tumblyug

크라우드 펀딩 성공 요인 분석 용사 성공 확률 예측

목차

About Project

팀원 소개 기획 의도

Data Collects

성공 데이터 수집실패 데이터 수집

2 Data Preprocessing

필요 데이터 추출 파생 변수 추가

3 EDA

성공 데이터 형태소 분석 평균 달성률 카테고리 비율 평균 제목, 인트로 글자수 목표 금액 4 Data Analysis

변수 별 상관 관계

Machine Learning
Algorithm

6 Conclusion

7 After the project

Project Outline

STEP 1.기획 및 설계

주제 선정/ 요구 사항 분석

Dataset 수집 계획 수립

업무 분할

STEP 2. Data 전처리

데이터 전처리

주요 변수 기술 통계

파생 변수 생성

변수 선택과 모델링

STEP 3. 모델링

모델 생성 / 평가

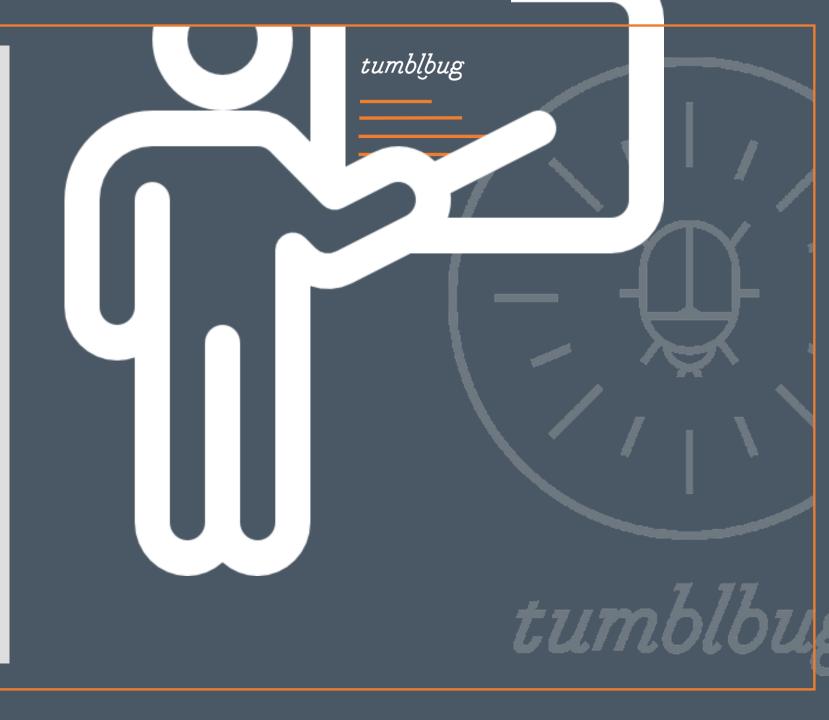
인사이트 도출

결과물 구현

About Project

Team Profile

Project goal



About Project

Team Profile



김정*

- Crawling
- Preprocessing
- Modeling



기윤*

- Crawling
- Preprocessing
- Modeling
- Prep for ppt



황겸*

- Crawling
- Preprocessing
- Modeling
- Github managing



두삼*

- Crawling
- Preprocessing
- Modeling
- Presentation



배주*

- Crawling
- Preprocessing
- Modeling
- Prep for ppt



유슴*

- Crawling
- Preprocessing
- · Modeling
- Presentation

About Project

Project Aim

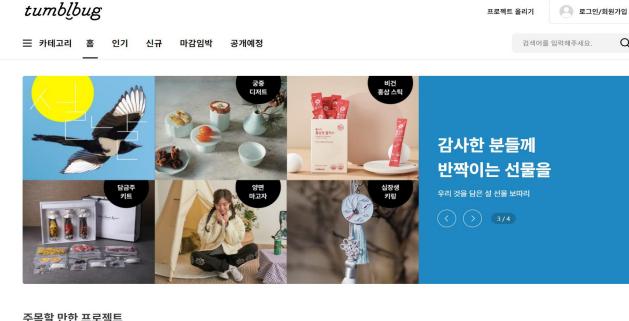
펀딩이란?

사업적 목적의 사람들에게

자금조달 & 구매/투자 수단

크라우드 펀딩 프로젝트 성공 요인 파악 # 이후 진행 프로젝트 성공 가능성 예측

텀블벅 (tumblbug)



주목할 만한 프로젝트



새하얀 눈속에서 핀 달콤하고 황홀한



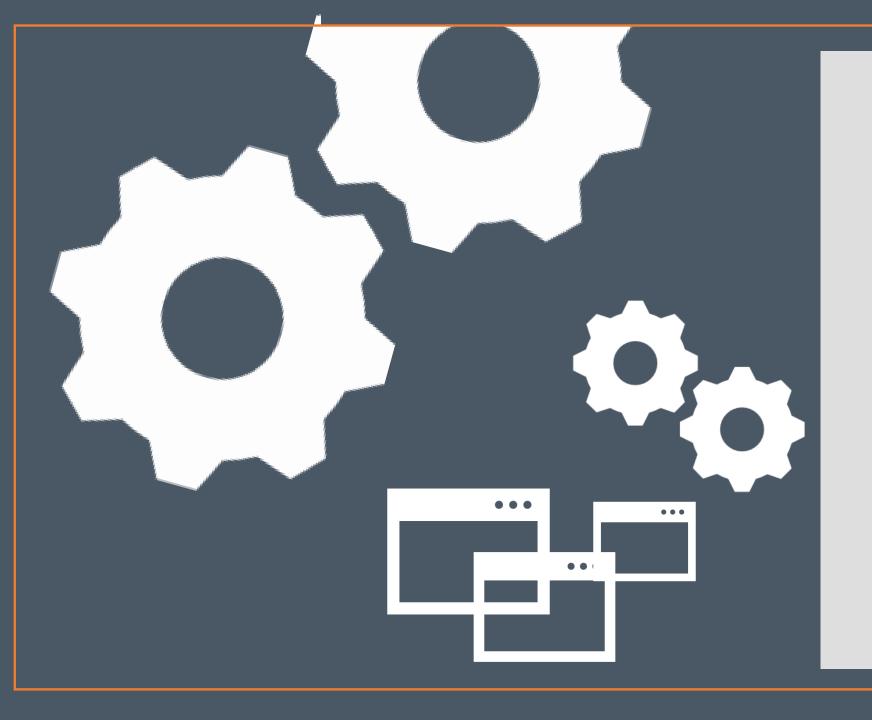
당신의 운명을 정하는 [YES/NO] 코인



보기에도, 읽기에도 좋은 도서용지



보일 첫 정규 앨범 <나쁜 마음> 제작



Data Crawling

Selenium

1. 달성율 100%이상의 성공 프로젝트 데이터 수집



이누야샤 OST 앨범 재발매 프로젝트

음악 | (주)뮤직파라디소

그 때 그 시절 그 노래 <이누야샤>의 음악을 다시 만듭니다.

476,850,391원 476%

① 38일 남음



따스한 별꽃이 반짝반짝 내려앉은 나무 향기<베르가못 향수>

향수·뷰티 | Nar.bY

베르가못 나무 한 그루를 통째로 녹여낸, 달콤하게 따뜻한 겨울용 시트러스 [동네꽃집]

5,804,000원 1160%

① 9일 남음



새하얀 눈속에서 핀 달콤하고 황홀한 향기 - 설향딸기 향수

향수·뷰티 | NATURFUM

새콤달콤하고 맑은 과즙이 듬뿍! 배어 나오는 설향딸기 향수로 일상을 달콤하게 물들여보세요.

