소상공인을 위한 창업 예측 서비스

日の大胆

고객의 실시간 반응 통계로 창업 성공률을 예측해 보세요!

목차소개

Table of Contents

 1
 추진 배경
 0
 4비스 소개

 • 문제 정의
 • 핵심 서비스
 • Trend 분석을 통한 프랜차이즈 점포 수 예측

 • 리얼타임 반응 리포트
 • 지역별 매장 입점 현황

 • 지역별 매장 입점 현황
 • Ul 구현

 기대 효과 및 발전 가능성
 • 데이터 기반 창업 전략

 • 할용 데이터
 • 한계점 및 개선 방향

15 | Lesson Learn

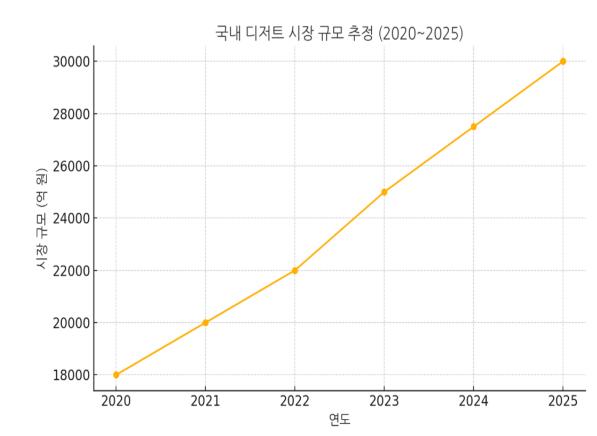
06

질의응답



01. 추진배경

높은 디저트 업종 폐업률



[국내 디저트 시장 규모] 2020년 약 1.8조 원 → 2025년 약 3조 출처: 중소벤처기업부, 식품산업통계, 업계 자료기반 추정



[음식점 페업률 및 폐업 건수] - 전체 음식점 중 상대적으로 더 큰 휴게업종 리스크

국내 디저트 시장 매년 약 10%의 안정적 성장세지만, 휴게음식점 폐업률은 여전히 매우 높음



>>> 폐업 주요 원인

- 업종 선정 실패가 가장 큰 원인
- 폐업을 고려하지 않는 이유는 차후 대안이 없기 때문



폐업 경험자들의 폐업 원인





01. 추진배경

>> 기존 서비스의 한계점

- 기존 상권분석 서비스는 유동인구, 매출 점포 수 같은 정량 데이터 (물리적 지표) 에 집중
- But, 매장의 성패는 '소비자 반응' 이 결정





