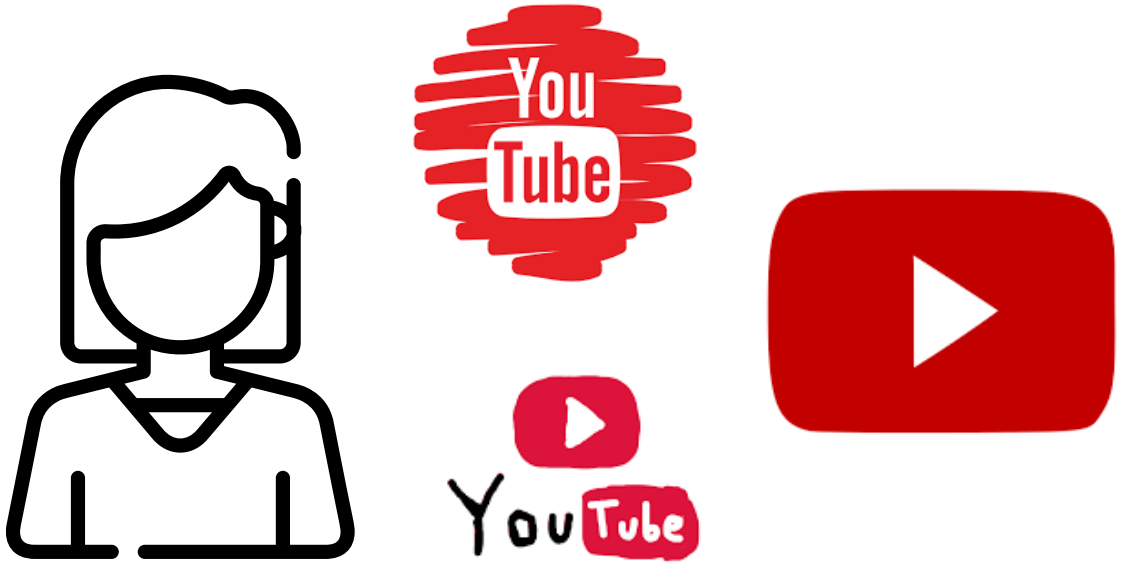


Who Is The Next Big YouTuber Star ?

이유경

INDEX

01 프로젝트 개요

- 선정 배경 및 목적

01 | 프로젝트 개요

- 선정 이유 및 목적

내가 좋아하는 7인의 YouTuber 의 영상 **데이터를 수집**하여 한번 분석해볼까?

그런데 구독자수, 날짜 이런 변수들은 수치화가 쉬울 듯 하나, 제일 중요한 **영상 제목**을 어떻게 **수치화** 시킬 수 있을까?

수치화 시킨 영상 제목을 포함한 다른 변수들을 모두 활용하여 **조회수 예측 모델**을 만들 수 있을까?

내가 원하는 키워드 혹은 컨텐츠가 포함된 영상 제목, 나의 현재 구독자 수, 업로드 후 기대 기간 등을 입력하면, **예상 조회수**를 볼 수 있을까?

내가 만약 YouTube를 시작한다면, **어떤 컨텐츠, 어떤 제목**으로 시작하면 높은 조회수를 얻을 수 있을까?

YouTube 영상 데이터를 수집한 후 조회수 예측 모델을 생성하여,
높은 조회수를 만들어낼 수 있는 영상 제목 키워드 및 컨텐츠를 알아보자 !

02 데이터 수집 및 전처리

- YouTube 데이터 크롤링
- 채널명/구독자수/업로드 날짜/조회수/좋아요/싫어요 처리
- 영상 제목 처리 (토큰화 및 TF-IDF생성)
- EDA