52,012,100원 10402%

① 2일 남음

feature

- 이동 URL
- 제목
- 카테고리



Selenium

1. 달성율 100%이상의 성공 프로젝트 데이터 수집

```
prev_height = driver.execute_script("return document.body.scrollHeight")
 3 While True:
       # 스크롤을 화면 가장 아래로 내린다
       driver.execute_script("window.scrollTo(0,document.body.scrollHeight)")
       # 페이지 로딩 대기
       time.sleep(4)
9
10
       # 현재 문서 높이를 가져와서 저장
11
       curr_height = driver.execute_script("return document.body.scrollHeight")
12
13
       if (curr_height == prev_height):
14
          break
15
       else:
16
          prev_height = driver.execute_script("return document.body.scrollHeight")
17
19 html_source = driver.page_source
20 driver.close()
   soup = BeautifulSoup(html_source, 'html.parser')
```

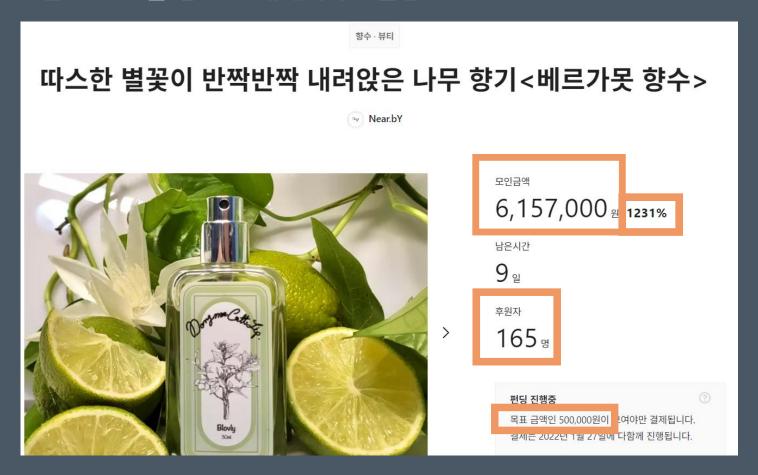
	title	category	url
0	프리랜서 작가로 한 달에 400~1000만 원 버는 법	출판	https://tumblbug.com/aronsong5
1	기독교 대천사부터 본디에이의 천사까지, <천 사소개록>	출판	https://tumblbug.com/angelus
2	데뷔 못 하면 죽는 병 걸림 1차 공식 굿즈	출판	https://tumblbug.com/testar_official_1st_goods
3	500가지 기호와 상징 <symbols> 심볼 사전</symbols>	출판	https://tumblbug.com/symbols300
4	월간커피, 군밤과 군고구마 같은 겨울 간식 콜 드브루 커피	푸드	https://tumblbug.com/brewsbrothers49

스크롤 다운 자동화

데이터 프레임에 저장

Beautiful Soup

2. 수집한 URL을 통해 상세페이지 크롤림



feature

- 이동 URL
- 제목
- 카테고리
- 모인 금액
- 달섬%
- 후원자 수
- 목표 금액

하나의 데이터 셋으로 병합

3. 기존 데이터 프레임과 새로 수집한 데이터 조인

url	category	title	
un	category	uue	
https://tumblbug.com/aronsong5	출판	프리랜서 작가로 한 달에 400~1000만 원 버는 법	0
https://tumblbug.com/angelus	출판	기독교 대천사부터 본디에이의 천사까지, <천 사소개록>	1
https://tumblbug.com/testar_official_1st_goods	출판	데뷔 못 하면 죽는 병 걸림 1차 공식 굿즈	2
https://tumblbug.com/symbols300	출판	500가지 기호와 상징 <symbols> 심볼 사전</symbols>	3
https://tumblbug.com/brewsbrothers49	푸드	월간커피, 군밤과 군고구마 같은 겨울 간식 콜 드브루 커피	4



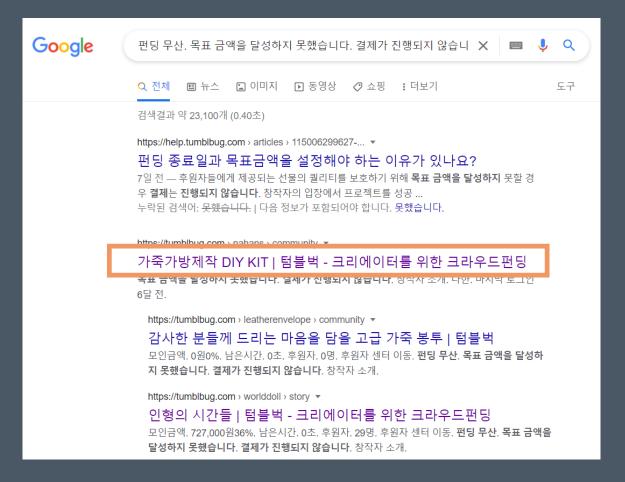
	상품명	현재 펀딩 금액	펀딩 퍼센트	목표금액	후원자 수
0	미니 난로 오브제로 경험하는 작고 따뜻한 불멍,향멍의 순간	9,143,000	914	배송 완료	131명
1	[웹툰 스케치업 배경] pc방과 각종 소품들!	4,570,000	914	500,000	244명
2	아프리카의 아름다움을 나누다 올로이 백 (통가죽 사파리백)	27,418,000	913	배송 완료	86명
3	돌아온 냥투 친구, 멍멍이 디자인화투 '멍투'	18,266,500	913	배송 완료	853명
4	사랑을 다루는 태도에 대한 단편집 <나는 널>	4,565,000	913	배송 완료	303명

	상품명	현재 펀딩 금 액	펀딩 퍼센트	목표금액	후원자 수	category	intro	url
0	프리랜서 작가로 한 달에 400~1000만 원 버는 법	21,305,777	4261	500,000	611명	출판	프리랜서 작가 실전 노하우와 팁 / 모든 글쓰기 분야 작법 설 명 / 정확한 페이 계산	https://tumblbug.com/aronsong5
1	기독교 대천사부터 본디에 이의 천사까지, <천사소개 록>	81,663,900	927	8,800,000	2,637 명	출판	기독교, 유대교 전설, 이슬람 천사들과 '본디에이의 천사 들'이라 불리는 부두교 로아	https://tumblbug.com/angelus
2	데뷔 못 하면 죽는 병 걸림 1차 공식 굿즈	423,791,000	4237	10,000,000	13,866 명	출판	[KW북스] 백덕수 작가님의 '데 뷔 못 하면 죽는 병 걸림' 첫 번 째 공식굿즈, 마	https://tumblbug.com/testar_official_1st_goods
3	500가지 기호와 상징 <symbols> 심볼 사전</symbols>	158,012,700	1053	15,000,000	4,257 명	출판	수천 년의 인류 역사에 걸쳐 500개가 넘는 기호, 상징\n고대 천문학부터 마법 악	https://tumblbug.com/symbols300
4	월간커피, 군밤과 군고구마 같은 겨울 간식 콜드브루 커피	6,730,000	1346	500,000	306명	푸드	'월간커피'라는 이름으로 매달 새로운 싱글 오리진 콜드브루 커피를 한정수량으로 선보	https://tumblbug.com/brewsbrothers49

실패 프로젝트

Google 검색을 통해 수집

1. 상세 페이지 URL 수집



텀블벅 페이지에서 실패 프로젝트를 보여주지 않음

→ Google에 〈펀딩 무산. 목표 금액을 달성하지 못했습니다. 결제가 진행되지 않습니다.〉 검색

> feature - 이동 URL

실패 프로젝트

기존 크롤러와 연동

2. 수집한 URL을 통해 상세페이지 크롤링



feature

- 이동 URL
- 제목
- 카테고리
- 모인 금액
- 달섬%
- 후원자 수
- - 목표 금액

Dataset Preprocessing



필요한 데이터만 추출

1. 중복 수집된 데이터 삭제

```
1 df1 = df1.drop_duplicates(['상품명'])
```

2. 성공 프로젝트에서 진행 중인 프로젝트 추출

```
1 df_INNER_JOIN = pd.merge(df1_cut, df2_cut, left_on='상품명', right_on='상품명', how='inner')
2 df_INNER_JOIN
```

3. <u>진행 중인 프로젝트를 제외한</u> 성공 프로젝트만 사용

```
1 list_of_single_column = list(df_INNER_JOIN['상품명'])
2 list_of_single_column
3
4 df_final = df1[~df1['상품명'].isin(list_of_single_column)]
5 df_final
```