소비자 감정 데이터 "감"에 의존한 창업 방식 데이 차트 최신 상권 분석 시장 반응

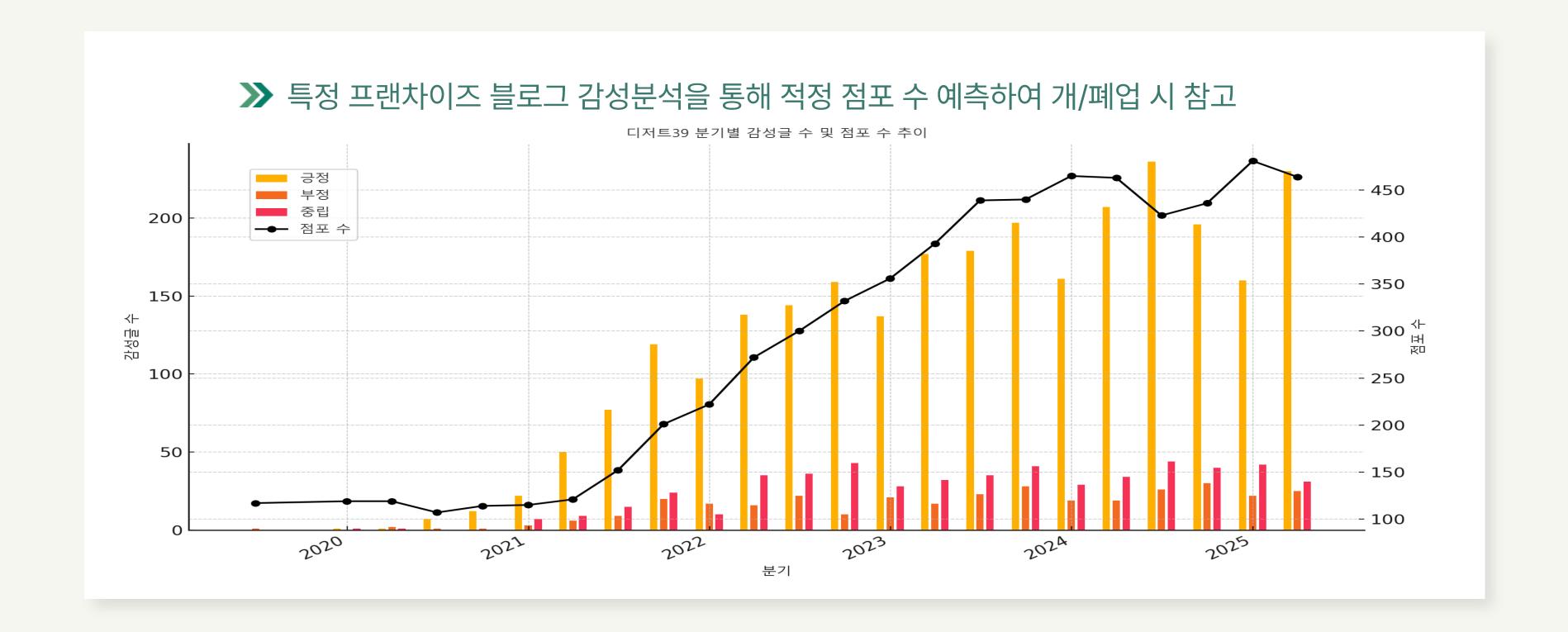


01 Trend 분석을 통한 프랜차이즈 점포 수 예측

02 리얼타임 반응 리포트

03 지역별 매장 입점 현황

Trend 분석을 통한 프랜차이즈 점포 수 예측



Trend 분석을 통한 프랜차이즈 점포 수 예측

>>> 월 단위 예측 모델 (Gradient Boosting)







