

크라우드 펀딩 성공 및 실패 요인분석

wadiz

2023.05.10 ~ 2023.05.26
(6인 팀 프로젝트)



기획의도

펀딩 사이트 내에 노출되는 정보를 바탕으로 펀딩의 성공과 실패에 미치는 영향력을 분석하고 성공여부 예측과 펀딩 성공 가이드 라인 제공을 통해서 새롭게 시작하는 개인 및 스타트업의 시행착오를 최소화 하기 위함

역할

데이터 크롤링, 자연어 처리, 이미지 색상 분석, 데이터 전처리, EDA

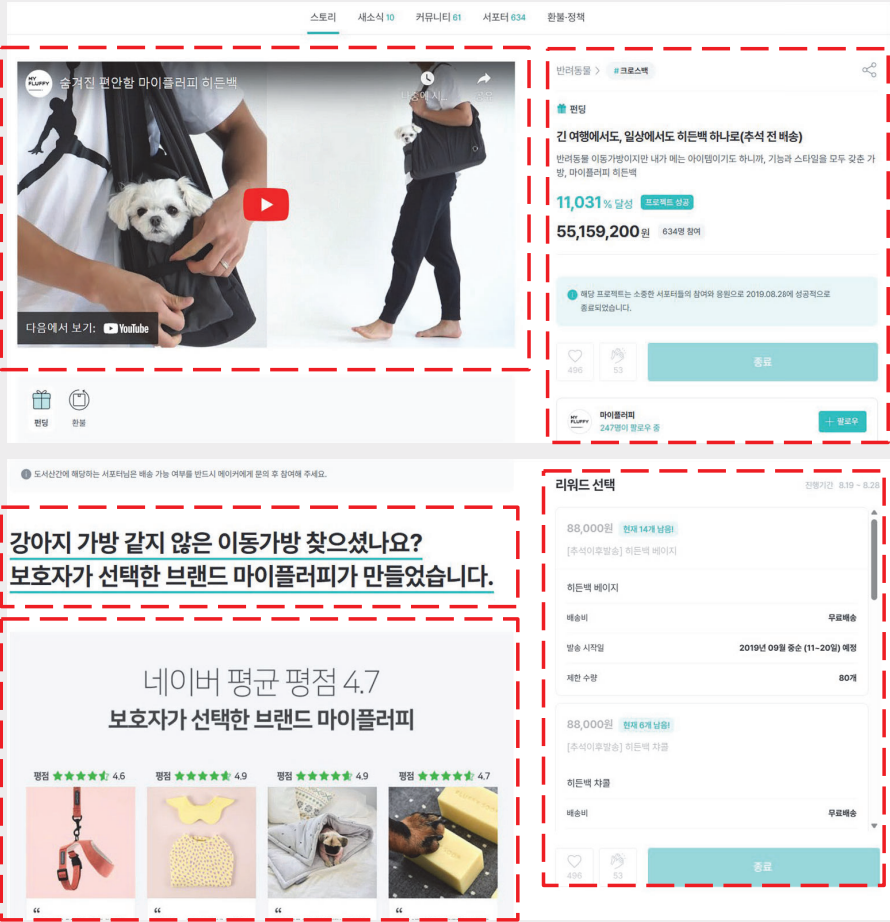
데이터셋

와디즈(wadiz) 펀딩 사이트 내에 노출되는 텍스트 및 이미지 데이터

결과 및 리뷰

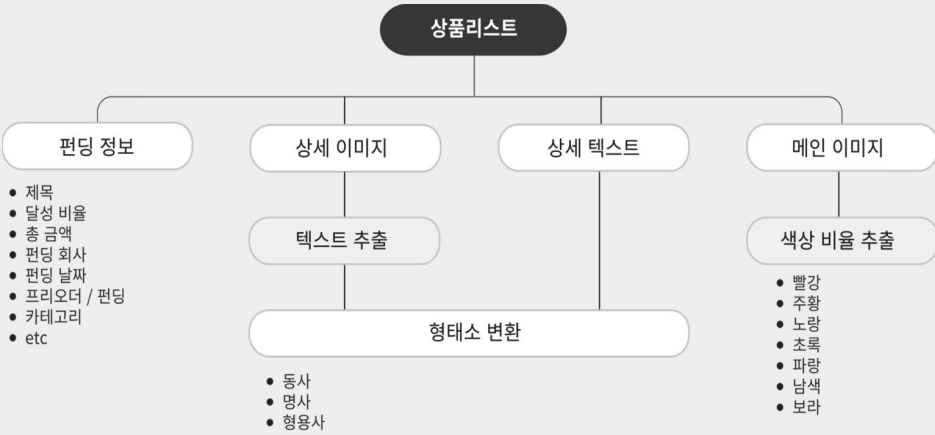
모델을 선정하고 변수의 중요도를 평가해서 불필요한 변수를 제거한 뒤, 하이퍼 파라미터 튜닝을 거쳐서 의미 있는 성과를 낼 수 있었고, 본 프로젝트를 좀 더 발전 시켜서 카테고리별 상품 설명과 