결측치 대체

4. '목표 금액'의 실패 프로젝트 부분 결측치 대체

	title	category	current_funding_amount	percentage	target_funding_amount	supporter
index						
0	짐볼로 만든 인 형, '돌리볼리'	캐릭터 · 굿즈	329000	8	4112500	6
1	함께 만드는 약 기 전문 웹진, <the strings=""></the>	출판	701000	14	5007100	27
2	감사한 분들께 드리는 마음을 담을 고급 가죽 봉투	디자인 문 구	0	0	0	0
3	마다가스카르 사진전	사진	2322000	38	6110500	54
4	인형의 시간들	출판	727000	36	2019400	29
9705	메카피아의 한 국 전통 자개 문 양 피젯 스피너	홈·리빙	1730000	173	10000	27
9706	카프카 <변신> 초판본과 독서 카드	출판	1730000	173	10000	114
9707	낯선 첫발을 내 딛는 이방인을 위한 <연남동 작 은 방>	출판	1730000	173	10000	54
9708	함께 행복하개! 유기견 팔찌	캐릭터 · 굿즈	865000	173	5000	45
9709	어버이날선물 한번에 준비하 기	홈·리빙	1730000	173	1000000	77

- 목표 금액(target_funding_amount)

: null → (모인 금액 / 달섬%) * 100

파생 변수 추가

5. 펀딩 성공여부, 제목 길이 변수 생성

г	title	category	current_funding_amount	percentage	target_funding_amount	supporter	result	title_len
index								
0	짐볼로 만든 인 형, '돌리볼리'	캐릭터 · 굿즈	329000	8	4112500	6	0	18
1	함께 만드는 약 기 전문 웹진, <the strings=""></the>	출판	701000	14	5007100	27	0	31
2	감사한 분들께 드리는 마음을 담을 고급 가죽 봉투	디자인 문 구	0	0	0	0	0	28
3	마다가스카르 사진전	사진	2322000	38	6110500	54	0	11
4	인형의 시간들	출판	727000	36	2019400	29	0	8
9705	메카피아의 한 국 전통 자개 문 양 피젯 스피너	홈·리빙	1730000	173	10000	27	1	25
9706	카프카 <변신> 초판본과 독서 카드	출판	1730000	173	10000	114	1	19
9707	낯선 첫발을 내 딛는 이방인을 위한 <연남동 작 은 방>	출판	1730000	173	10000	54	1	30
9708	함께 행복하개! 유기견 팔찌	캐릭터 · 굿즈	865000	173	5000	45	1	16
9709	어버이날선물 한번에 준비하 기	홈·리빙	1730000	173	1000000	77	1	16

- 펀딩 성공여부(result)

: 성공 (1) / 실패(0)

- 제목 길이(title_len)

파생 변수 추가

6. 카테고리 레이블링 후 변수로 추가

인코딩 클래스: ['공연' '디자인' '디자인 문구' '디지털 게임' '보드게임 TRPG' '사진' '영화 비디오' '예술' '웹툰 만화' '웹툰 리소스' '음악' '의류' '잡화' '주얼리' '출판' '캐릭터 굿즈' '테크 가전' '푸드' '향수 뷰티' '홈 리빙']



민코딩 변환값: [9 8 6 14 19 15 11 3 2 4 12 1 18 10 17 0 5 7 13 16]

데이터 정규화

5. 점규화 진행

변수들의 최솟값

current_funding_amount 0
percentage 0
target_funding_amount 0
supporter 0
result 0
title_len 8
category_label 0

변수들의 최댓값

current_funding_amount 347353000
percentage 12035
target_funding_amount 389400000
supporter 5883
result 1
title_len 33
category_label 19



변수들의 최솟값

current_funding_amount 0.0
percentage 0.0
target_funding_amount 0.0
supporter 0.0
result 0.0
title_len 0.0
category_label 0.0

변수들의 최댓값

current_funding_amount 1.0
percentage 1.0
target_funding_amount 1.0
supporter 1.0
result 1.0
title_len 1.0
category_label 1.0



Data EDA

성공 프로젝트 제목

1. 성공 프로젝트의 제목을 txt파일로 변환

	title
index	
149	웹툰과 일러스트를 위한 꽃브러쉬
150	오디오웹툰으로 만나는 <고래별_경성의 인어공 주> 시즌2
151	꼬마마법사 레미 공식 한정판 굿즈
152	만화가의 금융에세이 < 자산소득으로 생활하기 프로젝트>
153	당신을 지켜주는 영물, '메롱해치 자수 KIT'



'웹툰과 일러스트를 위한 꽃브러쉬,오디오웹툰으로 만나는 <고래별_경성의 인어공주> 시즌2,꼬마마법사 레미 공식 한정판 굿즈,만화가의 금융 에세이 <자산소득으로 생활하기 프로젝트>,당신을 지켜주는 영물, ₩'메롱해치 자수 KIT₩',파댕이가 돌아와따 ! [흑우 파댕이 인형],코디 걱정 없이 편하게, #'나봄 저고리 셔츠&허리치마#',[스케치업] 로판 일러스트를 위한 평민들의 거리,죽음을 연습하는 200가지 질문 - 「질문의 죽 |음.,,로맨스 판타지 웹툰에 도움이 되는 엑스트라 브러쉬,혼자하는 담력시험, 공포게임 「Last Light」,[AAF] 50인의 아티스트가 해석한 <식물 한작 아트북>,오스카 와일드 <행복한 왕자> 1888년 초판 복원,변호사가 알려주는 성폭력 피해자를 위한 법률가이드,웹툰, 일러스트를 위한 100 【종류의 다양한 문양들,[지갑송 장편소설] <소설 속 엑스트라 큐브 편> 애장판,매일 3분 라인이 살아나는 습관,내 손으로 하이드를 깨우다, 지 |킬 투 하이드 잉크,목격자의 진술로 사건을 해결하는 추리 게임 [미제 사건],[CachiBag]물과 햇빛 그리고 선인장으로 빚은 가치백,시간을 담은 포스터 만화책 <MEMO> 촐간,식물이 있는 우리, 타이페이식물원,길몽 팔아서 1억을 번다고요? 한국형 판타지 소설 <옥토>,우주만큼 복잡한 내 감정 찾기, 코스믹 마인드 오로라 카드,영롱한 펄과 함께 동화 속으로 초대할게요, 밀크웨이 입욕제,혹시 나도 탈모인가? 걱정한적 있다면! 헤 막안티헤어로스샴푸,스케치업 해안마을과 소방차, 이삿짐차,지구 멸망을 앞둔 두 여자의 사랑 소설 <멸망의 감정>,66100 2020 브라렛,30종의 비인간 존재가 되어보는 <상괭이와 춤을> 키트,맥베스의 향기를 찾았다 <넵튠과 아포리아>,[RE:PLA] 쓰레기도 줄여주고 플라스틱 업사이클까 |지??,웹툰&일러스트에 활용하기 좋은 <헤일로(후광)>소재,나의 졸업을 막으려는 누군가가 있다? [공포의 조별과제],대체 불가한 독보적인 컬러 의 아티스트 프롬 LP 발매,내년의 우린 좀 더 평안하고 행복하길 <차와향 리추얼키트>,사랑을 다룬 단편선, <기묘한 사랑에 관한 보고서>,강원 의 겨울산 모습을 닮은 세미 커스터마이징 가방, 산산백,서사의 주인공을 맡아보세요. 짙은 여운을 머금은 북퍼품,녹차 없는 녹차는 그만! 설 탕 없이 달달한 달콤 아몬드녹차,따듯한 햇살의 포근함을 드려요 [안전한 섬유 탈취제],폐플라스틱의 재탄생! 따뜻한 플라삭스,윌리엄 모리 스가 말하는 「아름다움을 만드는 일」,서울대생이 만든 공부습관 형성 플래너 & 뉴스레터,인생의 향기가 궁금할 때 <6가지 여정의 향>,극단 ₩'공백₩' 2021 창작뮤지컬 [레테의 정원],믿을 수 있는 건 무엇인가요?, 단편영화 <침묵>,지구돌봄을 위한 유쾌한 탐험! 생태놀이판 <토킹어스 >,할머니 할아버지를 위한 놀이책 <월간 실버>,달빛을 머금은 도자기, 월백 균요,연필스케치 컬러링 북 ₩'널 위해 핀 꽃₩',[시간 탐험집] 나와 당신의 서른 즈음에,크레용으로 그린, 따스한 공간이 담긴 2021년 패브릭달력,2022 나만의 색으로 채우는 ColorFill Diary,Band ₩'WILLY₩' 첫

