개인 오픈데이터 프로젝트 · 김규민

다음 데이터센터는 어디에 지어야 할까?

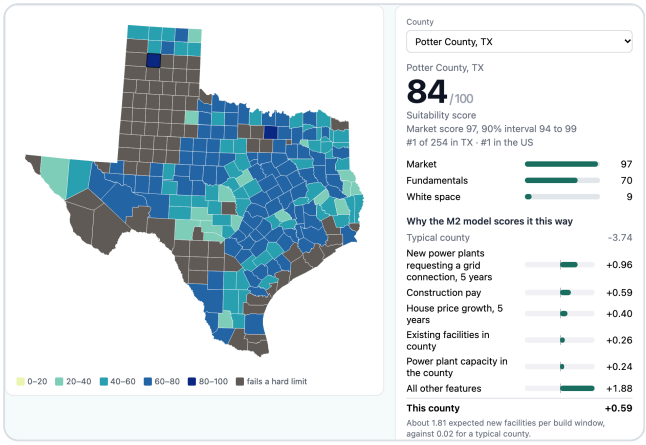
공공 데이터만으로 데이터센터 입지를 카운티 단위로 살펴본 프로젝트입니다. 실제로 데이터센터가 어디에 지어졌는지를 학습하고, 입지를 좌우하는 조건(전력, 토지, 재해, 물, 인력, 정책, 지역 수용성)과 결합해, 그 결과를 점수의 근거와 함께 인터랙티브 지도로 보여줍니다.

3,108
점수화한 카운티 (미국 본토 전체)

40+
결합한 공공 데이터셋

~80
카운티마다 측정한 조건

2
따로 점수화하는 세그먼트



인터랙티브 지도 (텍사스, 하이퍼스케일 세그먼트). 카운티 색은 점수, 회색은 탈락 조건에 걸린 곳입니다. 오른쪽 카드는 한 카운티를 설명하고, 어떤 조건이 점수를 올리고 내렸는지 보여 줍니다.

만든 이유

데이터센터 붐에 관한 뉴스를 보면서, 기업들이 어디에 지을지를 어떻게 정하는지 궁금해졌습니다. 전력과 통신 접근성이 가장 중요하지만, 부지는 주변 지역과도 맞아야 합니다. 노동시장, 토지와 주거 비용, 자연재해, 물, 지역 정치, 그리고 주민이 새 시설을 반기는지 아니면 모라토리엄으로 막는지까지요.

이런 요소를 한꺼번에 볼 수 있는 공개되고 투명한 방법을 찾지 못해서, 머신러닝과 공개 데이터만으로 가벼운 프로젝트가 어디까지 갈 수 있는지 직접 해 보기로 했습니다.

두 개의 세그먼트: 원하는 입지가 다르기 때문

A · 하이퍼스케일 및 AI

50 MW 이상 캠퍼스, 또는 하이퍼스케일러(Amazon, Google, Microsoft, Meta, Apple, Oracle)가 운영하는 시설. 전력 확보, 토지, 전력망 용량이 핵심입니다.

B · 코로케이션 및 기업용

더 작은 규모의 임대형 또는 기업 자체 시설. 고객, 광통신망, 숙련 인력, 대도시 접근성이 핵심입니다.

시사점

공개 데이터로도 꽤 멀리 간다

2016-2020년 데이터로만 학습한 모델이, 2021-2026년에 지어진 시설의 45-50%를 상위 10% 카운티에 담았습니다.

전력이 가장 중요하다

전력망 연결을 기다리는 신규 발전 용량과 주변 발전소 규모가 가장 강한 신호로, 토지나 인건비보다 앞섭니다.

화이트스페이스가 있다

조건은 좋은데 아직 시설이 적은 카운티, 즉 시장이 아직 들어가지 않은 곳이 있습니다.

지역 수용성이 빈칸이다

모라토리엄과 주민 반대는 실제 프로젝트를 좌우하지만, 이를 다룬 공개 데이터는 부족합니다.

작동 방식

공공 데이터에서, 근거를 따져 볼 수 있는 점수까지

1 · 수집
40개 이상의 공공 데이터. 출처와 라이선스를 기록하고 버전을 관리합니다.