02 | 데이터 수집 및 전처리

- YouTube 데이터 크롤링

▪ 구독자 10만 이상인 7인의 요리 YouTuber의 모든 영상에서 채널명/구독자 수/영상 제목/업로드 날짜/조회수/좋아요/싫어요 정보 크롤링하여 총 4568개의 데이터셋 구성

	youtuber	subscribers	name	upload_date	hits	likes_num	dislikes_num
475	강쉪 - YouTube	구독자 19.5만명	강쉪:) 카레 요리 / 얼큰 ~ 칼칼! 고구마 카레 만들기 = Sweet pot...	2017. 11. 1.	86회	4개	0개
1701	만개의레시피 10K Recipe - YouTube	구독자 35.6만명	그냥 콘치즈가 아니야! 소시지콘치즈 ;) [만개의레시피]	2019. 2. 27.	57,491회	92개	1개
289	강쉪 - YouTube	구독자 19.5만명	쉽고 간단한 ,느타리버섯 볶음 만드는법~ 담백하고, 풍부하게~ [강쉪]	2018. 8. 15.	182,960회	2.1천개	125개
787	만개의레시피 10K Recipe - YouTube	구독자 35.6만명	SUB) 흑임자소스와 연두부의 만남~! 연두부샐러드 :Soft Soybean Cur...	2019. 8. 21.	1,971회	34개	1개
3	강쉪 - YouTube	구독자 19.5만명	계란으로 할수있는 간단하고 맛있는 3가지 요리~ 3 kinds of egg cook...	2020. 3. 19.	262,820회	3천개	86개
294	강쉪 - YouTube	구독자 19.5만명	건새우 마늘볶음 만들기~ 마늘향이 솔솔~, 짭쪼롬하게, 맛있게~ [강쉪]	2018. 8. 9.	10,432회	135개	7개
354	강쉪 - YouTube	구독자 19.5만명	부추김치 담그는 법~ 쉽고 간단하고 맛있게 ~[강쉪]	2018. 5. 4.	93,110회	581개	34개
912	만개의레시피 10K Recipe - YouTube	구독자 35.6만명	🔥영양 만점 #생선조림 부터 #국물반찬까지 #오늘의식단 🔥 [만개의레시피]	2019. 7. 27.	3,233회	53개	1개
188	승우아빠 - YouTube	구독자 44.2만명	냉동 만두계를 평정하러 나타난 그 제품 리뷰	2019. 6. 17.	113,564회	1.8천개	43개
139	소소황 Cook & Eat - YouTube	구독자 12.7만명	지금 부산에서 가장 핫한 뉴웨이브 컨셉의 분식 맛집	2019. 7. 26.	2,311회	32개	3개
172	만개의레시피 10K Recipe - YouTube	구독자 35.6만명	냉동실에 식빵이 들어가 틈이 없어요!! 식빵레시피가 넘쳐서! [만개의레시피]	2020. 2. 3.	12,024회	253개	4개
2681	만개의레시피 10K Recipe - YouTube	구독자 35.6만명	상큼 깔끔하게 ♥ 토마토소스미역냉채 [만개의레시피]	2018. 7. 9.	528회	10개	0개
359	만개의레시피 10K Recipe - YouTube	구독자 35.6만명	두부젓국찌개 I 한식자격증 실기 I 시험시간 20분 [만개의레시피]	2019. 12. 8.	1,440회	24개	0개
2890	만개의레시피 10K Recipe - YouTube	구독자 35.6만명	친구들과 모였을때♥밀푀유나베 [만개의레시피]	2018. 5. 13.	3,496회	40개	0개
302	승우아빠 - YouTube	구독자 44.2만명	구독자 10000명, 아니 20000명, 아니 25000명 감사합니다..?!	2018. 12. 12.	16,121회	789개	6개

02 | 데이터 수집 및 전처리

- 채널명/구독자수/업로드 날짜/조회수/좋아요/싫어요 처리

- **youtuber(채널명)** : -> 범주화 시켜 **'youtuber_id'** 변수로 생성

- **subscribers(구독자수) / hits(조회수) / likes_num(좋아요) / dislikes_num(싫어요)** :