토큰화, 품사 태김

2. Konlpy를 이용하여 형태소 토큰화

```
((',', 'Punctuation'), 8266),
(('의', 'Josa'), 2602),
(('을', 'Josa'), 1856),
(('ć', 'Punctuation'), 1656),
(('"", 'Punctuation'), 1480),
(('ò,', 'Punctuation'), 1141),
(('한', 'Josa'), 875),
(('를', 'Josa'), 850),
((']', 'Punctuation'), 789),
(('에', 'Josa'), 679),
(('에', 'Josa'), 663),
(("",", 'Punctuation'), 663),
(("",", 'Punctuation'), 663),
(('에', 'Josa'), 639),
(('웹툰', 'Noun'), 630),
```

3. 품사별로 가장 많이 등장한 상위 1500개 선별

형용사 [('함께', 173), ('품', 70), ('다', 55), ('없이', 37), ('우', 23), ('같이', 19), ('딱', 14), ('암', 12), ('언제나', 12), ('어느', 12), ('그만', 11)]

동사 [('하다', 1151), ('담다', 400), ('만들다', 295), ('되다', 176), ('보다', 123), ('많다', 122), ('오다', 105), ('만나다', 93), ('매다', 92), ('싶다', 91), ('타다', 81), ('나르다', 80), ('해주다', 72), ('찾다', 67), ('가다', 67), ('쓰다', 64), ('먹다', 62), ('즐기다', 62), ('추다', 60), ('지키다', 59), ('알다', 59), ('크다', 57), ('떠나다', 51), ('돌아오다', 49), ('쉬다', 49),

명사 [('위', 663), ('웹툰', 630), ('스케치업', 516), ('배경', 433), ('당신', 343), ('나', 322), ('뱃지', 298), ('집', 284), ('이야기', 280), ('우리', 276), ('프로젝트', 269), ('세트', 243), ('내', 205), ('고양이', 194), ('북', 190), ('여행', 175), ('여성', 175),

7点 [(',', 8266), ('<', 1656), (""", 1480), ('>,', 1141), (']', 789), ("',", 663), ('-', 617), (':', 555), ('>', 552), ('!', 550), ('.', 454), (',[', 439), ('&', 420), ("", 408), ('[', 398), ('!,', 348), ('],', 293), ('(', 278), (

WordCloud로 성공 키워드 시각화

4. NLTK를 이용하여 빈도수 상위 50개까지의 형용사, 동사, 기호 표시



만나는 잃어버린꿈꾸는 끝내는지친 비 쉽게 따라 보는 하여 뜨겁고자 할 했던 된 보는 기를 하여 다 지쳐주는 이어 유지 이는 어린하며 보세요 되는 이어 유지 이는 어린하며 보세요 되는 이어 유지 이는 어린하며 보세요 되는 수 있는 이는 그는 사는 이 그를 하지 되어 그 그 이 그는 나는 이 그를 어려면 보세요 되는 사는 이 그를 어려면 되는 사는 이 기후 아서 사는 는 사는 이 기후 아서 사는 는 사는 이 기후 알려주는 분 보는 심어 오르 이는 이 하게 들 알려주는 분 보는 심어 오르 이는 이 하게 들 알려주는 분 보는 심어 오르 이는 이 하게 들 알려주는 분 보는 심어 오르 이는 이 하게 들 알려주는 분 보는 심어 오르 이는 이 하게 들 알려주는 분 보는 심어 오르 이는 이 하게 들 알려주는 분 보는 심어 오르 이 하게 들 알려주는 보는 심어 오르 이 하게 들 알려주는 분 보는 심어 오르 이 하게 들 알려주는 보는 심어 오르 이 하게 들 알려주는 보는 심어 오르 이 하게 들 알려주는 보는 심어 오르 이 나를 보는 심어 오르 이를 보고 있다.

형용사

동사

WordCloud로 성공 키워드 시각화

4. 명사는 불용어 처리 후 상위 50개까지 표시

```
1 stop_words = ["제", "월", "일", "조", "때", "그", "이", "바", "및", "안", "정", "위", "관", "수", "프로젝트", "내", "종", "속", "를"
2 token_ko = [word for word in token_ko if word not in stop_words]
```

WordCloud로 성공 키워드 시각화

4. 명사는 불용어 처리 후 상위 50개까지 표시

우리, 여행, 집, 고양이, 커피와 같이 <u>개인적인 가치를</u> 반영한 제품들이 많음

웹툰, 스케치업, 이야기, 달력, 일러스트, 북 등 <u>참작 예술품</u>들의 펀딩이 성공적임을 확인

WordCloud로 성공 키워드 시각화

<u> 4. 명사는 불용어 처리 후 상위 50개까지 표시</u>

시간꽃커피 책사람 향기봄 원 분 다이첫 역성 배 이 여성 배 이 시간 고양이첫 선물 기계임 보이 고양이첫 보이 가장 선물 기계임 보이 되는 다른 지 시장 기억 향수 우리 제작 세트 일러스트 판타지

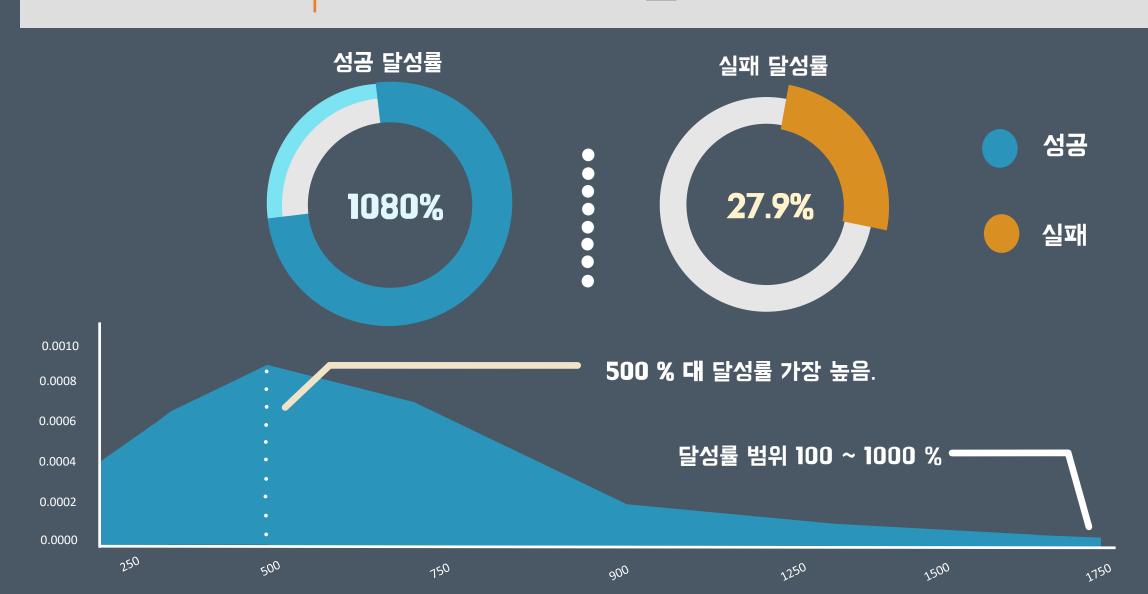
우리, 여행, 집, 고양이, 커피와 같이 <u>개인적인 가치를</u> 반영한 제품들이 많음

