CCS 프로젝트의 운영 안정성 예측을 위한 머신러닝 접근

송리안

대전대신고등학교(Daejeon Daeshin High.) A.C.T.(KE)

ABSTRACT: 본 연구는 Global CCS Institute 에서 공개한 407 개 CCS 프로젝트의 메타데이터를 기반으로, 프로젝트의 운영 활성화 상태(Active/Non-Active)를 예측하는 데이터 기반 모델을 구축하였습니다. 연구 과정에서 위도·경도와 같은 지리적 정보와 DOE 지원, 지역 파트너십 참여 여부 등 제도적 요인을 주요 설명 변수로 활용하였습니다. 최종적으로 랜덤포레스트와 XGBoost 라는 두 가지 머신러닝 알고리즘을 적용하여 모델을 학습시킨 뒤, ROC-AUC 및 변수 중요도 분석을 통해 각 모델의 예측 성능과 주요 영향 요인을 비교 분석하였습니다.

I. 서론

탄소 포집 및 저장(Carbon Capture and Storage, CCS)은 기후변화 대응의 핵심 기술 중 하나로, 산업 공정에서 발생한 이산화탄소를 포집해 압축 후 지하 깊은 지층에 저장함으로써 대기 중 배출을 줄이는 원리를 갖는다. CCS 기술은 포집(Capture), 운송(Transport), 저장(Storage)의 세 단계를 포함하며, 각 단계의 기술 효율성과 환경적 안정성이 전체 시스템의 성공 여부를 결정한다.

기본 이론적 배경으로, 저장층은 일반적으로 다공성 암석(예: 사암)이며 그 위를 밀폐층(cap rock)이 덮고 있어 CO_2가 상부로 새어 나가지 않도록 한다. 주입된 CO_2는 초임계유체(supercritical fluid) 상태로 존재하며, 압력, 온도, 공극률, 투수율 등 물리적 변수에 따라 거동이 달라진다. 시간이 경과함에 따라 용해, 모세관 포획, 광물화 반응 등의 과정을 통해 장기 고정이 이루어지며, 이 과정을 '지질학적 격리(geological sequestration)'라고 한다. 대표적인 상용화 사례로 노르웨이의 Sleipner 프로젝트와 캐나다의 Weyburn 프로젝트가 있다.

지질 저장 안정성(stability)은 지층 구조의 연속성과 밀폐성, 지진 활동, 단층 및 균열의 존재, 주입 압력과 온도 변화 등 다양한 지질·역학적 요인에 의해 결정된다. 주입 압력이 과도하게 상승할 경우 지층 파열이나 단층 미끄러짐이 발생할 수 있어 CO_2 누출(leakage) 가능성이 커지므로, 압력 관리와 실시간 모니터링이 중요하다. 이러한 안정성 평가는 수치해석 모델링, 지진계 및 센서 기반 데이터 수집과 함께 머신러닝을 활용한 리스크 예측 모델을 결합하여 수행하는 것이 바람직하다.

연구의 목적은 다음과 같다. 첫째, 전 세계 CCS 프로젝트 데이터를 활용하여 지리적·제도적 요인이 프로젝트 운영 안정성에 미치는 영향을 정량적으로 분석한다. 둘째, 머신러닝 기법을 적용해 데이터 기반 리스크 예측 모델을 개발함으로써 향후 지질 저장소의 안전성 평가 및 정책적 의사결정 지원에 기여할 수 있는 틀을 제시한다.

본 연구의 필요성은 다음과 같다. 현재 CCS 기술의 상용화는 빠르게 진행되고 있지만, 저장소의 장기적 안정성 검증과 모니터링 체계는 아직 미흡하다. CO_2 누출과 지진 유발 가능성 등 안전성 문제는 사회적 수용성을 저해하는 주요 요인으로 작용하고 있으며, 이에 대한 과학적 검증과 예측 체계의 확립이 시급하다. 따라서 기존의 현장 중심 평가 방식에 더해, 대규모 데이터 분석과 인공지능 기술을 활용한 사전 위험 예측 체계의 도입이 필요하다.