고민중인 브랜드에 대한 리얼타임 반응 리포트



>> 서비스 작동 프로세스



 \rightarrow















키워드 입력

블로그 자동 수집

감정 분석

데이터 시각화

인사이트 도출

사용자가 분석할 브랜드 또는 키워드 입력 최근 7일간 관련 블로그 게시글 자동 수집 및 정리

게시글 내 문장을 단위로 긍정 / 부정 감정 분류

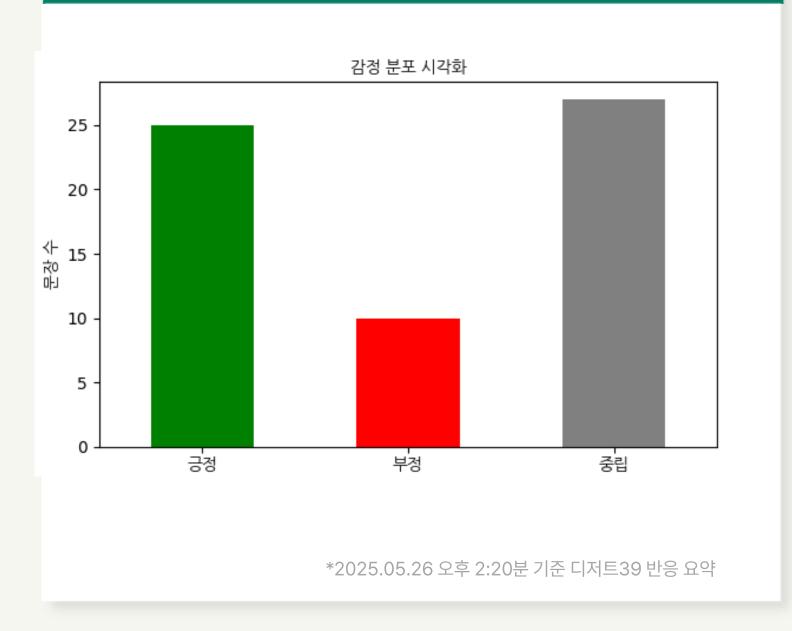
감정 분포와 핵심 키워드 시각화 소비자 반응과 브랜드 이슈 직관적확인



리얼타임 반응 리포트

>> 감성 반응 요약

브랜드에 대한 최근 7일간의 소비자 반응을 정량화해 감정 반응을 시각화

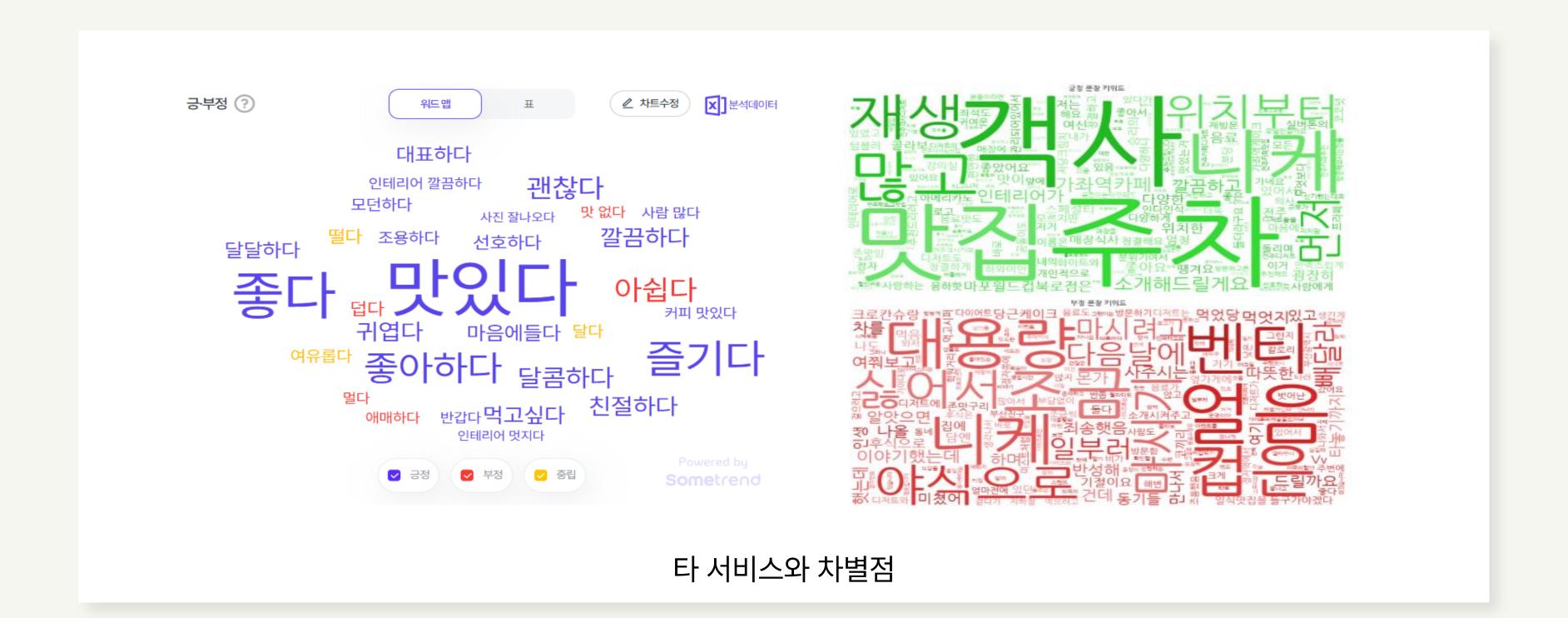


>> 소비자 언급 키워드

워드클라우드를 통해 소비자들이 주목하는 브랜드 이슈와 키워드 도출

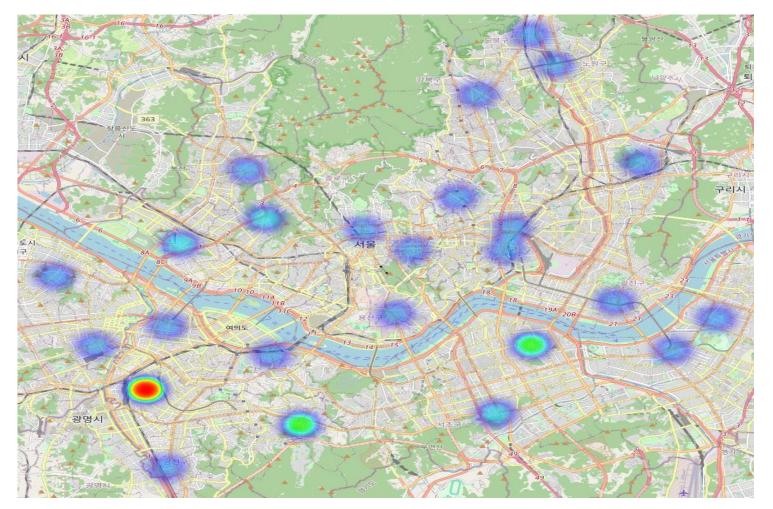








>> 지역별 매장 분포 및 경쟁도 분석

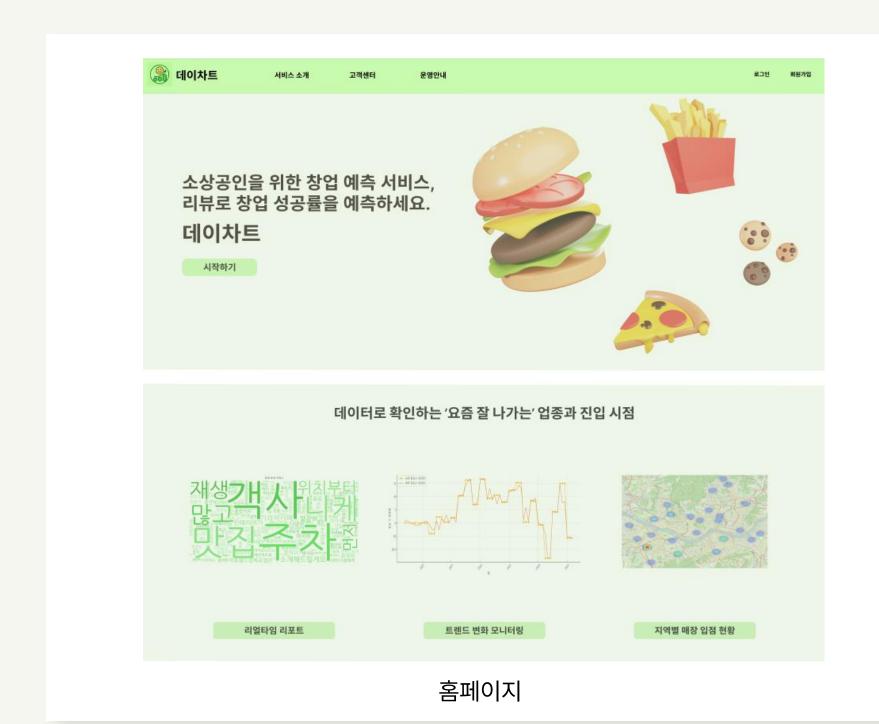


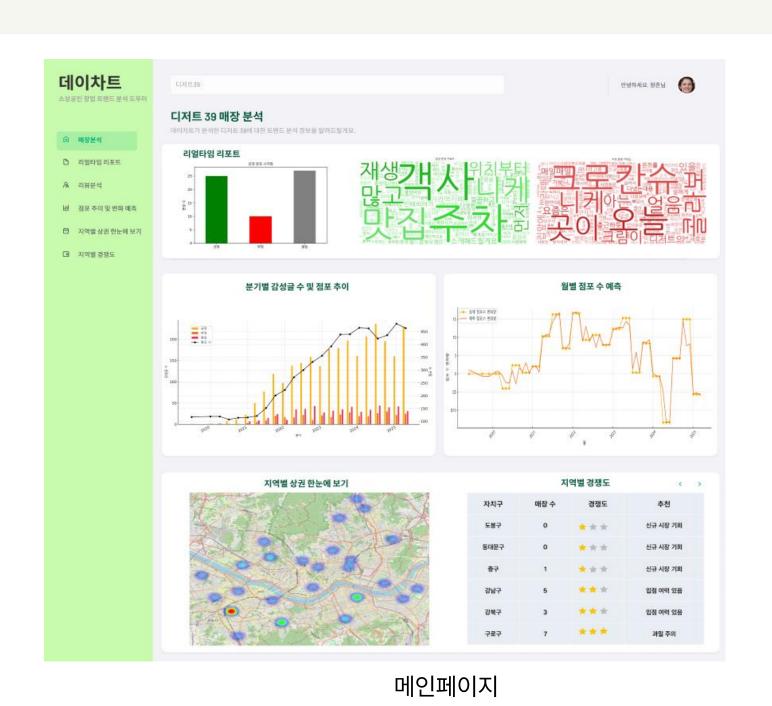