웹툰, 스케치업, 이야기, 달력, 일러스트, 북 등 <u>참작 예술품</u>들의 펀딩이 성공적임을 확인

코로나 이후 모든 게 집에서 이뤄지는 **'홈코노미 현상'** 반영

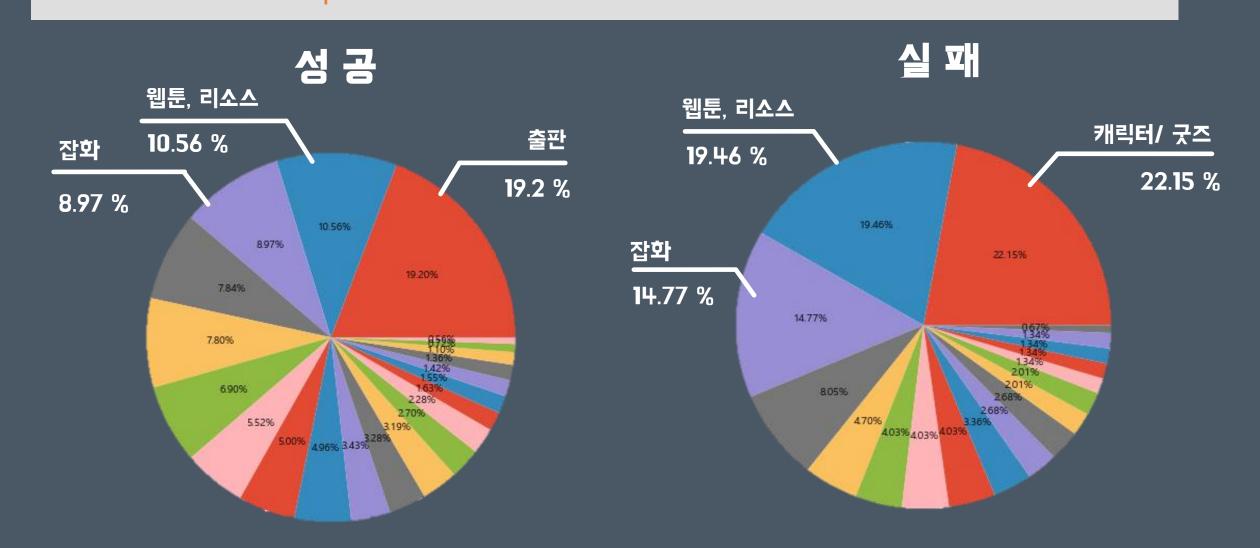
EDA

평균 달성률



EDA

카테고리



EDA

카테고리

섬공



출판

>



웹툰, 리소스

>



잡화

실패



캐릭터/ 굿즈



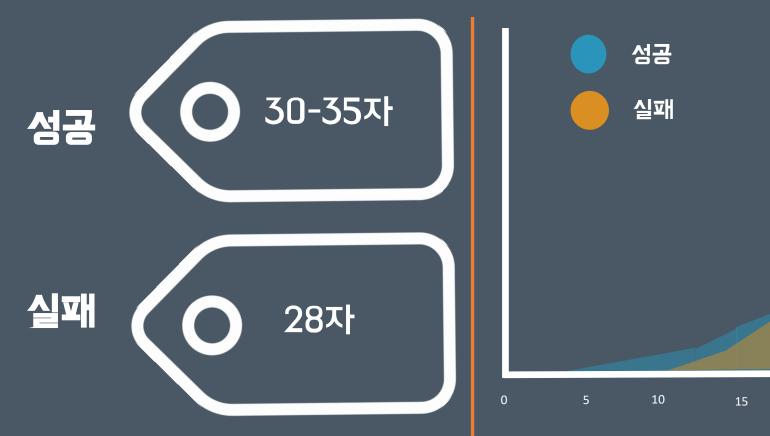
잡화

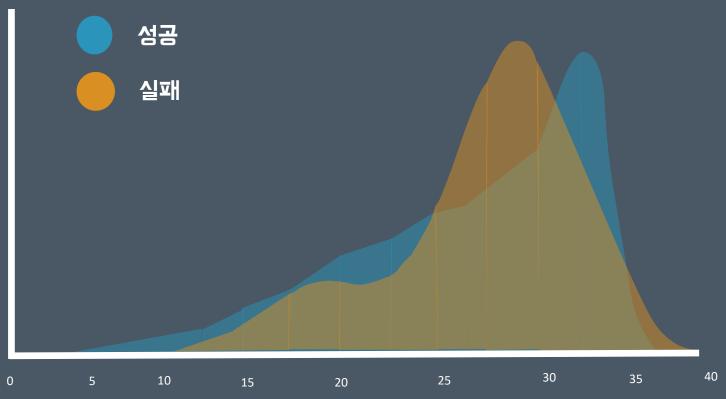


웹툰, 리소스

EDA

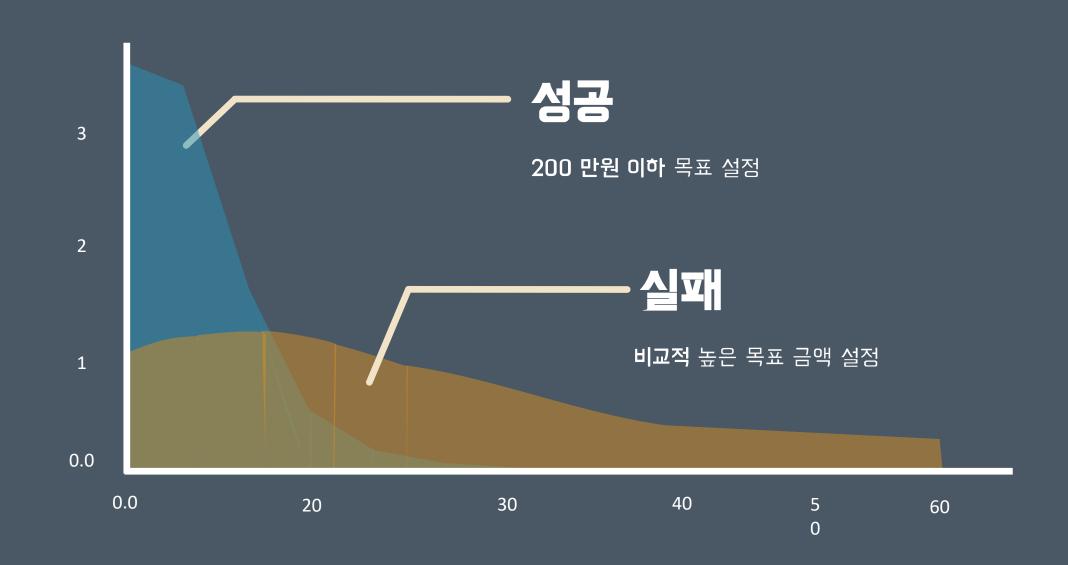
제목 글자 수



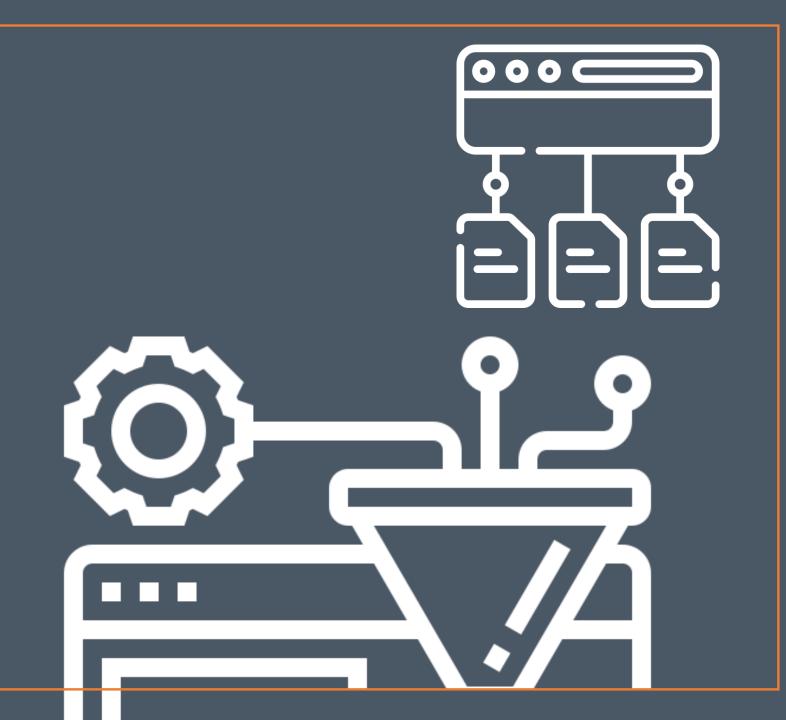


EDA

목표 금액



Data Modeling



독립 변수

목표 금액

카테고리

제목 길이

종속 변수

모인 금액

달섬 %

후원자 수

펀딩 성공 여부

통계 분석

독립변수 - 목표 금액 종속변수 - 펀딩 성공 여부

삼관 분석

스피어만

correlation coefficient: -0.1658

p-value: 1.04773e-61

켄달타우

correlation coefficient: -0.1375

p-value: 6.8882e-61