2 · 측정
카운티마다 약 80개 지표. 2015년, 2020년, 현재 시점.

3 · 학습
어떤 조건 뒤에 신규 건설이 따라왔는지 모델이 학습합니다.

4 · 점수화
0-100 점수 4종과 탈락 조건 확인.

5 · 설명
지도, 순위, 그리고 이유를 보여 주는 카운티 카드.

어떤 데이터가 들어가나

분야	측정하는 것	주요 출처
시설	알려진 데이터센터와 프로젝트의 위치, 규모(MW), 운영사, 상태, 연도	Compute Atlas, OpenStreetMap
전력	산업용 전기요금, 정전, 지역 발전량, 송전선·변전소까지 거리, 전력망 연결을 기다리는 프로젝트	EIA-860·861, LBNL Queued Up, OpenStreetMap 전력망, NREL
현재 전력망 접근성	전력을 받기까지 예상 기간, 전력망이 얼마나 붐비는지, 대용량 고객용 요금제	LBNL 대기열 (요청 38,000건), 전력회사 요금 신고서, OpenEI URDB
재해	자연재해 위험, 홍수 위험 지역, 지진, 산불, 산사태, 싱크홀이 생기기 쉬운 지반	FEMA NRI·홍수지도, USGS, USFS
경제·토지	인구, 소득, IT·건설 인력과 임금, GDP, 주택가격, 임대료, 농지 가격	Census ACS, BLS, BEA, FHFA, USDA
기후·물	냉방이 얼마나 필요한 기후인지, 여름철 더위, 물 사용량, 물 부족 정도	NOAA, USGS, WRI Aqueduct
통신·정책	광통신 커버리지, 도로, 공항, 대기질 기준 충족 여부, 판매세 감면, 신규 프로젝트 금지 조치	FCC, Census TIGER, FAA, EPA, 주 법령

방법

핵심 예측은 **앞으로 5년 남짓 동안 한 카운티에 새 데이터센터가 몇 개나 들어설 것인가**이며, 세그먼트별로 따로 예측합니다. 카운트 모델(Poisson 목적함수)이 실제 두 번의 건설 물결에서 배웠습니다. 2015년 조건에서 2016-2020년에 지어진 것, 그리고 2020년 조건에서 2021-2026년에 지어진 것입니다. 열 가지 모델(규제 회귀, 랜덤 포레스트, XGBoost, CatBoost, 신경망, LightGBM)을 비교했고, Optuna로 튜닝한 LightGBM이 가장 좋았습니다.

모든 점수에는 이유가 붙습니다. SHAP 값으로 어떤 조건이 점수를 올리고 내렸는지 보여 주고, Market 점수에는 90% 구간이 함께 표시됩니다.

네 가지 점수 (0-100)

- **Market**: 다음 신규 건설이 어디서 일어날지에 대한 모델의 예상.
- **Fundamentals**: 조건만 놓고 본 입지 점수. 각 조건을 얼마나 중요하게 볼지 사용자가 정할 수 있고, 처음 값은 모델에서 나옵니다.
- **Suitability**: 위 두 점수를 섞은 종합 점수. 섞는 비율은 슬라이더로 정합니다.
- **White space**: 조건은 좋은데 주변에 기존·계획 시설이 적은 곳.

전력·전력망 정보

전력 공급까지 걸리는 시간

과거 전력망 연결 요청 38,000건이 실제로 몇 년 걸렸는지를 생존모형으로 학습해, 카운티별로 프로젝트가 전력을 받기까지 몇 년 기다리는지를 예상 범위와 함께 추정합니다.

전력망 혼잡도

반경 100 km 안에 계획된 데이터센터 전력 수요를 주변 발전소, 곧 들어설 가능성이 높은 신규 발전소, 고압 송전선 수와 비교합니다.

대용량 고객용 요금제

공개 신고서에서 읽어 낸 전력회사 요금제 55개: 최소 규모, 계약 기간, 최소 월 요금, 보증금 조건. 장기 전력 비용 계산에 들어갑니다.

잘 맞을까?