이 연구는 이러한 기술적·사회적 요구를 충족하기 위한 기초 연구로서, 공개된 데이터를 활용해 머신러닝 기법이 지질 저장 안정성 평가에 어떻게 적용될 수 있는지를 실증적으로 검토하고자 한다.

II. 연구 방법

1. 데이터셋과 변수 선정

연구에는 Global CCS Institute 의 "CCS Map Data Jan 2023" CSV 파일을 사용하였다. 원자료는 417 개 프로젝트, 39 개 컬럼으로 구성되어 있으며, 이 중 본 연구에서는 다음 여섯 개의 설명 변수를 선택하였다.

- Latitude: 프로젝트 위치의 위도
- Longitude: 프로젝트 위치의 경도
- DOE Support: 미국 에너지부(Department of Energy)의 지원 여부
- Exact Checkbox: 위치 좌표의 정확도 표시(정확한 좌표 여부)
- Paper: 관련 논문·자료 존재 여부
- Regional Partnership: 지역 파트너십 프로그램 참여 여부

목표 변수(Target)는 프로젝트의 전체 상태를 의미하는 Overall Status 로 설정하고,

- Active = 1
- 그 외 상태(예: Hold, Planned 등) = 0

으로 이진 분류 문제로 변환하였다.

Overall Status 에 결측값이 있는 10 개 행을 제거한 후, 최종 분석에는 407 개의 프로젝트 데이터가 사용되었다. 위도·경도에는 각각 19 개의 결측값이 존재했으며, 평균값으로 대치하였다.

2.데이터 분할 및 학습 설정

전처리된 데이터셋은 설명 변수 6 개, 목표 변수 1 개로 구성되어 있으며, 전체 데이터의 특성을 고려하여 학습용과 테스트용 데이터로 8:2 의 비율로 분할하였다. 분할 과정에서는 클래스 불균형을 방지하기 위해 Stratified Sampling 방식을 적용하였다. 또한, 데이터의 신뢰성을 확보하기 위해 학습 전에 변수 간의 단위 차이를 확인하고 필요에 따라 표준화(Standardization) 과정을 실시하였다.

모델 학습 과정은 다음과 같은 단계로 체계적으로 진행되었다. 먼저, 데이터의 특성에 따라 비선형적 패턴을 잘 포착할 수 있는 트리 기반 모델을 선택하였다. 이를 위해 두 가지 알고리즘, 즉 랜덤포레스트 (RandomForestClassifier)와 XGBoostClassifier 를 비교하였다. 두 모델 모두 다수의 결정트리를 앙상블하여 예측 성능을 향상시키는 구조를 가지고 있으나, XGBoost 는 부스팅(boosting) 기법을 적용하여 오차를 반복적으로 보정하는 특징을 지닌다.

RandomForestClassifier 는 n_estimators 를 200 으로 설정하여 충분한 트리 개수를 확보하였고, 각 트리의 깊이(max_depth)는 10 으로 제한하여 과적합(overfitting)을 방지하였다. class_weight 를 “balanced”로 지정해 클래스 간 비율 불균형 문제를 완화하였으며, 5-Fold Stratified 교차검증을 통해 모델의 일반화 성능을 평가하였다.

XGBoostClassifier 의 경우, n_estimators 를 300 으로 설정하여 반복 학습의 충분성을 확보하고, learning_rate 를 0.05 로 두어 학습 안정성을 높였다. 또한 subsample=0.8, colsample_bytree=0.8 로 설정해 각 학습 단계에서 데이터와 변수의 일부를 무작위로 선택함으로써 모델의 다양성과 강건성을 유지하였다. Objective 는 “binary:logistic”으로 지정하여 이진 분류에 최적화된 형태로 학습을 수행하였다.