목표 금액과 펀딩 성공 여부는 매우 약한 음(-)의 <u>상관관계</u>

단일 선형 회귀 분석

R-squared: 0.014549

모형이 데이터를 잘 설명하지 못함

통계 분석

독립변수 - 목표 금액 종속변수 - 달성 %

삼관 분석

스피어만

correlation coefficient: 0.0279

p-value: 0.005

켄달타우

correlation coefficient: 0.0156

p-value: 0.02209

목표 금액과 달성 % 매우 약한 <u>양(+)의 상관관계</u>

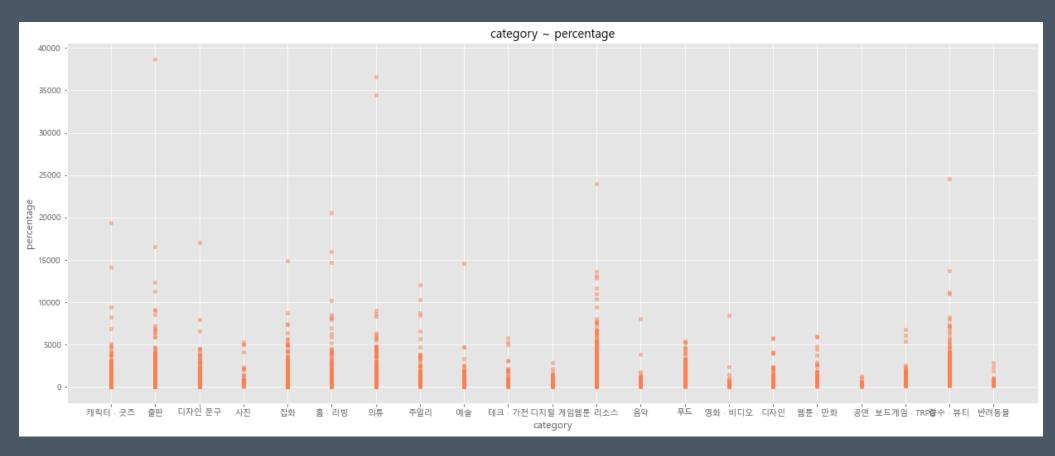
단일 선형 회귀 분석

R-squared: -0.0005

모형이 데이터를 잘 설명하지 못함

통계 분석

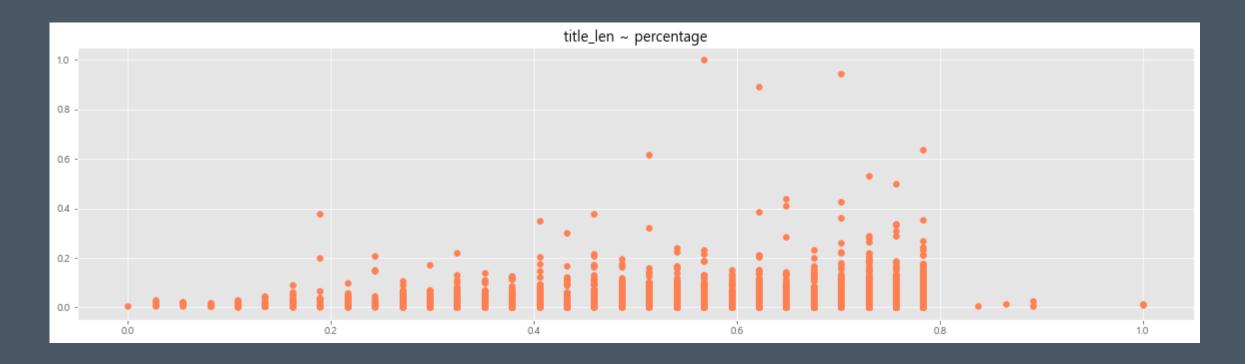
독립변수 - 카테고리 종속변수 - 달성 %



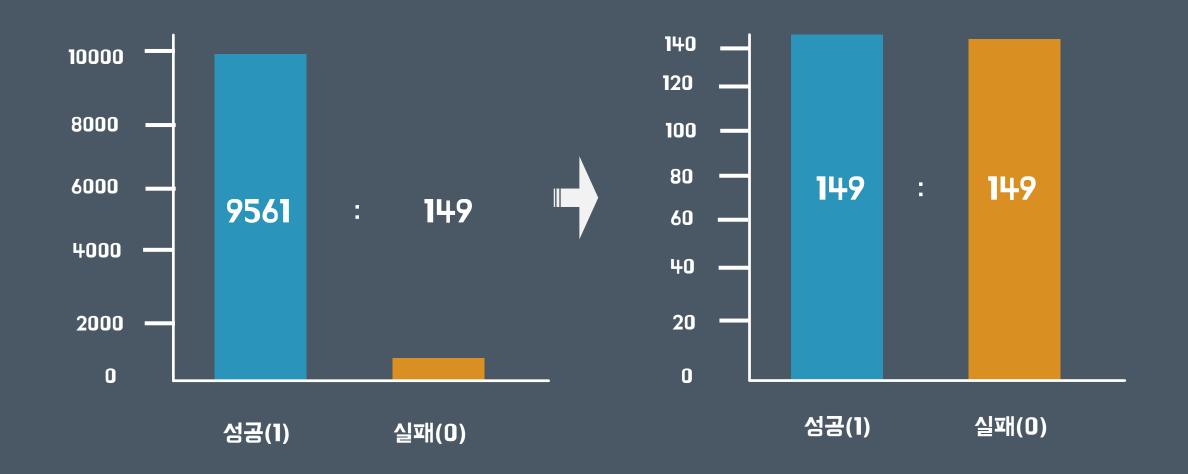
출판 , 웹툰 리소스, 잡화 등 특정 카테고리가 높은 펀딩 %를 달섬하는 겸햠이 있지만 성공/실패에 유의미한 영향을 미치지는 않음

통계 분석

독립변수 - 제목 길이 종속변수 - 달성 %



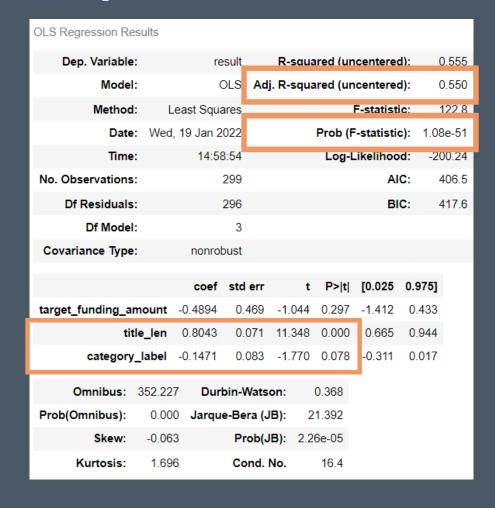
제목의 글자 수가 많을 수록 높은 펀딩 %를 달섬하는 겸햠이 있지만 성공/실패에 유의미한 영향을 미치지는 않음