실제로 쓰일 방식 그대로 검증했습니다

핵심 검증: 과거 데이터로 미래를 맞히기

2016-2020년 건설 데이터로만 모델을 학습시키고, 2015년 데이터로 모든 카운티의 순위를 매긴 뒤, 2021-2026년에 실제로 무엇이 지어졌는지 확인했습니다. 카운티를 무작위로 고르면 상위 10%에는 신규 용량의 약 10%가 들어갑니다.

32%

신규 하이퍼스케일·AI 용량(MW)의 32%가 모델이 꼽은 상위 10% 카운티에 있었습니다. 무작위의 약 **3배**입니다 (시설 수 기준 45%).

57%

신규 코로케이션·기업용 용량(MW)의 57%가 상위 10% 카운티에 있었습니다. 무작위의 약 **6배**입니다 (시설 수 기준 50%).

수치 뒤의 검증 절차

- **처음 보는 지역으로 시험.** 학습할 때 지역(200 km 구역)을 통째로 빼 두었다가 그 지역으로 시험합니다 (공간 교차검증).
- **공정한 모델 선택.** 가장 강한 네 가지 방법(LightGBM, XGBoost, 랜덤 포레스트, CatBoost)을 똑같은 노력으로 튜닝해 같은 시험으로 비교했고, 매번 LightGBM이 1위였습니다.
- **측정된 개선.** 첫 버전 대비, 처음 보는 지역에서의 예측 오차가 17%(하이퍼스케일), 7%(코로케이션) 줄었습니다.
- **모든 결정을 기록.** 모델링 선택, 기본값, 알려진 약점을 이유와 함께 모두 문서로 남겼습니다.

설계 원칙

- **탈락 조건**에 걸린 카운티는 숨기지 않고 이유와 함께 회색으로 표시합니다: 25 km 안에 고압 송전선(230 kV) 없음, 강한 지진 위험, 홍수 위험 지역 30% 이상, 심각한 물 부족, 신규 프로젝트 금지 조치. 원하면 전력을 받기까지 오래 걸리는 곳도 걸러 냅니다.
- **책임 있는 사용.** 인종·민족 정보는 모델 입력으로 쓰지 않고, 편향 점검에만 씁니다. 지역 반대 여론은 "해결해야 할 우려 사항"으로 보여 주며, "저항이 가장 적은 곳"을 찾는 용도로 쓰지 않습니다.
- **공개 데이터.** 모든 출처는 공개 데이터이며, 각 출처의 라이선스를 기록합니다.

한계

- **필지가 아닌 카운티 단위.** 지역을 좁혀 줄 뿐, 부지를 고르지는 않습니다.
- **공공 데이터만 사용.** 상용 시설 데이터베이스를 쓰면 코로케이션 쪽 그림이 더 정확해집니다.
- **데이터가 부족한 곳은 추정치.** 데이터센터가 전력을 받기까지 걸리는 기간은 공개 자료가 없어, 발전소가 전력망에 연결되기까지 걸린 기간으로 대신 추정합니다. 요금제 조건 대부분은 신고서에서 읽어 낸 것이라 "낮은 신뢰도"로 표시됩니다.
- **총량 예측이 아닌 순위.** 점수는 활동이 어디서 더 일어날지를 말할 뿐, 미국 전체에 몇 MW가 지어질지를 예측하지 않습니다.

살펴보기

인터랙티브 지도

가벼운 버전을 브라우저에서 바로 볼 수 있습니다:
www.kyuminkim.com/projects/dc-site

더 자세한 버전

비용 계산기, 카운티 비교, 시설 지도 등 더 자세한 내용을 담은 버전은 요청하시면 보여 드립니다.

연락처

질문, 아이디어, 협업:
kyumin.kim.econ@gmail.com

김규민 개인 프로젝트 (Lorenzo Swank와 함께 개발) · 2026년 11월. 시설 데이터: Compute Atlas (CC BY 4.0), OpenStreetMap 기여자 (ODbL, 모델 입력으로만 사용). 전력망 연계 데이터: LBNL Queued Up, GridTracker (CC BY 4.0). 점수는 후보를 좁히는 데 도움을 주는 참고 자료이며 투자 조언이 아닙니다.