  만명,천명,회,개수 등 문자로 표현된 것들을 모두 숫자로 변환

- **upload_date(업로드 날짜)** : [오늘 날짜 - 업로드 날짜] = 00일 전 -> **'days_after_upload'** 변수로 추가

	youtuber_id	name	subscribers	days_after_upload	likes_num	dislikes_num	hits
475	6	강쉪:) 카레 요리 / 얼큰 ~ 칼칼! 고구마 카레 만들기 = Sweet pot...	195000000	887	400	0	86
1701	3	그냥 콘치즈가 아니야! 소시지콘치즈 ;) [만개의레시피]	356000000	404	9200	1	57491
289	6	쉽고 간단한 ,느타리버섯 볶음 만드는법~ 담백하고, 풍부하게~ [강쉪]	195000000	600	210000	125	182960
787	3	SUB) 흑임자소스와 연두부의 만남~! 연두부샐러드 :Soft Soybean Cur...	356000000	229	3400	1	1971
3	6	계란으로 할수있는 간단하고 맛있는 3가지 요리~ 3 kinds of egg cook...	195000000	18	30000	86	262820
294	6	건새우 마늘볶음 만들기~ 마늘향이 솔솔~, 짭쪼롬하게, 맛있게~ [강쉪]	195000000	606	13500	7	10432
354	6	부추김치 담그는 법~ 쉽고 간단하고 맛있게 ~[강쉪]	195000000	703	58100	34	93110
912	3	🔥영양 만점 #생선조림 부터 #국물반찬까지 #오늘의식단 🔥 [만개의레시피]	356000000	254	5300	1	3233
188	1	냉동 만두계를 평정하러 나타난 그 제품 리뷰	442000000	294	180000	43	113564

02 | 데이터 수집 및 전처리

- 영상 제목 처리 (토큰화 및 TF-IDF생성)

- 기본적인 불용어 및 특수문자 1차 처리
- my_stopwords 리스트 생성하여 2차 처리
- 처리 된 제목에서 명사들만 저장

	name	clean_name	clean_name_noun
1198	🔥 넌 커리? 난 카레~! 카레랑 밥이랑 ★ 카레라이스 레시피 ★ 🔥 [만개의레시피]	넌 커리 난 카레 카레 밥 이랑 카레라이스	[넌, 커리, 난, 카레, 카레, 밥, 카레라이스]
2306	펑리수 달콤한 파인애플쨈이 쏘옥! [만개의레시피]	펑 리수 달콤하다 파인애플 째다 쏘다 옥	[펑, 리수, 파인애플, 옥]
2963	애호박조개전 ☆ 크으~ 땡긴다 땡겨! [만개의레시피]	애호박 조 개전 크으 땡기다 땡기다	[애호박, 조, 개전]
998	🔥 홍합이 입벌리는 최고의 순간 홍합탕 레시피 🔥 [만개의레시피]	홍합 입 벌리다 최고 순간 홍합 탕	[홍합, 입, 최고, 순간, 홍합, 탕]
2375	미니크루아상 샌드위치 귀엽고맛있어 ... ♡ [만개의레시피]	미니 크루아상 샌드위치 귀엽다 맛있다	[미니, 크루아상, 샌드위치]
1637	🔥 집에서 먹는 중국요리 레시피 ☆ 오늘의식단 🔥 [만개의레시피]	집 먹다 중국요리 오늘 식단	[집, 중국요리, 오늘, 식단]
3034	쪽파손질법 ♡ 야무지게 활용하자! [만개의노하우]	쪽파 손질 법 야무지다 활용 하다 노하우	[쪽파, 손질, 법, 활용, 노하우]
142	당신의 요리실력을 망치는 인터넷 레시피의 문제점	당신 요리실 력 망치다 인터넷 문제점	[당신, 요리실, 망치, 인터넷, 문제점]
3060	돌체라떼 ♥ 연유에 우유까지 [만개의레시피]	돌체 라떼 연유 우유 까지	[돌체, 라떼, 연유, 우유]
149	떡볶이를 천상계로 끌어올리는 셰프의 킥 두가지	떡볶이 천상계 끌다 올리다 셰프 킥 두 가지	[떡볶이, 천상계, 리다, 셰프, 킥, 두, 가지]
2673	생크림머핀 ☆ No버터 폭신폭신♥ [만개의레시피]	생크림 머핀 No 버터 폭 신 폭신	[생크림, 머핀, 버터, 폭, 신, 폭신]
1372	SUB) 유부초밥에 즐겼다면 ? 고소한 크래미유부초밥 ★ [만개의레시피]	유부초밥 즐기다 고소하다 크다 래미 유부초밥	[유부초밥, 래미, 유부초밥]
3486	열무김치 맛있게 담는법 초보자도 하나하나 담을 수 있게 쉬워요 심방골주부	열무김치 맛있다 담다 법 초보자 하나 하나 담 수 있다 쉽다	[열무김치, 법, 초보자, 하나, 하나, 담, 수]
4368	고추장찌개 만들기~ 얼큰하고 진하게~ [강쉪]	고추장 찌개 만들기 얼큰하다 진하다	[고추장, 찌개, 만들기]
613	Sub) 타르트지의 기초 ! 파트슈크레로 베이직한 딸기타르트 만들기 : Strawb...	타르트 지 기초 파트슈크레 베이직 한 딸기 타르트 만들기 Strawberry Tar	[타르트, 기초, 파트, 슈크레, 베이직, 딸기, 타르트, 만들기]

