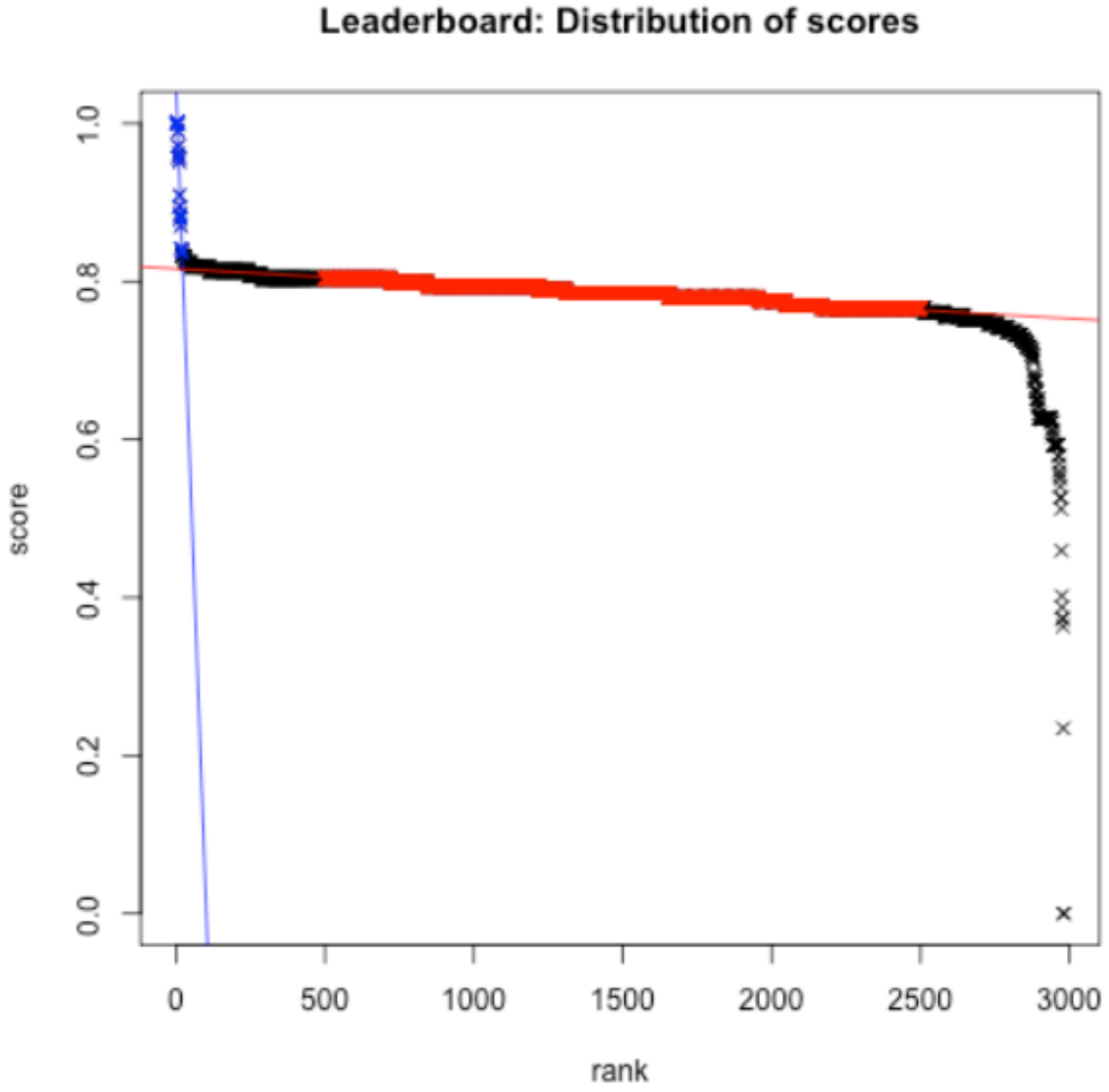
타이타닉 간단 Review

김도완

19.09.30


Leaderboard: Distribution of scores
score
rank
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
3000