다중 선형 회귀

독립변수 - 목표 금액, 카테고리, 제목 길이 종속변수 - 펀딩 성공 여부

1. OLS Regression (잔차제곱법)



설명력 : 55%

모혐의 적합도: 1.083e-51 → 유의

유의한 변수: 제목 길이 , 카테고리

y = 0.8043* 제목길이 - 0.1471*카테고리

다중 로지스틱 회귀

독립변수 - 목표 금액, 카테고리, 제목 길이 종속변수 - 펀딩 성공 여부

2. 범주형(0 or 1)인 종속변수 → 더 나은 성능을 위해 다중 로지스틱 회귀 실시

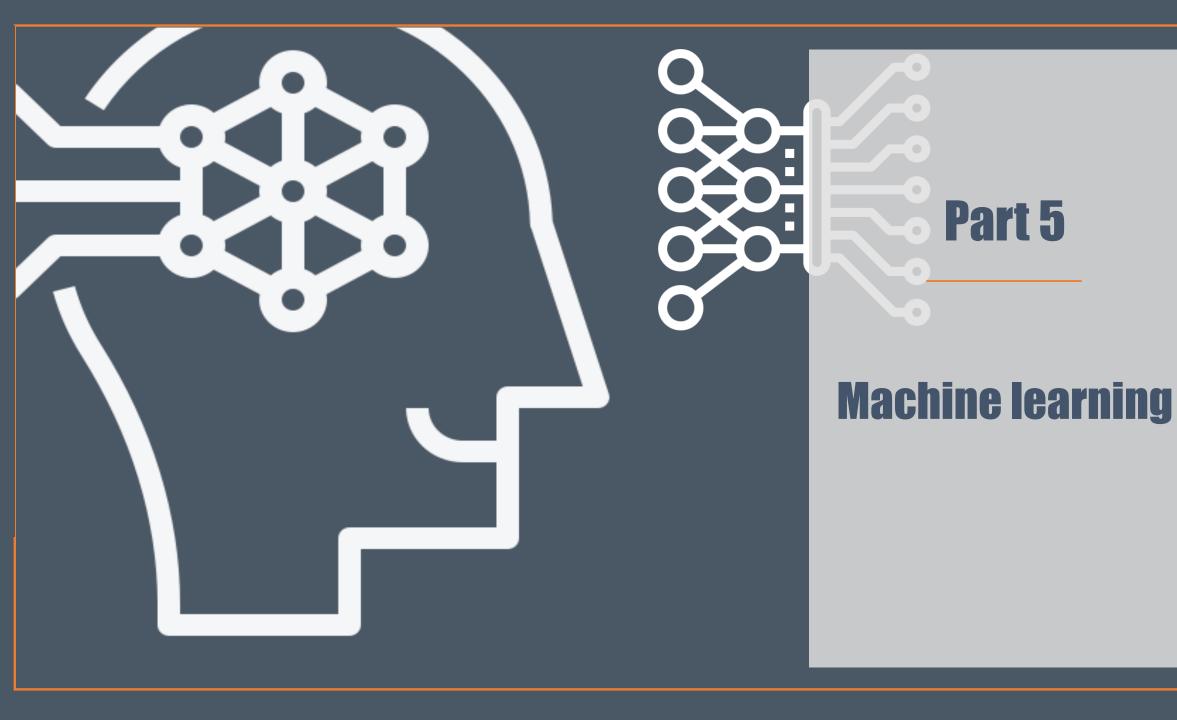
	precision	recall	f1-score	support
0.0 1.0	0.58 0.70	0.74 0.54	0.65 0.61	42 48
accuracy macro avg weighted avg	0.64 0.65	0.64 0.63	0.63 0.63 0.63	90 90 90

설명력: 63%

학습용 데이터셋 정확도: 66%

검증용 데이터셋 점확도: 63%

y = -0.488*목표금액 + 1.882*제목길이 + 1.088*카테고리



성공 실패 여부

1. 성공 / 실패 개수 차이 확인

1-1. 숫자

result count
0 0 149
1 1 9561



성공 실패 여부

2. 제목에서 띄어쓰기, 특수문자, 염어 제외 (한글만 선택)

		title	category	current_funding_amount	percentage	target_funding_amount	supporter	description	result
	ndex								
ĺ	0	짐볼로만든인형돌리볼리	개릭터·굿즈	329000	8	4112500	6명	0	0
	1	함께만드는악기전문웹진	출판	701000	14	5007100	27명	0	0
	2	감사한분들께드리는마음을담을고급가죽봉투	디자인 문구	0	0	0	0명	0	0
	3	마다가스카르사진전	사진	2322000	38	6110500	54명	0	0
	4	인형의시간들	출판	727000	36	2019400	29명	0	0

3. 필요한 자료만 남김 (title)

dfl = df[["title","result"]]

	title	result
index		
0	짐볼로만든인형돌리볼리	0
1	함께만드는악기전문웹진	0
2	감사한분들께드리는마음을담을고급가죽봉투	0
3	마다가스카르사진전	0
4	인형의시간들	0
9705	메카피아의한국전통자개문양피젯스피너	1
9706	카프카변신초판본과독서카드	1
9707	낯선첫발물내딛는이방인물위한연남동작은방	1
9708	함께행복하개유기견팔찌	1
9709	어버이날선물한번에준비하기	1
9710 rd	ows × 2 columns	

성공 실패 여부

Ψ. White space 데이터를 empty value 로 변경한글이 없는 경우 공란으로 처리

```
df1['title'] = df1['title'].str.replace('^ +', "")
df1['title'].replace(", np.nan, inplace=True)
print(df1.isnull().sum())
```

title 56

result 0

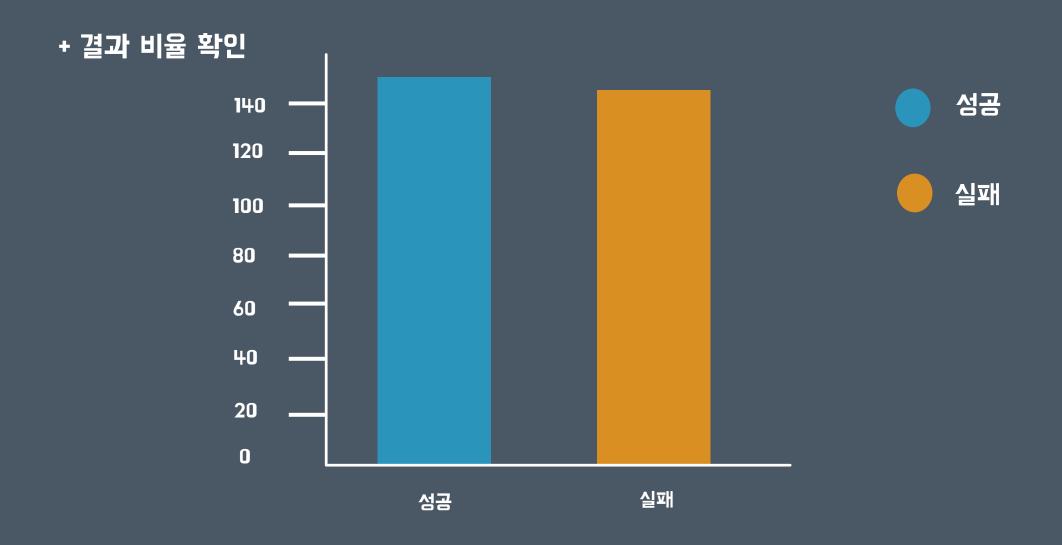
dtype int64

5. Nan title 제거

dfl = dfl.dropna(how = 'any')

index		
0	짐볼로만든인형돌리볼리	0
1	함께만드는악기전문웹진	0
2	감사한분들께드리는마음을담을고급가죽봉투	0
3	마다가스카르사진전	0
4	인형의시간들	0
9705	메카피아의한국전통자개문양피젯스피너	1
9706	카프카변신초판본과독서카드	1
9707	낯선첫발을내딛는이방인을위한연남동작은방	1
9708	함께행복하개유기견팔찌	1
9709	어버이날선물한번에준비하기	1

성공 실패 여부



성공 실패 여부

6. 불용어 처리

```
stopwords =
['의','가','이','은','들','는','좀','잘','강','과','도','를','으로','자','에','와','한','하다']
```

- + 성공과 실패의 데이터 개수 차이가 많이 나서 trαinm test 로 나누기 어려움
 -> 일괄적으로 trαin 에 넣고 진행함.

성공 실패 여부

7. 토큰의 점수화

from tensorflow.keras.preprocessing.text import Tokenizer

tokenizer = Tokenizer()

tokenizer.fit_on_texts(X_train)

print(tokenizer.word_index)

Ct': 24, '되다': 25, '함께': 26, '판타지': 37, '알혁': 28, '선물': 29, '그': 30, '북': 31

5, '년': 36, '책': 37, '일상': 38, '없다': 39, '다': 40, '일러스트': 41, '지': 42, '세상'

7, '향수': 48, '브러쉬': 49, '마음': 50, '설': 51, '사랑': 52, '대': 53, '않다': 54, '속'
되': 59, '여': 60, '적': 61, '앵콜': 62, '절': 63, '시': 64, '있다': 65, '물다': 66, '석'

(CT': 71, '디 : 72, '키트 : 73, 시간 : 74, '봄': 75, '세': 76, '사람': 77, '양': 78, '

(2, '전': 83, '기록': 84, '만': 85, '감성': 86, '디자인': 87, '만나다': 88, '엽서': 89, '?

