통계 없이 관리자의 관찰에 의해 측정됩니다.

Fill in the blanks

ANALYSIS PROJECT > PHASE > Planning

	보험사기자 예측	My Project
Objective	보험 사기자로 예상되는 사람들 분류	
Plan Point	사기자 예측	
Goal (Quantitative, Qualitative)	보험사기자 예측률 향상	
Analytic (Classification, Estimation, Prediction, Association, Clustering)	일반인과 사기자로 분류 (Classification)	
Assess Criteria (Quality)	F-measure를 이용한 모델 검증	
Required Data	고객 데이터	
Teaming	팀 편성	
Time Box Schedule	프로젝트 일정표	
Risk Management	일정지연, 팀원손실 등 위험관리	

Data preparing

Data preparing Phase

Data preparing

- ~~데이터 정의~~
- ~~레거시 시스템 이해~~
- ~~데이터 수집 및 저장~~
- 데이터 탐색
- 데이터 시각화
- ~~데이터 품질 확인~~
- ~~데이터 저장소 디자인~~

02

Data Acquirement

- **Structured/Unstructured Data**
- **Data Item Definition**
- **Legacy ERD/Meta Data/Process**
- **Data Life Cycle**
- **Data Storing Design (RDB, NoSQL, HDFS, …)**

Data Quality

- **Understanding Missing & Outlier Data**
- **EDA**
- **Consistent Data Quality**

Fill in the blanks

ANALYSIS PROJECT > PHASE > Data preparing

	보험사기자 예측	My Project
Structured/Unstructured Data	데이터 재구조화, 데이터 병합	
Data Item Definition	테이블 및 변수 정의서 참고	
Legacy ERD/Meta Data/Process	관계 없음	
Data Life Cycle	2016년 데이터	
Data Storing Design (RDB, NoSQL, HDFS, …)	csv 파일 형식	
Understanding Missing/Outlier	NA, NULL 값 등 결측치 처리	
EDA	기본 통계량 확인, 데이터 탐색	
Consistent Data Quality	관계 없음	

Data analyzing

Data analyzing Phase

Data analyzing

- ~~테스트 디자인 생성~~
- 가설수립 및 검증
- 변수 선택
- ~~비정형 데이터 분석~~
- 정형 데이터 분석
- 모델 선택
- 모델 구축
- 모델 평가

03

Model

- **Handling Missing & Outlier Data**
- **Selection Variables, Derived Data**
- **Advanced Analytics Results (Classification, Estimation, Prediction, Association, Clustering)**
- **Model Evaluation**
- **Considering Overfitting**
- **Model Algorithm Description**

Assessing Model

- **Model Assessing Criteria**
- **Training Data, Test Data, Validation Data**

Fill in the blanks

ANALYSIS PROJECT > PHASE > Data analyzing

	보험사기자 예측	My Project
Understanding Missing & Outlier Data	결측치, 이상치 처리	
Selection Variables, Derived Data	변수 선택(PCA, SVMRFE 등)	
Advanced Analytics Results	NNet, SVM, XGBoost 등을 이용한 분류(Classification)	
Model Evaluation	F-measure를 이용한 모델 검증	
Considering Overfitting		
Model Algorithm Description	NNet, SVM, XGBoost	
Model Assessing Criteria	F-measure가 얼마인가?	
Training Data, Test Data, Validation Data	Training Set 70%, Test Set 30%	