모델 학습 후에는 각 알고리즘의 예측 결과를 Accuracy(정확도), Recall(재현율), ROC-AUC (수신자 조작 특성 곡선 아래 면적) 지표를 통해 종합적으로 평가하였다. 또한 ROC 곡선과 변수 중요도 (Feature Importance) 시각화를 통해 각 모델이 어떤 변수를 중점적으로 활용하여 예측을 수행하는지를 분석하였다. 이러한 과정을 통해 모델의 신뢰성과 해석 가능성을 동시에 확보하였다.

III. 결과

1.기술 통계 및 상관분석

전체 407 개 프로젝트 데이터를 분석한 결과, 위도는 평균 36.49°, 경도는 평균 -24.87°로 나타나 북반구 지역에 프로젝트가 집중되어 있었다. 이는 CCS 프로젝트가 기술 인프라와 경제력이 높은 선진국 중심으로 분포한다는 점과 부합한다. 목표 변수인 Overall Status 의 평균값은 약 0.624 로, 전체의 62%가 Active 상태였다. 이는 대부분의 프로젝트가 아직 운영 중이거나 일정 수준의 활동을 유지하고 있음을 시사한다.

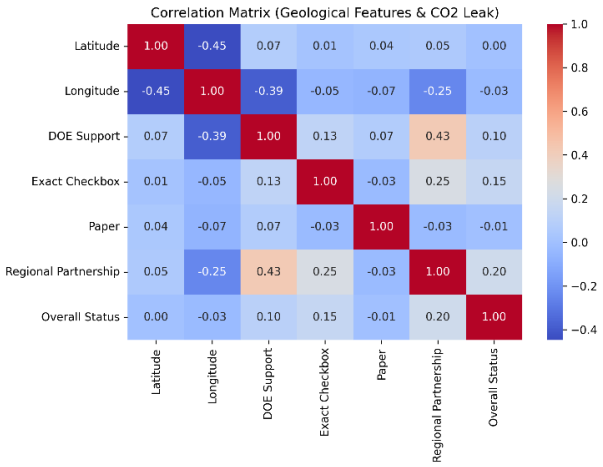

그림 1 상관분석

상관관계 분석 결과(그림 1), 위도와 경도는 -0.45 로 음의 상관을 보였으며, 이는 유럽 및 북미 지역의 지리적 군집을 반영한다. DOE Support 와 Regional Partnership 간에는 0.43 의 양의 상관관계가 확인되어, 정부 지원이 지역 협력 구조와 연계되어 이루어지는 경향을 보였다. 또한 Regional Partnership 와 Overall Status 의 상관계수는 0.20, Exact Checkbox 는 0.15, DOE Support 는 0.10 으로 나타나, 정책적·제도적 요인이 프로젝트 활성화와 일정한 상관성을 가지는 것으로 분석되었다. 이러한 결과는 CCS 프로젝트가 단순한 기술적 문제를 넘어 제도적 협력과 지원 체계에 의해 크게 영향을 받는다는 점을 뒷받침한다.

2.RandomForest 모델 성능 분석

RandomForest 모델은 5-Fold 교차검증 결과 AUC 평균값이 0.589 로 나타났으며, 테스트 데이터셋에서 Accuracy 0.561, Recall 0.608, AUC 0.583 을 기록했다. 이는 모델이 완벽히 높은 성능을 보이지는 않지만, 무작위 분류보다 안정적인 예측력을 보유함을 의미한다. 특히 활성(1) 클래스의 재현율이 0.61 로 비교적 높은 편이었는데, 이는 모델이 실제로 운영 중인 프로젝트를 더 잘 탐지하는 경향을 나타낸다.