매장 분포를 히트맵으로 시각화하여 밀집도 파악

자치구	매장수	인구수	5만명당 매장수	경쟁도	추천이유
도봉구	0	139002	0	* \$ \$	신규 시장 기회
동대문구	0	174846	0	* \$ \$	신규 시장 기회
중구	1	64863	0.770855495	* & &	신규 시장 기회
강북구	3	143249	1.047127729	★★☆	입점 여력 있음
강남구	5	245139	1.019829566	★★☆	입점 여력 있음
구로구	7	183970	1.902484101	***	과밀 주의

인구 5만명당 매장 수 기준 경쟁도 분석 1.0 미만 (저밀지역) / 1.0 이상 1.8 미만 (적정 밀도) / 1.8 이상 (고밀지역)

UI 구현







>>> 파인튜닝

>>	Step1. 데이터 수집 및 전처리	Step2. 지도학습 라벨링	Step3. 1차 파인튜닝	Step4. 추가 학습 데이터 생성	Step5. 추가 파인튜닝	Step6. 실험 관리 및 버전 트래킹
기능 요약	네이버 블로그 크롤링해 감정 키워드 리뷰 추출 의미 없는 문장 제거 후 CSV 저장	긍정, 부정, 중립 감정 수작업 분류해 키워드 기반 점수로 자동 보조	- kcbert-base 모델로 2~3분류 구현 - Lexicon 점수 