가설 수립 리스트

변수 선택 및 추가 시나리오

개인 특성에 따른 가설

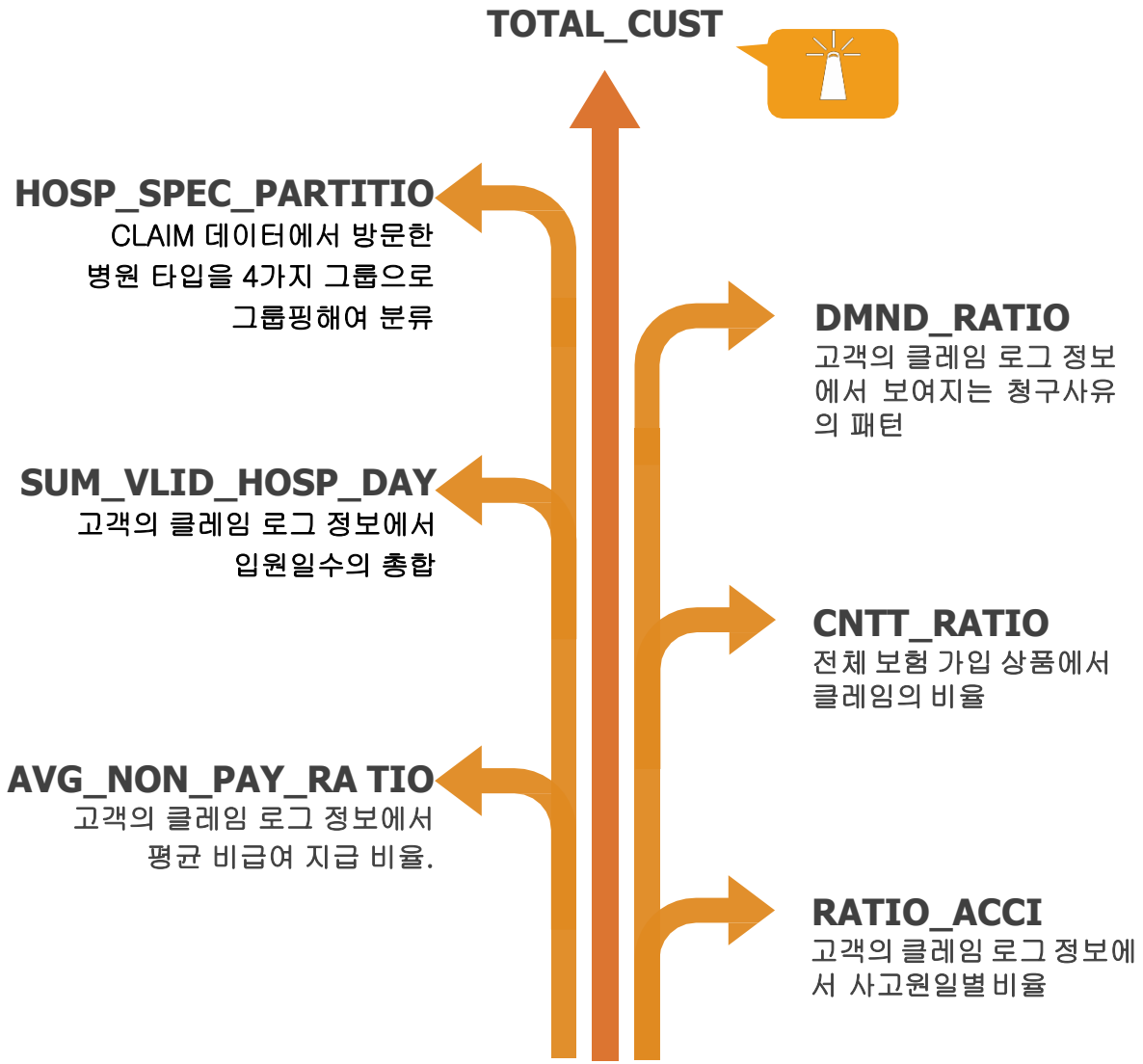
- ★ 고객의 성별, 나이에 따라 사기자 비율이 달라지지 않을까?
- ★ **FP**경력에 따라 사기자 비율이 달라지지 않을까?
- ★ 지역에 따라 사기자 비율이 달라지지 않을까?
- ★ 주거 타입에 따라 사기자 비율이 달라지지 않을까?
- ★ 고객 소득이 낮을 수록 사기자 비율이 낮지 않을까?