02 | 데이터 수집 및 전처리

- 영상 제목 처리 (토큰화 및 TF-IDF생성)

- TF-IDF 계산하여 각 영상 제목에 들어있는 단어들의 특이한 정도를 확인한 후,
  TF-IDF로 표현된 각 영상의 단어들이 hits(조회수)에 얼마큼 영향을 끼치는 지 알아야 함
- 이를 수치화 시키기 위해 hits(조회수)에 모든 단어들의 impact의 정도(비중)을 찾으려고 함

① TF-IDF 계산하기 위한 과정

모든 영상에 나온 단어 리스트
(27119개)

→

중복 단어 제거 (4131개)
✓ **2번 이상 나온 단어들 (2235개)**
3번 이상 나온 단어들 (1594개)

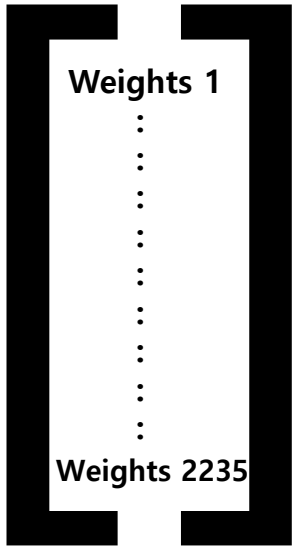
→

영상을 구분할 수 있는
핵심 단어만 남긴 데이터셋

pre_total_word_list	clean_name_most_freq2	clean_name_most_freq3
요리사 자리 모이 면 일 오리 다리 콩피	요리사 자리 모이 면 일 오리 다리	요리사 자리 모이 면 일 오리
요리 남자 집들이 음식	요리 남자 집들이 음식	요리 남자 집들이 음식
종료 제 여러분 요리	종료 제 여러분 요리	제 요리
요리사 자리 모이 면 일	요리사 자리 모이 면 일	요리사 자리 모이 면 일
요리 남자 파티 음식	요리 남자 파티 음식	요리 남자 파티 음식

youtuber_id	subscribers	days_after_upload	hits	clean_name_most_freq2
0	556000000	9	245137	요리사 자리 모이 면 일 오리 다리
0	556000000	30	630196	요리 남자 집들이 음식
0	556000000	43	160997	종료 제 여러분 요리
0	556000000	59	1962881	요리사 자리 모이 면 일
0	556000000	71	720225	요리 남자 파티 음식

중복만 제거하면 너무 큰 차원의 TF-IDF 가 생길 것으로,
3번 이상 나온 단어들을 사용할 경우에는 위에서 보듯
없어지는 단어들이 많아 제목이 구분되지 않을 것으로 판단



02 | 데이터 수집 및 전처리

- 영상 제목 처리 (토큰화 및 TF-IDF생성)

② predict_hits – hits 를 최소화하는 weights (각 단어들이 hits에 영향을 끼치는 impact) 구하기

아래의 식과 opt.minimize를 통하여 weights 를 구함

법	요리	만들기	가지	볶음	...	로그	피오	설렁탕	습격	인기가요
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
7.733684	16.853662	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0
0.000000	8.426831	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0

X

Weights 1
⋮
Weights 2235

+

subscribers
195000000
356000000
195000000
356000000
195000000
195000000
195000000
356000000
442000000
127000000
356000000