혼합 방식 적용 - 감정 + 신뢰도 출력	- 브랜드 비교 문장 템플릿 자동 생성기 개발 - 긍/부 표현 포함 문장 생성	- 총 5회 파인튜닝 진행 - 중립 포함 3분류로 전환 - 하이퍼파라미터 조정 및 실험 반복	- wandb로 실험 기록 및 시각화 - 모델별 버전 관리 (예: v2, v2-5)
도입 목적	- 분석 적합 문장 확보 - 학습용 데이터 정제	- 고품질 라벨링을 통한 모델 학습 정확도 향상	- 문맥 기반 + 키워드 기반 감정 판단 결합 - 신뢰도 높은 감정 분류	- 추가학습을 위한 다양한 문장 유형 확보	- 감정 분류 성능 고도화 - 다양한 문장 유형 반영	- 실험 재현성 확보 및 성능 비교 체계화
탈용 도구	- Selenium - ChromeDriver - webdriver-manager - re	- 수작업 라벨링 - 키워드 기반 필터	- Transformers - Torch - AutoModel (Sequence Classification)	- Python 기반 자동 생성기 - Excel 저장	- Kcbert - Wandb - Torch - learning_rate 등 튜닝	- wandb (Weights & Biases), - 프로젝트 버전명 시스템