고객 로그 정보에 따른 가설

- ★ 사기자들은 보험 클레임을 많이 걸지 않을까?
- ★ 사기자들이 특정 병원 종류를 선호하지 않을까?
- ★ 금감원 유의 병원에 갔으면 사기자로 의심해야 할까?
- ★ 사기자들은 자기부담금 비율이 낮지 않을까?
- ★ 사기자들이 선호하는 청구사유가 있지 않을까?
- ★ 입원일수가 증가할 수록 사기자 비율이 증가하지 않을까?
- ★ 사기자들이 선호하는 보험상품 종류가 있지 않을까?

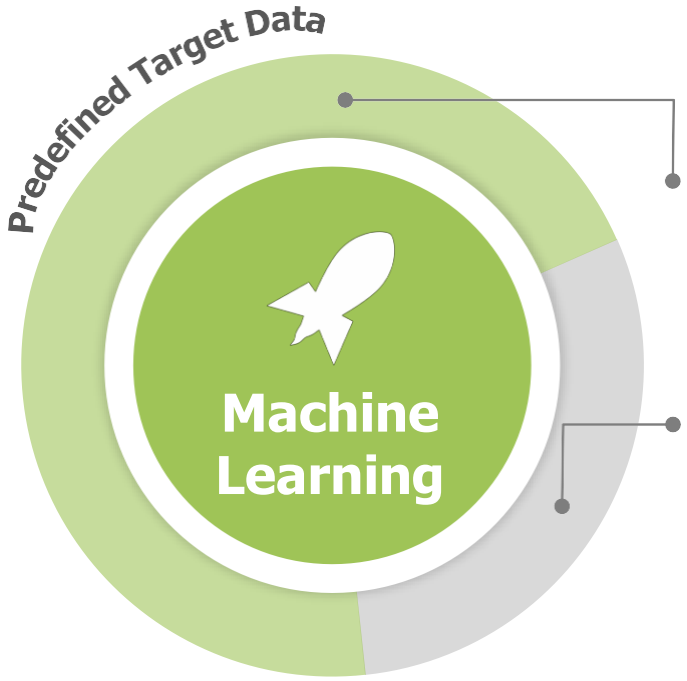
가설에 따른 파생변수 추가

원래 데이터가 가지고 있지 않은 변수를 파악하고 추가하는 것은 데이터 분석에 있어서 매우 중요



Data Sampling

ANALYSIS PROJECT > PHASE > Data analyzing



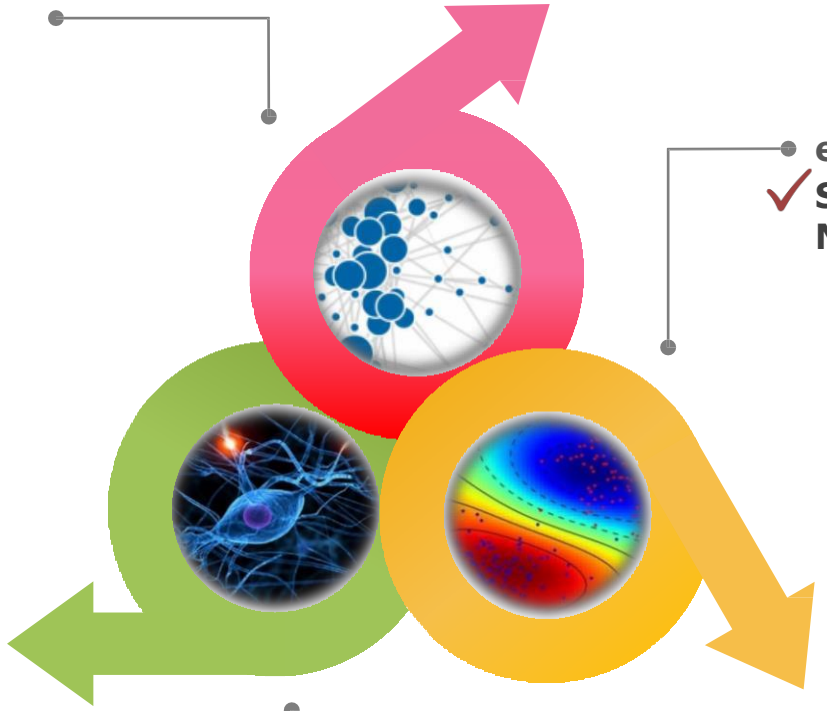
70%
Training Set
학습 모델을 생성하기 위한 데이터 셋

30%
Test Set
검정 통계량 확인을 통하여 모델의 적합도를 확인하기 위한 데이터 셋

Training Set & Test Set
이미 분류되어 있는 데이터 셋에 대하여 트레이닝 셋과 테스트 셋으로 나누고 트레이닝 셋을 이용하여 학습 모델을 생성한다.
테스트셋과 예측 데이터를 비교한 검정 통계량 확인을 통하여 모델의 적합도를 확인한다.

ML(Classification) Model

모델링을 위해 어떤 패키지와 어떤 모델 알고리즘을 사용 할 것인가?