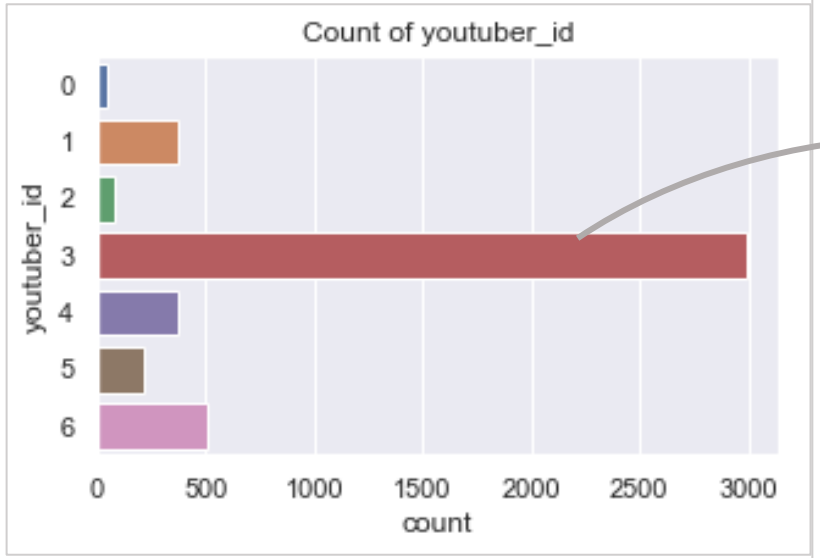
=

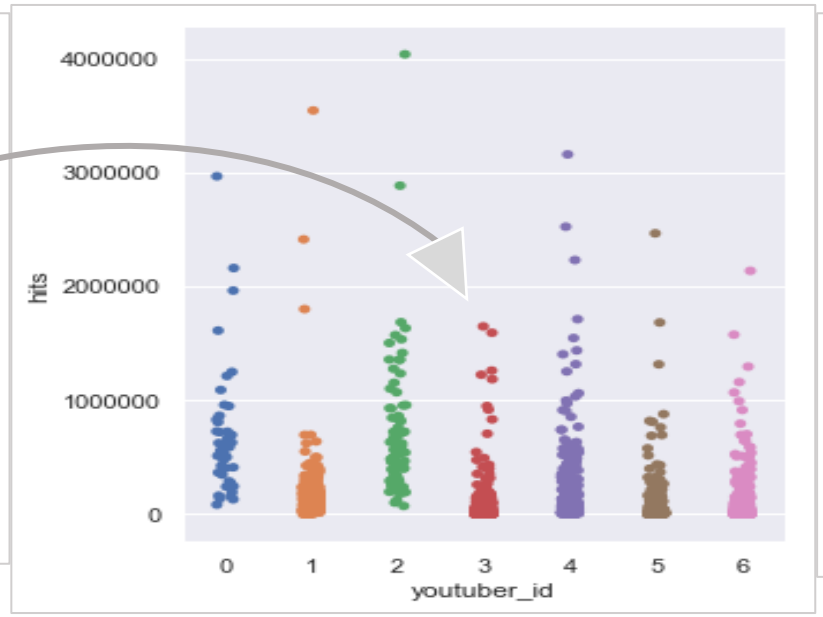
hits
86
57491
182960
1971
262820
10432
93110
3233
113564
2311
12024
528