>>> 데이터 크롤링

>>	^{Step1.} 브랜드명 입력 (Input)	Step2. 네이버 블로그 크롤링	Step3. Input관련 감정 문장 추출	Step4. CSV 저장	Step5. 분석
기능 요약	사용자 분석할 브랜드명 입력 main() 실행 시 전체 프로세스 자동 시작	네이버 블로그 내 최근 7일 이내의 관련 게시글 자 동 수집	크롤링한 문장에서 브랜드에 대한 감정 표현 문장만 추출해 긍정 / 부정 / 중립 구분	문장, 감정, 신뢰도, 링크 등을 CSV 형식으로 저장	- 시각화 결과 생성 - ML 활용 점포 수 예측
도입 목적	- 입력된 브랜드 기준으로 전체 분석 흐름 시작 - 최종 파일명 도출	실시간 사용자 반응 확보	감정 분포 정량화를 위해 핵심 문장만 선별 추출	후속 리포트 및 활용을 위한 데이터 보존	데이터 통한 인사이트 도출
활용 도구	- Python 모듈화 - subprocess - Os	- Selenium - webdriver-manager - Re	-Transformers -Torch -BERT 기반 파인튜닝 모델 (sentiment-model-v2-5)	- pandas	- Matplotlib - Wordcloud - Nbconvert - 한글 폰트 설정 - 머신러닝



>>> ML : Gradient Boosting

모델	R² (설명력)	MSE (평균 오차)
Linear Regression	0.429	3.07
Random Forest	0.634	2.50
XGBoost	0.685	2.36
Gradient Boosting	0.692	2.30

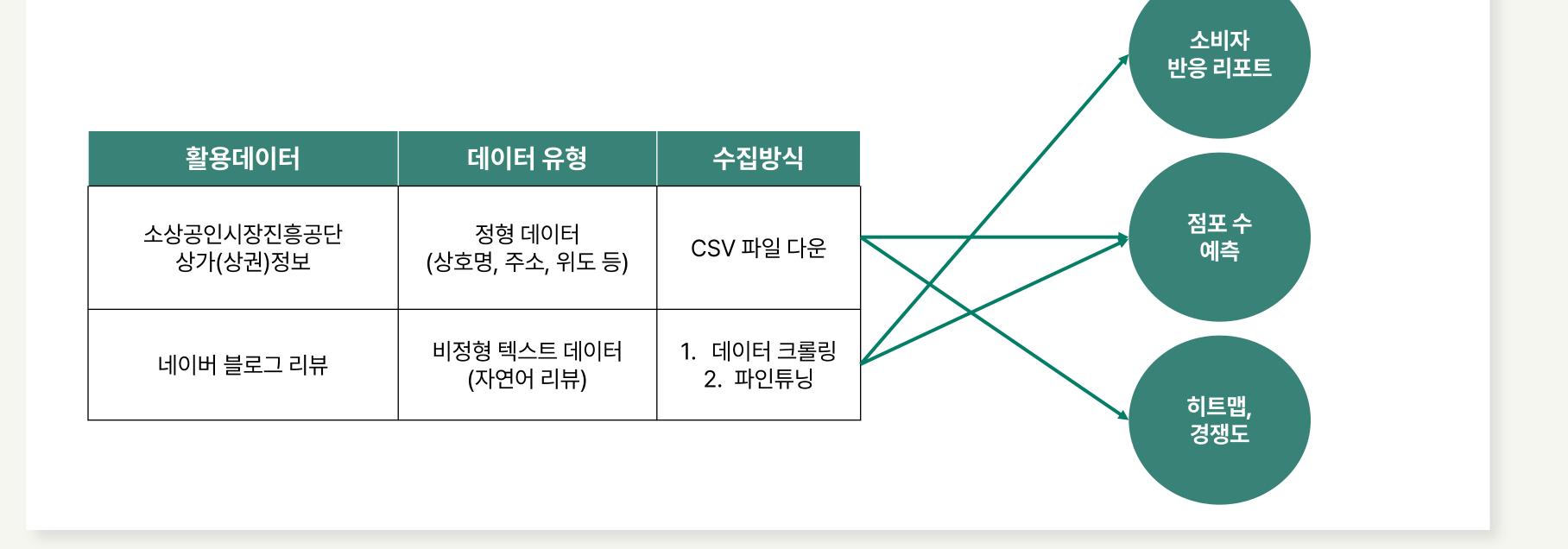
^{*} Gradient Boosting: 손실 함수를 기준으로 오차를 줄이는 방향(gradient)에 따라 약한 학습기(weak learner) 를 순차적으로 추가하여 최종 예측 모델을 점진적으로 향상시키는 앙상블 학습 기법