xgboost
Extreme Gradient Boosting

randomForest Rando
m Forest Model

adabag
Adaptive Boosting

party
Tree Model

nnet
Neural Network Lo
gistic Regression

e1071
Support Vector Machine
Naive Bayes Model

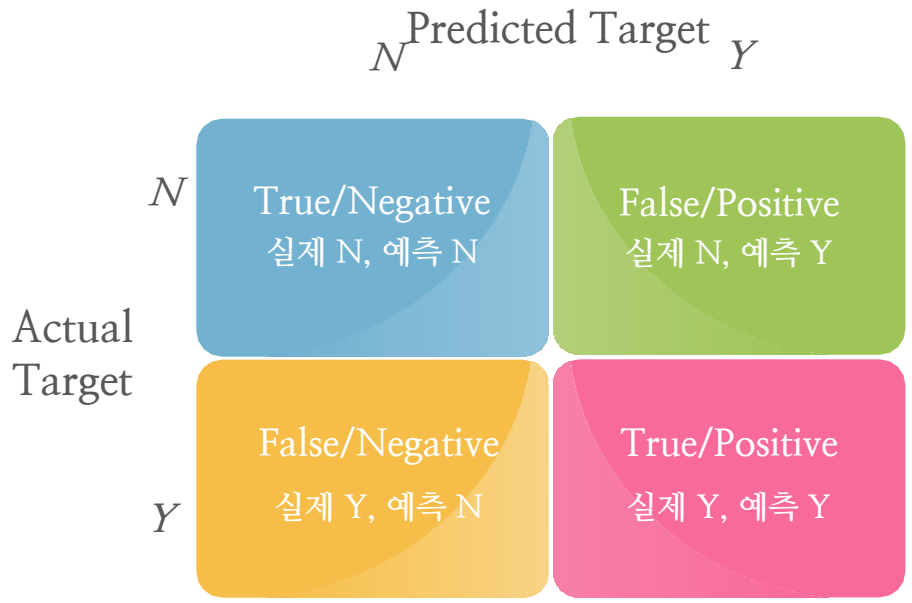
MASS
Linear Discriminant Model

h2o
Deep Learning

class
k-Nearest Neighbor

분류 모델의 평가 측도

ANALYSIS PROJECT > PHASE > Deploying



혼동 행렬(Confusion Matrix)

다양한 모델 평가 방법들

- TCO(Total Cost of Ownership)
- ROI(Return Of Investment)
- IRR(Internal Rate of Return)
- NPV(Net Present Value)
- PP(Payback Period)
- RMSE(Root Mean Square Error)
- MAPE(Mean Absolute Percentile Error)

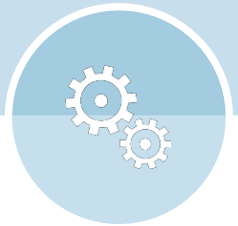
오분류에 따른 위험성이 다를 경우, 민감도(Sensitivity), 특이도(Specificity), 정밀도(Precision), F-measure를 사용