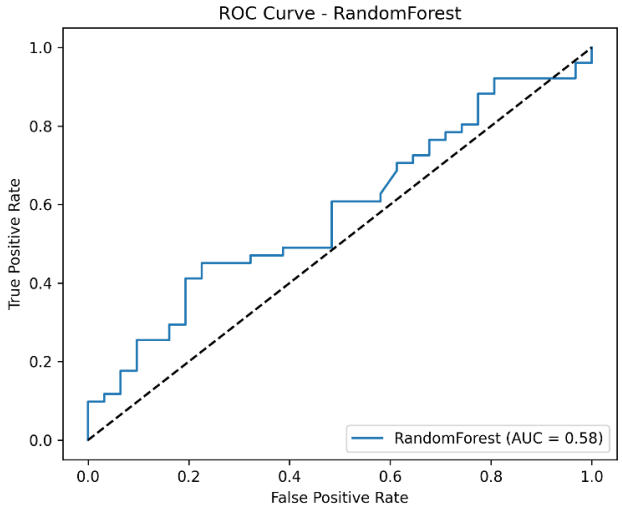

그림 2 random forest ROC 곡선

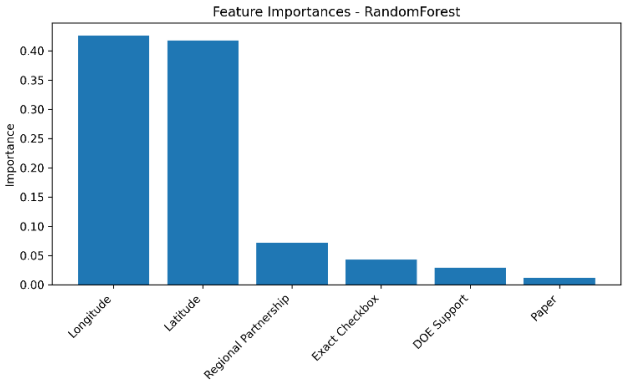

그림 3 random forest 변수 중요도 분석

정밀도-재현율 균형을 살펴보면 비활성(0) 클래스의 정밀도는 0.43, 활성(1) 클래스의 정밀도는 0.66 으로, 두 클래스 간의 예측 편차가 존재했다. ROC 곡선(그림 2)은 무작위 기준선보다 상단에 위치하여 일정 수준의 판별력을 확보하고 있으며, 이는 랜덤포레스트가 지리적 변수의 영향 패턴을 일부 포착한 것으로 해석된다. 변수 중요도 분석(그림 3)에 따르면, Longitude(0.426)와 Latitude(0.418)가 전체 예측 성능의 대부분을 차지했으며, Regional Partnership(0.072), Exact Checkbox(0.043), DOE Support(0.029), Paper(0.012)가 그 뒤를 이었다. 이 결과는 지리적 위치가 모델의 주요 판단 기준으로 작용함을 시사하며, 특정 지역의 프로젝트가 활성 상태일 가능성이 높다는 공간적 패턴을 반영한다.

3.XGBoost 모델 성능 분석

XGBoost 모델은 RandomForest 보다 높은 예측력을 보였다. 교차검증 및 테스트 결과, Accuracy 0.659, Recall 0.765, AUC 0.663 을 기록하며 모든 주요 평가 지표에서 향상된 성능을 보였다. 특히 Recall 이 0.76 으로 나타난 점은 모델이 Active 상태의 프로젝트를 놓치지 않고 탐지할 가능성이 높다는 것을 의미한다. ROC 곡선(그림 4)은 RandomForest 대비 기준선에서 더 크게 이탈하여 모델의 분류 경계가 더욱 명확해졌음을 시각적으로 보여준다.