이미지 파일, 펀딩 기간 등을 입력하면 펀딩의 성공 여부를 예측 하고 성공 확률을 높이기 위해서 '긍정어/부정어/기술용어'의 사용 빈도와 이미지의 색상 등에 대한 개선점을 가이드 해주는 프로그램 또는 알고리즘을 만들 수 있을 것으로 예상됨

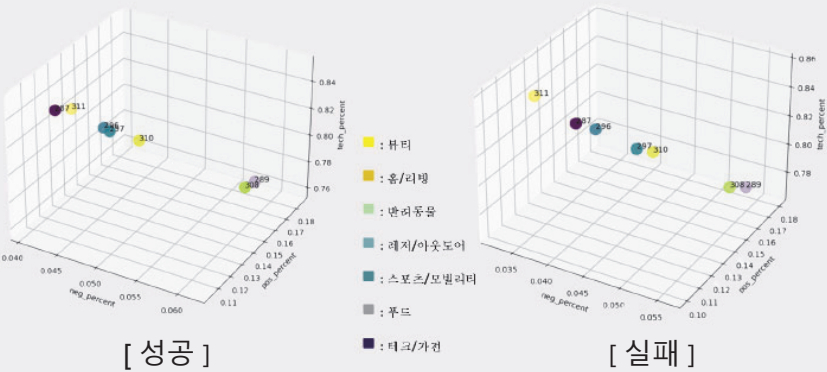
Second Project 크라우드 펀딩 성공 및 실패 요인분석



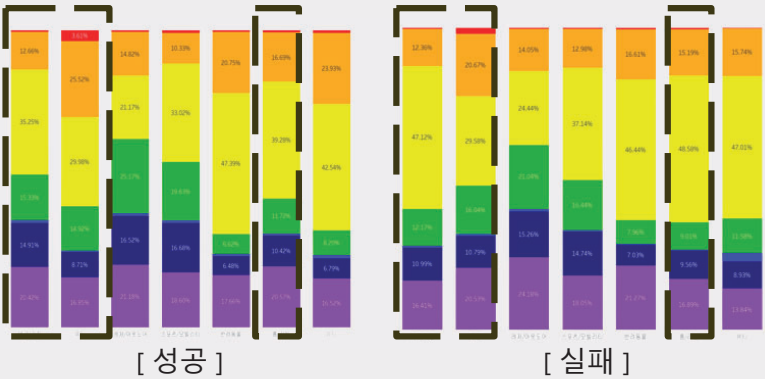
데이터 수집

와디즈(wadiz)의 펀딩 페이지에서 Selenium, BeautifulSoup을 이용하여 이미지 및 텍스트를 크롤링하고 크롤링한 이미지는 Google Cloud Vision API를 이용하여 이미지에서 텍스트 추출을 진행





카테고리별 텍스트 분석 [긍정 / 부정 / 기술용어 비율]



카테고리별 색상 분석 [7색 기준]
(빨강 / 주황 / 노랑 / 초록 / 파랑 / 남색 / 보라)