메트릭	계산식	의미
정밀도 (Precision)	$\frac{TP}{FP+TP}$	Y으로 예측된 것 중 실제로도 Y인 경우의 비율
민감도 (Recall)	$\frac{TP}{FN+TP}$	실제로 Y인 것들 중 예측이 Y로 된 경우 비율(=Sensitivity)
정확도 (Accuracy)	$\frac{TP+TN}{TP+FP+FN+TN}$	전체 예측에서 옳은 예측의 비율
특이도 (Specificity)	$\frac{TN}{FP+TN}$	실제로 N인 것들 중에서 예측이 N으로 된 경우의 비율
오류율 (FP Rate)	$\frac{FP}{FP+TN}$	Y가 아닌데 Y로 예측된 비 율.(=error rate) 1-Specificity와 같은 값
F-measure	$2 * \frac{(Precision * Recall)}{(Precision + Recall)}$	Precision과 Recall의 조화 평균. 시스템의 성능을 하나의 수치로 표현하기 위해 사용하는 점수. 0~1사이의 값을 가짐. Precision과 Recall 두 값이 골고루 클 때 큰 값을 가짐.
Kappa	$K = \frac{Paccuracy - P(e)}{1 - P(e)}$	코헨의 카파. 두 평가자의 평가가 얼마나 일치하는지 평가하는 값. 0~1사이의 값을 가짐. P(e)는 두 평가자의 평가가 우연히 일치할 확률. 코헨의 카파는 두 평가자의 평가가 우연히 일치할 확률을 제외한 뒤의 점수.

Lab

ANALYSIS PROJECT > PHASE > Data analyzing

Step 1



R Review

R Review Machine Learning

R_Review.R
Machine Learning.R

Step 2

Preprocessing

데이터 전처리
데이터 탐색
데이터 탐색 후 변수 제거

Lab_1-1.R
Lab_1-2.R
Lab_1-3.R

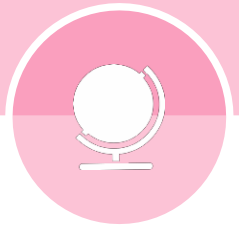
Step 3

Model Test

결측치 추정
파생 변수 추가
변수 선택 기계학습 학습 모델 평가

Lab_2-1.R
Lab_2-2.R
Lab_2-3.R
Lab_2-4.R
Lab_2-5.R

Step 4



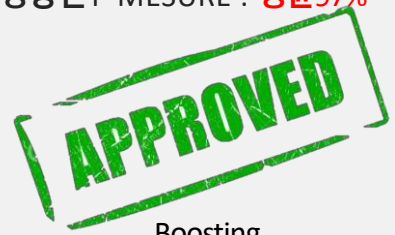
Modeling

Training Set & Prediction Data Set ML을 이용한 분류 모델 검증

Lab_3-1.R
Lab_3-2.R
Lab_3-3.R

모델 선택

ANALYSIS PROJECT > PHASE > Data analyzing

Neural Net	XGBoost	Random Forest
R 'nnet' PACKAGE	R 'xgboost' PACKAGE	R 'randomForest' PACKAGE
학습 시간: 15~25min	학습 시간: 5~30sec	학습 시간: 5~10min
사기자 F-MESURE : 평균60%	사기자 F-MESURE : 평균62%	사기자 F-MESURE : 평균58%
정상인 F-MESURE : 평균96%	정상인 F-MESURE : 평균97%	정상인 F-MESURE : 평균96%
		
Neural network 인간뇌의신경회로망에서수행하는정보처리방식을모방한알고리즘	Boosting 잘못분류된개체들에집중하여새로운분류규칙을만드는단계를반복하는방법	Bagging 단순복원임의추출법을통해원자료로부터크기가동일한여러개의표본자료를 추출해각모델링하여평균이나보팅하 는방법

VS VS

Comments

샘플링을여러번시행해평균값을계산함