음': 94, '시리즈': 95, '등': 96, '기억': 97, '파무치': 98, '다이어리': 99, '메다': 100, '
'모든': 105, '에코': 106, '귀걸이': 107, '초콜릿': 108, '가지': 109, '냥': 110, '색롭다': '종다': 115, '잡지': 116, '연': 117, '해주다': 118, '향': 119, '단행본': 120, '번째': 121

[*오늘': 13]

[*오늘': 13]

[*오늘': 13]

[*오늘': 13]

[*오늘': 13]

|'이야기': 13, '세트': 14, '나': 15, '우리': 다': 24, '되다': 25, '함께': 26, '판타지': |5, '년': 36, '책': 37, '일상': 38, '없다':

+ 단어 집합 크기 확인

vocab_size = len(tokenizer.word_index)

print("단어 집합의 크기:", vocab_size)



단어 집합의 크기 : 12081

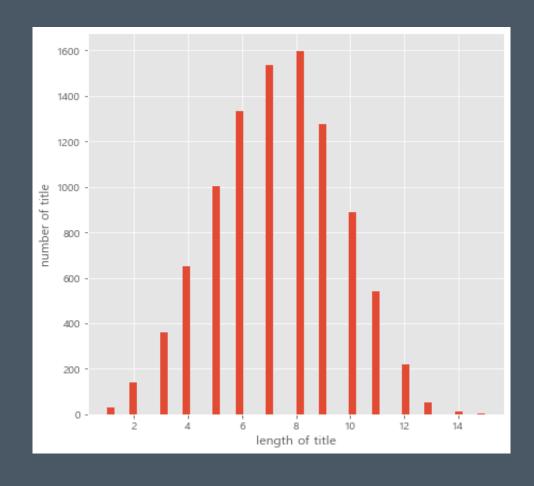
성공 실패 여부

8. 제목 평균 길이 파악

- 성공, 실패를 y_train 에 넣은 후 작업

제목의 최대 길이 : 1 5

제목의 평균 길이 : 7. 303



성공 실패 여부

9. 패딩

- 최대치 에 맞춰서 작업

```
Epoch 1/15

121/121 [==============] ETA : OS LOSS :0.1117- ACC: 0.9776

Epoch 00001 : val_acc improved from inf to 1.00000, saving model to best_model.h5

:
```



성공 실패 여부

10. 예측

sentiment_predict("아름다운 우리나라")

66.39 % 확률로 성공 제목 입니다.

sentiment_predict("함께 만드는 웹진")

56.09 % 확률로 실패 제목 입니다.

sentiment_predict("다이어리")

51.35 % 확률로 성공 제목 입니다.

sentiment_predict("너를 만나는")

68.66 % 확률로 성공 제목 입니다.

성공 실패 여부

실제 사례를 넣어서 예측

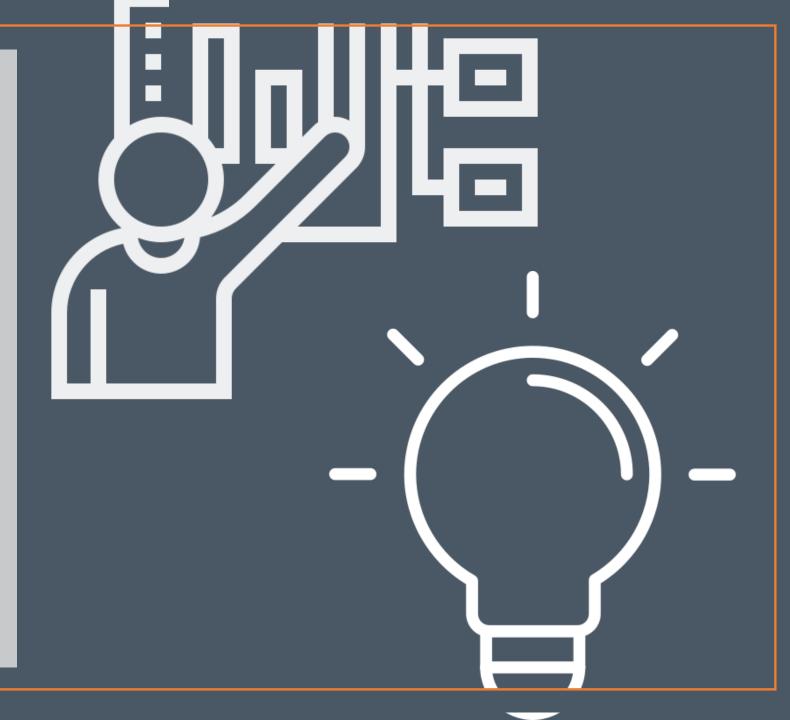
sentiment_predict("새하얀 눈속에서 핀 달콤하고 황홀한 향기 설향 딸기 향수")

64.69 % 확률로 실패 제목 입니다.

실제로는 성공한 펀딩 (2022.01.19 기준 11736% 달성)



Conclusion



Conclusion

전처리 데이터로 분석 진행

독립 변수 : 목표 금액 / 카테고리 / 제목 길이

종속 변수 : 모인 금액 / 달성률 / 후원자 수 / 펀딩 성공 여부

-> 모두 약한 관계를 보이며, 모형의 데이터 설명 가능성이 부족

데이터 불균형 처리 후 재 진행

다중 선형 회귀:

설명력: 55 %

유의

다중 로지스틱:

설명력: 63 %

여러가지의 변수를

종합적으로 고려 할 때

유의미한 회귀식 도출



펀딩 성공 여부에는 목표금액, 카테고리, 제목의 길이 가 종합적으로 영향을 미침

예측:

자연어 처리(제목 글자수) 를 통한 예측 외에 변수 고려에 어려움 있었음,

그러나, 데이터 불균형과 자체 크기의 한계로 **과적합** 우려.

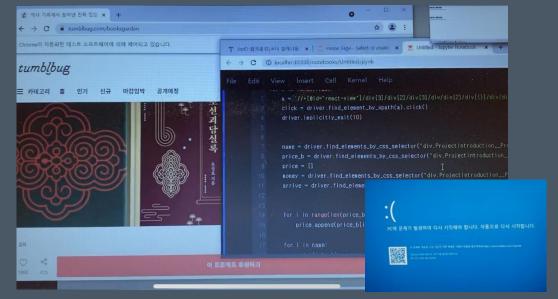
→ 향후 추가 데이터 수집으로 향상된 정확도의 분석 및 예측 가능할 것

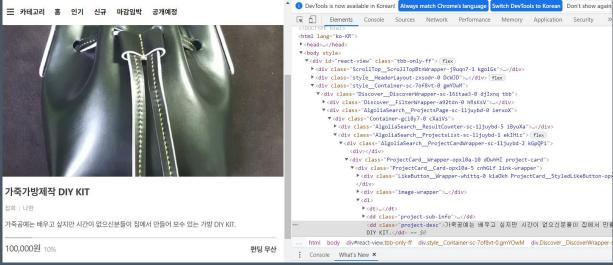
After the Project



After the project ..

아쉬웠던 점





약 1만개 이상의 데이터 수집 시도 시 크롤링이 제대로 진행되지 않고 컴퓨터가 다운되는 문제 동적 페이지 특성 상 <u>크롤링하기 어려운 부분</u>이 일부 있었음

After the project ...

데이터 분석을 해보며

좋은 데이터가 들어가야 분석하기 원활해짐을 느꼈다. 분석 뿐 만 아니라 데이터를 정리 할 때에도 '좋은' 데이터가 되도록 가공하는 단계에 더욱 신경을 써야겠다는 점을 느낌.

또한 <mark>충분한 양의 데이터를</mark> 위해 수집 방법의 효율성을 위해 방향성과 방법론 적인 면 둘 다 신경 써야겠다 라고 생각함.

Team 1 코코

Thank you