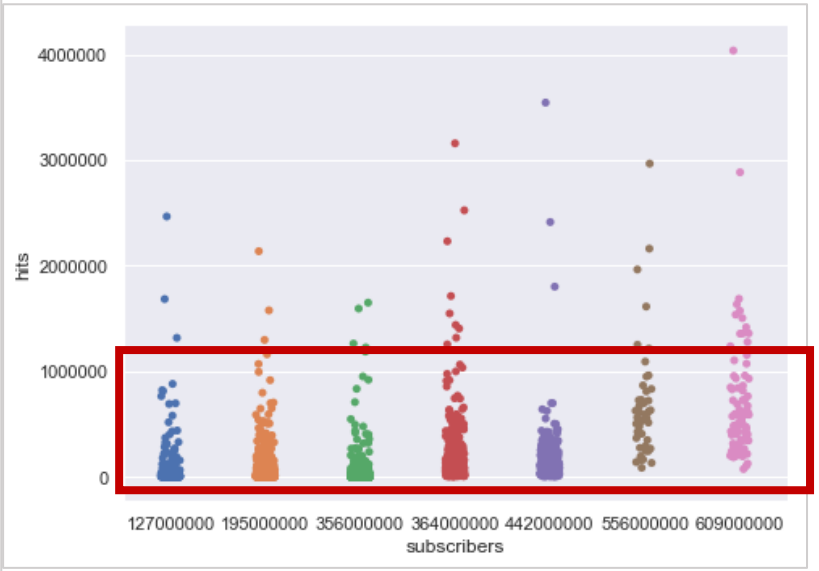
➡ ‘name*weights’ 변수 추가 -> 이를 영상 제목 단어의 특이한 정도와 조회수에 끼치는 임팩트로 생각해 봄

02 | 데이터 수집 및 전처리

- EDA







Youtuber의 영상 개수의
차이가 많이 보임

영상이 많다고,
조회수가 높은 것이 아니다,
-> 키워드가 중요한 것 같다

구독자수와 조회수는
양의 상관관계가 있지만,
대부분의 영상은 1,000,000에
몰려 있다는 것을 볼 수 있다.

03 모델 생성/예측/평가

- 모델 예측 비교 및 평가

03 | 모델 생성/예측/평가

- 모델 예측 비교 및 평가

모델 / 예측값	OLS	GBoosting	XGBoost	RF	LASSO(0.1)
Test	0.310	0.358	0.355	0.358 (mse)	0.309
Training	0.315	0.562	0.581	0.421	0.301

- 전반적으로 낮은 예측력을 보이나, Training 예측값에서 50% 이상은 예측할 수 있다는 것을 확인
  -> 만약 데이터를 10,000개 이상으로 구성했을 시, 보다 나은 예측력을 기대할 수 있을 것

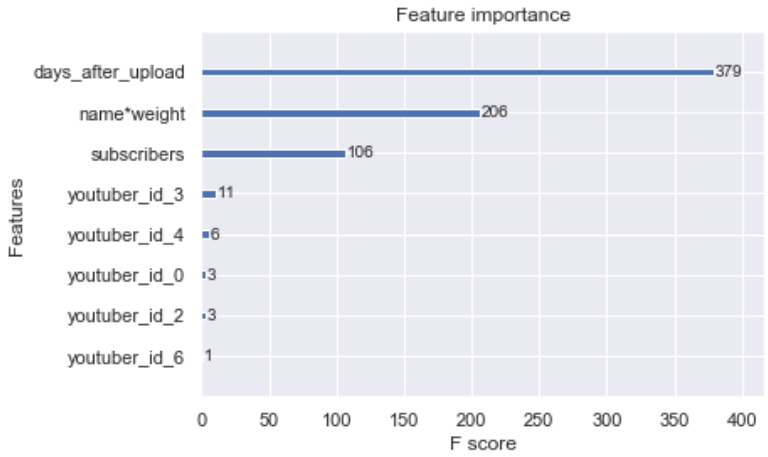
- 가장 높은 예측력을 보이는 모델은 그래디언트 부스팅과 랜덤포레스트이며,
  정확한 **predict**가 필요한 목적인 경우, 이 모델을 사용하는 것이 적절

- 하지만, 일반 OLS모델을 사용하는 경우에도 비슷한 정도의 예측력을 보임
  -> **Inference**가 중요하다면 **OLS 모델**을 사용하는 것도 적합해 보임

03 | 모델 생성/예측/평가

- 모델 예측 비교 및 평가

XGBoosting



GBoosting

	Feature	Importance
0	subscribers	0.501
1	days_after_upload	0.246
2	name*weight	0.088
6	youtuber_id_3	0.077
7	youtuber_id_4	0.063
9	youtuber_id_6	0.018
3	youtuber_id_0	0.004
5	youtuber_id_2	0.002
4	youtuber_id_1	0.001
8	youtuber_id_5	0.001