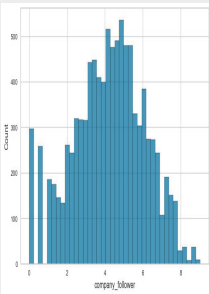
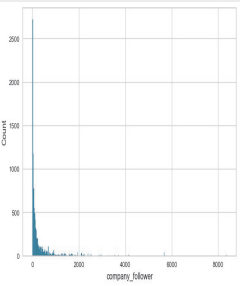
텍스트 분석

펀딩 상세 페이지의 이미지에서 추출한 텍스트와 작성된 텍스트를 ‘긍정어/부정어/기술용어’로 분류하여 3차원 그래프상에 표시하고 카테고리별 성공/실패 사례를 나누어서 텍스트 비율을 분석


이미지 색상 분석

펀딩 상세 페이지에 첨부되어있는 이미지의 색상을 7가지 색으로 분류하고 카테고리별 성공/실패 사례를 나누어서 각 색상의 비율을 분석

Second Project 크라우드 펀딩 성공 및 실패 요인분석



log 변환 적용



Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.9029	0.9238	0.8791	0.9003	0.9473	0.3011	0.3384
rf	Random Forest Classifier	0.9017	0.9130	0.9025	0.9145	0.9472	0.2415	0.2834
et	Extra Trees Classifier	0.8983	0.7884	0.8928	0.9050	0.9489	0.1205	0.1830
gbm	Gradient Boosting Classifier	0.8969	0.8201	0.9091	0.9203	0.9440	0.2869	0.3044
ada	Ada Boost Classifier	0.9726	0.9025	0.9329	0.9259	0.9392	0.2841	0.2881
dt	Decision Tree Classifier	0.8374	0.8138	0.8873	0.8220	0.9042	0.1996	0.2021
svm	SVM: Linear Kernel	0.7843	0.0000	0.7038	0.9409	0.9469	0.1916	0.2477
knn	K Neighbors Classifier	0.7349	0.7061	0.7527	0.8402	0.8358	0.1963	0.2233
lda	Linear Discriminant Analysis	0.7083	0.7773	0.7072	0.9539	0.8117	0.2025	0.2610
ridge	Ridge Classifier	0.7050	0.0000	0.7069	0.9539	0.8115	0.2023	0.2907
lr	Logistic Regression	0.7043	0.7859	0.7053	0.9535	0.8103	0.1990	0.2574
nb	Naive Bayes	0.5517	0.7570	0.5187	0.9854	0.6745	0.1260	0.2142
qda	Quadratic Discriminant Analysis	0.1754	0.6949	0.0821	0.9807	0.1505	0.0169	0.0877
dummy	Dummy Classifier	0.1023	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

Pycaret
상위 5개 모델 비교

[변수 중요도]

1.부절여 계수	65
2.공정여 계수	39
3.기술여 계수	25
5.카테고리	22
6.판당 기간	20
7.배출여기 기간	18
8.공시 할조율	17
9.배당 시간계수	17
10.배당여 사진 or 영상	13
11.판당 종류	12
12.판당 평균 소비액	10
13.시작년도	10
14.시작날	9
15.시작일	7
16.종료년도	6
17.종료날	6
18.종료일	6
19.시작 계절	5
20.종료 계절	5
21.판당 여부	5
22.제출길이	5
23.배당사진 색깔(빨)	4
24.배당사진 색깔(노)	4
25.배당사진 색깔(노)	3
26.배당사진 색깔(초)	2
27.배당사진 색깔(파)	1
28.배당사진 색깔(노)	1
29.배당사진 색깔(노)	1
30.배당사진 색깔(노)	0



F1_macro = 0.628
Accuracy = 0.907



F1_macro = 0.734
Accuracy = 0.929

전처리

변수 분포가 치우쳐 있어서 로그 변환과 이상치를 제거한 뒤, Pycaret 라이브러리를 이용하여 상위 5개의 모델을 대상으로 하이퍼 파라미터 튜닝을 진행하여 비교한 결과, ‘튜닝된 Light GBM’을 최종 모델로 선정

하이퍼 파라미터 튜닝

이상치 제거 및 하이퍼 파라미터 튜닝을 통해서 아래와 같이 개선된 것을 확인함

	f1 Score	Accuracy
튜닝 전	0.628	0.907
튜닝 후	0.734	0.929
증감량	+ 0.106	+ 0.022