- 단순 성별분류 : 0.76 (gender submission)
- 베이스라인 : 0.76 ~ 0.78
- 0.8 이상 ?
  - 모델 파라메타 튜닝, 모델 앙상블
  - feature engineering

- 참고 커널1: Titanic Top 4% with ensemble modeling
- 참고 커널2: https://kaggle.com/karell/kakr-1st-titanic-gender-model-with-blending

단순 EDA

- 수치형 컬럼 : 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare'
  - SibSp(형제/자매수) : 형제,자매가 많을수록 생존확률이 낮아지는 경향
  - Parch(부모/자식수) : 소규모 가족(1,2)이 대체로 생존확률 높다. 4이상의 대가족과 0독신은 낮다.
  - Age : 0~5세 어린아이들의 생존확률 높다.
  - Fare : 비싼 요금일수록 생존확률 높다.
  - Pclass : 높은 좌석등급일수록 생존확률 높다.

- 문자형 컬럼 : 'Name', 'Sex', 'Ticket', 'Cabin', 'Embarked'
  - Sex : 여성의 생존확률이 뚜렷이 높다.
  - Ticket : unique수가 너무 많음.
  - Cabin(좌석) : NaN값 너무 많음. 무의미
  - Embarked : C의 생존확률이 높다. 하지만 C의 1등급좌석비율이 크기 때문 무의미할수도
  - Name : 이름의 Title은 성별과 연관, 특이한 호칭의 여성과 아이는 생존률 높음.

이름 + 티켓 = 가족

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0.0	3	Rice, Master. Eugene	male	2.0	4	1	382652
0.0	3	Rice, Master. Arthur	male	4.0	4	1	382652
0.0	3	Rice, Master. Eric	male	7.0	4	1	382652
0.0	3	Rice, Master. George Hugh	male	8.0	4	1	382652
0.0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652
NaN	3	Rice, Master. Albert	male	10.0	4	1	382652

이름 : Family name, Title. First name
유추 : 가족그룹 / (아이, 직업, 지위)

가족의 생존확률은 비슷할 것이다. 가정

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1.0	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia...	female	38.0	1	5	347077
0.0	3	Asplund, Master. Clarence Gustaf Hugo	male	9.0	4	2	347077
1.0	3	Asplund, Miss. Lillian Gertrud	female	5.0	4	2	347077
1.0	3	Asplund, Master. Edvin Rojj Felix	male	3.0	4	2	347077
NaN	3	Asplund, Master. Filip Oscar	male	13.0	4	2	347077
NaN	3	Asplund, Mr. Carl Oscar Vilhelm Gustafsson	male	40.0	1	5	347077
NaN	3	Asplund, Mr. Johan Charles	male	23.0	0	0	350054
NaN	3	Asplund, Master. Carl Edgar	male	5.0	4	2	347077