>>> 매장 분포 기반 입지 제안

Step1.	•	Step3.	Step4.
데이터 정제 및 필터링		매장 분포 히트맵	유망 지역 및 경쟁도
Pandas를 활용해 CSV 파일을 통합하고 결측값을 제거 '디저트39' 키워드를 포함한 상호명만 필터링하여 분석 대상 데이터 확보	정제된 데이터를 기반으로 서울시 25개 자치구별 '디저트39' 매장 수를 집계 구별 분포를 수치화하여 브랜드의 지역 확산 정도와 밀집도를 정량적으로 분석	Folium 기반으로 매장 위치 데이터를 시각화하고, 자치구별 밀집도를 히트맵 형태로 지도에 표시 지역의 과잉 또는 미진출 여부를 한눈에 파악 가능	서울시 자치구별 인구 5만명당 매장 수를 기준으로 경쟁도를 산출, 진입에 유리한 지역 제시 고밀/적정/저밀 구간으로 나눠 현재 진출 수준 진단







04. 기대 효과 및 발전 가능성

데이터 기반 창업 전략

>>> 프랜차이즈 창업을 위한 정량적 데이터 제공

- 브랜드에 대한 소비자 감정 반응 시각화
 - 부정 키워드 증가 등 이상징후 감지 후 사전 경고 및 개선 유도
- 창업 시점 예측
 - 브랜드별 소비자 반응 변화 추이를 분석해 창업 최적 시점 예측
- 지역 상권 진단
 - 인구 대비 매장 수 경쟁도 분석 후 과밀/적정/저밀 상권 추천



"소상공인을 위한

성공적인 창업, 운영 지원"



04. 기대 효과 및 발전 가능성

≫ 한계점

- 정밀한 인사이트 도출 한계
- 데이터 출처가 블로그에 국한되어 있어 SNS 채널 확대 필요
- 매출 지표 연 단위 제공으로 사용 제한, 최근 추이와 연계 어려움



≫ 개선방향

- SNS 리뷰 크롤링 확장: 인스타그램, 유튜브 등 리뷰 포함
- 감정 분석 모델의 파인튜닝 고도화로 정확도 향상
- 외부 매출 데이터 연동 (배달의민족, 국민카드 등)



05. Lesson Learn

- 주제 관련, 데이터 확보가 어려움
 - → 데이터 확보 가능성을 고려하여 주제 선정 필요
- 여러 개의 요구사항을 한꺼번에 GPT에 코드 생성 요청 시, 디버깅이 어려움
 - → 요구사항을 하나씩 반영하는 것이 디버깅이 용이하고 코딩 속도 향상
- GPT를 통한 데이터분석 시 데이터 정제/가공에 오류가 많이 발생
 - → 반복적인 데이터분석시 원본 데이터로 분석하는 대신, 정제된 데이터를 저장하여 사용
- 중복작업이 발생하고 일부 업무(보고서작성)에 시간이 부족함
 - 프로젝트 초기, 작업범위/업무량 기반에 정확한 R&R 선정 필요

* Roles and Responsibilities



06. 질의응답