- 'name*weight' 와 다른 변수들 간의 Importance 비중의 차이는 크다고 볼 수 있음
- 하지만, 'name*weight'의 값을 수정해 나간다면 모델 예측력에 조금 더 큰 기여를 할 수 있을 것으로 생각

  -> 영상 제목 안에 들어가는 단어들을 몇 차례 더 정제가 필요함

  ex) 중복 되는 단어들 삭제하여, 한 영상 안에 'name*weight'가 잘못 계산된 것으로도 판단함

03 | 모델 생성/예측/평가

- 모델 예측 비교 및 평가

요리 동영상 핵심 표현 키워드 Top 100

요리와 함께 활용 가능한 키워드 Top 100

- 또한 영상 제목 안에 명사만 남기는 것이 아닌, top 30 부사 혹은 형용사를 포함하여야 더욱 정확하게 각 영상 제목의 특징을 잡을 수 있을 것으로 생각함

  ex) 유튜브 검색을 할 때 '건강한', '간단한' 등의 형용사들도 많이 사용한다는 것을 놓침

04 Word Cloud

- 영상 제목 형태소/품사 분석
- Word Cloud 형성

04 | Word Cloud

- 영상 제목 형태소/품사 분석

▪ name(영상 제목)을 형태소/품사 분석 -> 총 39946개로 분리됨

형태소

Noun	26160
Verb	4033
Alpha	3463
Adjective	2119
Josa	822
Adverb	506
Exclamation	155
Modifier	80
Suffix	41
Foreign	39
Punctuation	25
Conjunction	12
Eomi	10
Determiner	4
VerbPrefix	3
PreEomi	1

품사

법	647
만들다	584
다	427
한	342
가지	340
먹다	258
간단하다	247
볶음	242
밥	224
맛	206

dtype: int64

부라자	1
소갈비	1
어벤져스	1
케어	1
아줌마	1
예술인	1
에그인헬	1

ex)

	0	1
0	들	Verb
1	자리	Noun
2	모이	Noun
3	면	Noun
4	벌어지다	Verb
5	일	Noun

	0	1
0	보쌈	Noun
1	삼겹살	Noun
2	수육	Noun
3	굴	Noun
4	무생채	Noun
5	황금	Noun
6	Pork	Alpha
7	wraps	Alpha
8	oyster	Alpha
9	kimchi	Alpha

04 | Word Cloud

- Word Cloud 형성

▪ name(영상 제목) -> Word Cloud -> 인기(최다사용) 요리 컨텐츠/영어 단어 컨텐츠 등 찾아보기

04 | Word Cloud

- Word Cloud 형성

▪ name(영상 제목) -> Word Cloud -> 요리 영상 제작 시 활용 가능한 표현 찾아보기

요리와 함께 활용 가능한 키워드 Top 100

요리 동영상 핵심 표현 키워드 Top 100

05 Lesson Learn

- 내 스스로 생각해낸 것(Originality)
- 과제를 통해 배운점

05 | Lesson Learn

- 내 스스로 생각해낸 것(Originality) / 과제를 통해 배운점

- 텍스트 데이터가 들어있는 변수와 다른 변수들을 종합하여 무언가를 예측/분류 해보기 위한 시도
- 텍스트 데이터를 벡터화 혹은 수치화를 시켜야 할 때, 다른 영향 변수들과의 관계를 고려해보려고 함
- 임의의 초기 weight를 시작으로, (실제 값 – 예측 값)을 minimize하는 weight를 활용하고자 함 (그러나, 이 방법이 맞을 지는 모르겠습니다..)

- 혼자 처음부터 프로젝트를 구상하는 것을 시작으로 직접 데이터셋 구성, 데이터 전처리 및 변수 변형, 모델 생성/예측/평가를 해볼 수 있는 기회가 되었습니다. 막연하게 궁금했던 일상의 작은 호기심을 직접 분석해보고 나아가 예측할 수 있는 영역까지 직접 해볼 수 있던 것이 즐거웠습니다. 기획했던 것들을 해 나아가는 과정에서 많은 에러가 발생하여 시간이 오래 걸렸지만, 하나 하나 검색하며 고쳐 나가서 마무리할 수 있던 것 역시 정말 좋은 경험이 된 것 같습니다. 감사합니다!

THANK YOU