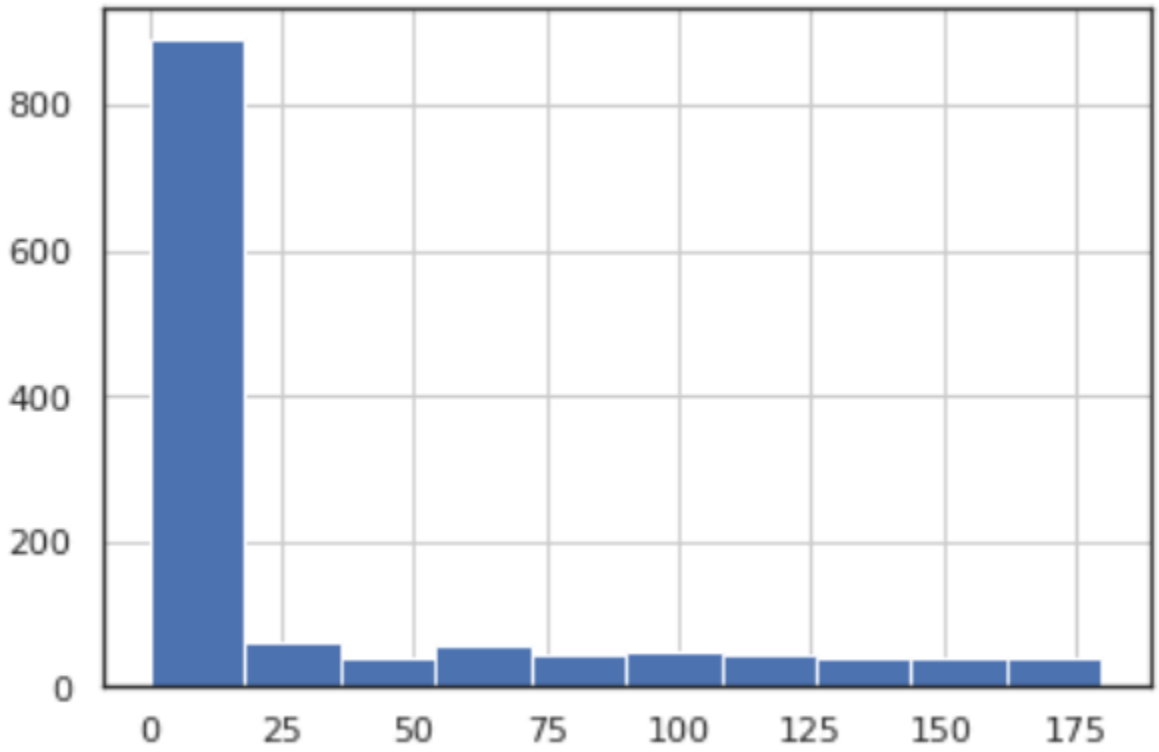
Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1.0	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742
1.0	3	Johnson, Miss. Eleanor Ileen	female	1.0	1	1	347742
0.0	3	Johnson, Mr. William Cahoone Jr	male	19.0	0	0	LINE
0.0	3	Johnson, Mr. Alfred	male	49.0	0	0	LINE
0.0	3	Johnson, Mr. Malkolm Joackim	male	33.0	0	0	347062
1.0	3	Johnson, Master. Harold Theodor	male	4.0	1	1	347742

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0.0	2	Davies, Mr. Charles Henry	male	18.0	0	0	S.O.C. 14879
1.0	2	Davies, Master. John Morgan Jr	male	8.0	1	1	C.A. 33112
0.0	3	Davies, Mr. Alfred J	male	24.0	2	0	A/4 48871
NaN	3	Davies, Mr. John Samuel	male	21.0	2	0	A/4 48871
NaN	3	Davies, Mr. Evan	male	22.0	0	0	SC/A4 23568
NaN	3	Davies, Mr. Joseph	male	17.0	2	0	A/4 48873
NaN	2	Davies, Mrs. John Morgan (Elizabeth Agnes Mary...	female	48.0	0	2	C.A. 33112

Feture Engineering

- 그룹핑 : First Name + Ticket[:-2]
- 후처리: FamilySize=0 or 그룹원1명 -> Alone(혼자) 처리

	Name	Ticket	Fsize	Family_Ticket
	Braund, Mr. Owen Harris	A/5 21171	1	ALONE_PRD
	Cumings, Mrs. John Bradley (Florence Briggs Th...	PC 17599	1	Cumings-PC 175
	Heikkinen, Miss. Laina	STON/O2. 3101282	0	ALONE
	Futrelle, Mrs. Jacques Heath (Lily May Peel)	113803	1	Futrelle-1138
	Allen, Mr. William Henry	373450	0	ALONE



188개의 그룹이 나왔으며
Alone으로 추정되는 사람은 832명이다.
832 / 1309 = 0.63
family size가 0인 사람으로 Alone계산시
790 / 1309 = 0.60

	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Fsize	Title	
	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	1	Mr	
	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	1	Mrs	C
	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	0	Miss	
	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	1	Mrs	
	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	0	Mr	

	Survived	Pclass	Sex	Age_band	Fare_band	Fsize	Title	Family_Ticket
0	0.0	3	0	2	0	1	0	0
1	1.0	1	1	3	4	1	1	1
2	1.0	3	1	2	1	0	2	2
3	1.0	1	1	3	4	1	1	3
4	0.0	3	0	3	1	0	0	2

* 중요: Family_Ticket컬럼에 대해 One-hot encoding 한다.