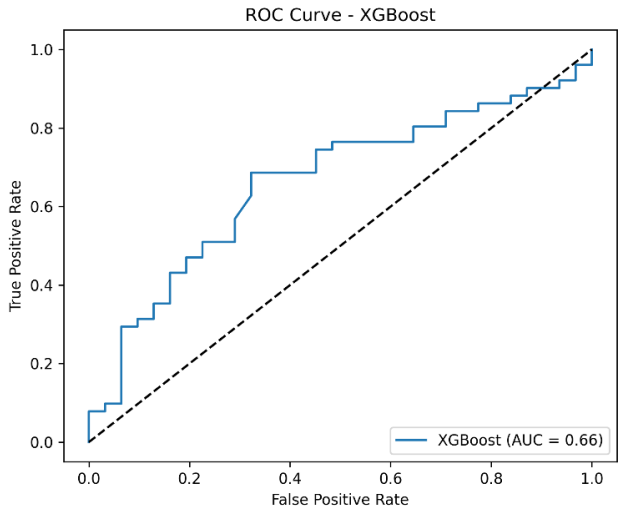

그림 4 XGBoost ROC 곡선

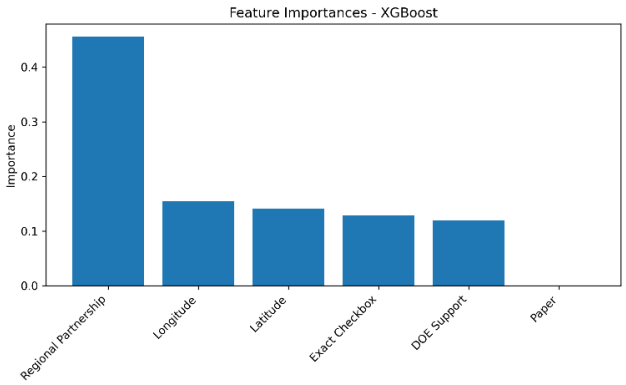

그림 5 XGBoost 분석

변수 중요도 결과(그림 5)에서는 Regional Partnership (0.456)이 단일 변수 중 가장 큰 비중을 차지하였으며, Longitude(0.155), Latitude(0.141), Exact Checkbox(0.129),

DOE Support(0.119)가 뒤를 이었다. Paper 변수는 중요도가 0 으로 나타나 프로젝트 문헌 정보는 상태 분류에 거의 영향을 미치지 않는 것으로 확인되었다. 이러한 결과는 제도적 요인, 특히 지역 파트너십 참여가 프로젝트의 지속성 및 활성화에 결정적인 역할을 할 수 있음을 시사한다.

4. 모델 비교 및 종합 평가

두 모델을 종합적으로 비교하면, XGBoost 는 Random-Forest 보다 정확도와 재현율, AUC 등 모든 지표에서 우수했다. RandomForest 는 단순한 패턴 인식에 강점을 보였으나 지리적 변수에 과도하게 의존하여 제도적 요인의 영향력을 충분히 반영하지 못했다. 반면 XGBoost 는 부스팅 구조를 통해 미세한 변수 간 상호작용을 학습하며, 지역 파트너십 및 정책 지원의 영향력을 효과적으로 포착하였다.

이 결과는 CCS 프로젝트의 성공 여부가 단순히 위치적 요인에만 달려 있지 않고, 제도적 협력 네트워크와 정책 지원 체계의 결합에 의해 크게 좌우된다는 점을 데이터 기반으로 검증한 것이다. 따라서 향후 안정성 예측 연구에서는 지질학적 특성과 더불어 제도적·정책적 맥락을 함께 고려한 복합 모델 접근이 필요하다.

IV. 논의

1.정책·제도적 요인의 중요성

상관분석과 XGBoost 변수 중요도 분석 결과를 종합하면, DOE Support 와 Regional Partnership 같은 제도적 지원 요인이 CCS 프로젝트의 활성 상태와 완만하지만 일관된 양의 관계를 보인다는 점이 드러났다. 특히 XGBoost 모델에서 Regional Partnership 가 가장 중요한 변수로 나타난 것은, 지질 특성 정보가 없어도 "어떤 제도적 네트워크에 속해 있는가"라는 정보만으로도 프로젝트 지속 여부를 어느 정도 예측할 수 있음을 의미한다. 이는 기술적 요인뿐 아니라 정책·거버넌스 구조가 CCS 프로젝트 성패에 핵심적으로 작용함을 시사한다.