188개의 컬럼 생성

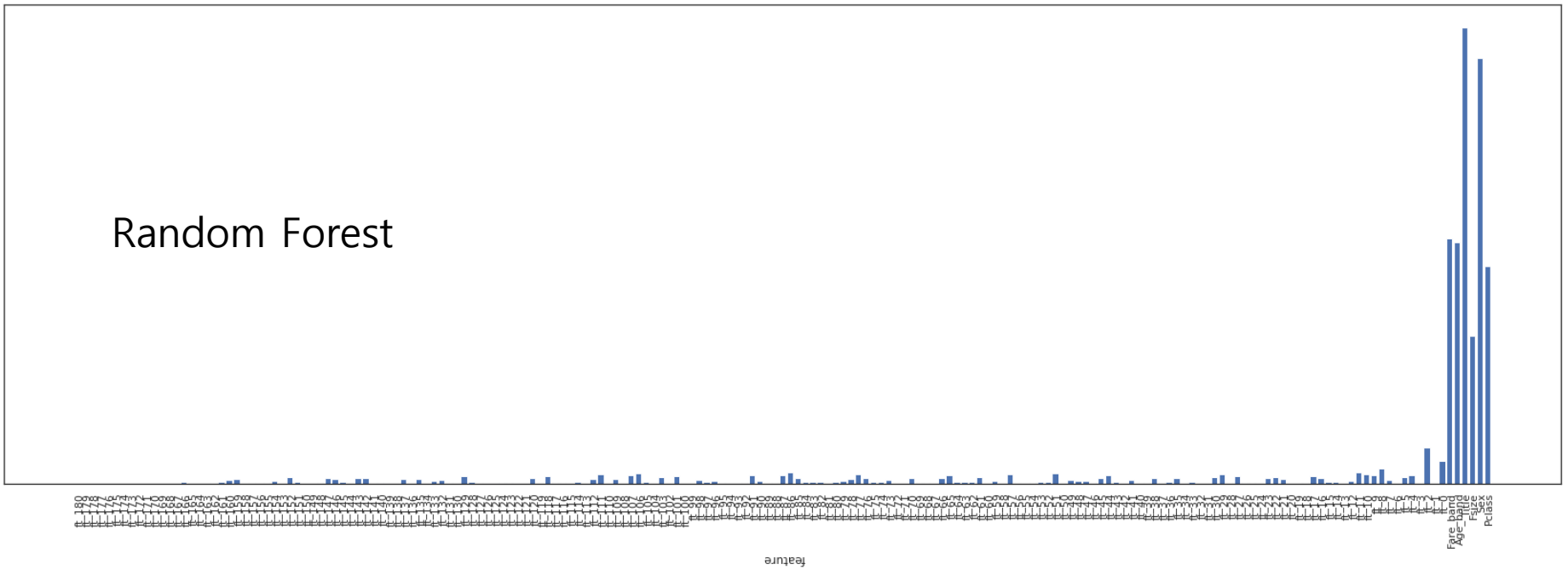
	Survived	Pclass	Sex	Fsize	Title	Age_band	Fare_band	ft_0	ft_1	ft_2	...	ft_171	ft_172	ft_173	ft_174	ft_175	ft_176	ft_177	ft_178	ft_179	ft_180
0	0.0	3	0	1	0	2	0	1	0	0	...	0	0	0	0	0	0	0	0	0	0
1	1.0	1	1	1	1	3	4	0	1	0	...	0	0	0	0	0	0	0	0	0	0
2	1.0	3	1	0	2	2	1	0	0	1	...	0	0	0	0	0	0	0	0	0	0
3	1.0	1	1	1	1	3	4	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0.0	3	0	0	0	3	1	0	0	1	...	0	0	0	0	0	0	0	0	0	0

baseline : Pclass, Sex, Age_band, Fare_band, Fsize, Title
특성추가 : Family_Ticket

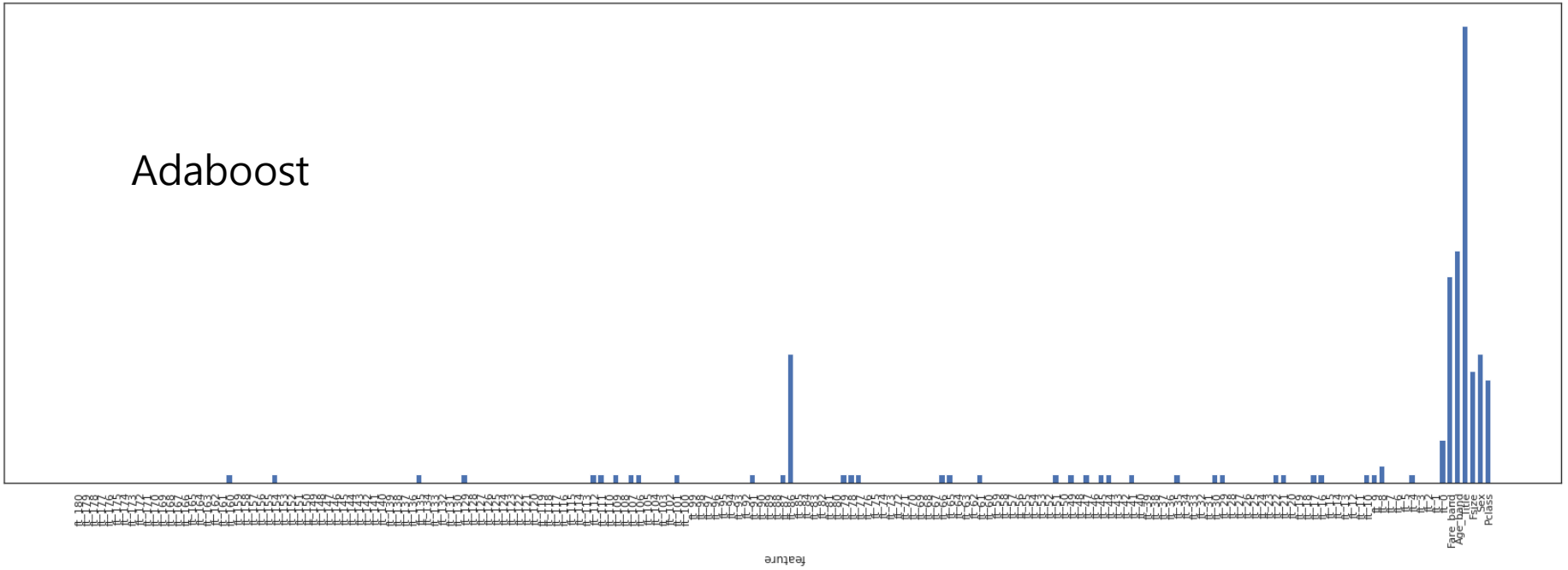
	특성추가 전 (baseline)	특성추가 후
RandomForest	0.75 ~ 0.78	0.79 ~ 0.80
AdaBoost	0.77 ~ 0.79	0.81
SVC		

약 0.2점이 증가하는 효과

Random Forest



Adaboost



462	**dowaari**		0.81818	12	13h

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Getting Started · Ongoing · tutorial, tabular data, binary classification

462/10794
Top 5%

- 주제: 장고를 이용한 웹개발 스터디
- 기간: 10월 ~ 12월 (12주), 10/6일 미팅
- 시간: 일 14:00~ 17:00 (주말, 오전/오후 변동가능)
- 장소: 서면 또는 센텀 (논의 후 결정)
- 비용: 무료 (스터디룸예약시 비용발생)

- 스터디 방법
  - 진행자 1명이 발표하고 함께 실습하는 스타일, 모르는 부분은 함께 고민하기
  - 짧은시간내에 최대한 많은 웹 만들어보기 , 오픈소스 필사
  - 참가 난이도: 초중급이상 (파이썬 코딩가능하며 장고기초지식은 있어야함.)

- 커리큘럼(계획)
  1. 장고걸스 튜토리얼 따라하기 (3주) – 진행자2명
  2. 인터넷 또는 책에서 예제 선정하여 공부 (4주) – 진행자 2명
  3. 온라인강의 선정하여 공부 (4주) – 진행자 2명
  4. 자신만의 웹구현, 개인별 주제 정한 후 발표 – 모두

- 문의는 오픈채팅 개설 , 참가신청 google폼으로