2.모델 성능의 한계

한편 두 모델의 AUC 값이 0.58(RandomForest), 0.66 (XGBoost)에 그쳤다는 점은 이 데이터와 변수 구성만으로는 프로젝트 상태를 정확하게 예측하기 어렵다는 한계를 보여준다.

그 이유는 다음과 같이 해석할 수 있다.

- 입력 변수들은 위도·경도와 제도적 참여 여부 등 "메타데이터"에 해당하며, 실제 저장소의 지질 특성(심도, 공극률, 투수율, 주입 압력 등)을 전혀 포함하지 않는다.
- 목표 변수인 Overall Status 는 지질학적 안정성뿐 아니라 경제성, 정책 변화, 기업 전략 등 다양한 요인이 뒤섞여 결정되는 결과 변수이기 때문에, 단순 메타데이터로 설명하기 어렵다.

따라서 본 연구의 결과는 "지질 저장 안정성" 그 자체를 직접 예측한다기보다, "현재 공개된 메타데이터만으로 어느 정도까지 프로젝트의 운영 상태를 추정할 수 있는가"라는 관점에서 이해하는 것이 타당하다.

3. 지질 데이터 기반 확장 가능성

이번 분석에서 사용한 파이프라인(데이터 전처리–상관분석–RandomForest/XGBoost 학습–ROC·변수 중요도 해석)은 입력 변수만 교체하면 그대로 활용할 수 있다. 향후 실제 저장소의 심도, 저장층 압력·온도, 공극률·투수율, 누출 여부 기록 등이 포함된 데이터셋을 확보한다면,

- 목표 변수: CO_2 누출 여부 또는 누출률
- 설명 변수: 지질·운영 변수(심도, 압력, 주입량, 단층 구조 등)

으로 재구성하여, 진정한 의미의 "지질 저장 안정성 리스크 예측 모델"로 확장할 수 있을 것이다.

V. 결론

본 연구는 Global CCS Institute 의 공개 데이터를 활용하여 CCS 프로젝트의 메타데이터만으로 운영 상태를 예측할 수 있는 머신러닝 기반 분석 모델을 구축하였다. 이를 통해 위치적·제도적 요인이 프로젝트의 활성 여부에 미치는 영향을 정량적으로 파악하고, 정책적 지원 구조가 기술적 안정성에 미치는 영향을 실증적으로 분석하였다.

모델 비교 결과, XGBoost 가 RandomForest 보다 전반적으로 우수한 성능을 보였으며(Accuracy 0.659, Recall 0.765, AUC 0.663), 변수 중요도 분석에서 Regional Partnership 가 가장 핵심적인 요인으로 확인되었다. 이는 지역 파트너십 참여가 프로젝트의 운영 안정성 및 지속 가능성을 결정하는 중요한 제도적 메커니즘임을 시사하며, DOE Support 와 Exact Checkbox 또한 프로젝트의 관리 체계와 직접적으로 연결되어 있다는 점을 보여준다.

이러한 결과는 CCS 프로젝트의 성패가 단순히 기술적 또는 지리적 조건에 의존하지 않고, 제도적 네트워크와 정책적 협력 수준에 의해 크게 좌우된다는 점을 데이터로 뒷받침한다. 즉, 정부의 지속적 재정 지원, 지역 단위 협력체계 구축, 정확한 위치 기반 데이터

관리가 CCS 사업의 성공적 추진에 핵심적이라는 것이다.

한편 본 연구는 메타데이터 기반 분석이라는 점에서 지질학적 변수(예: 저장 심도, 압력, 공극률, 단층 구조 등)를 직접 반영하지 못하는 한계를 지닌다. 그러나 이번 연구에서 확립한 데이터 전처리–모델 학습–성능 평가–변수 중요도 해석의 파이프라인은 향후 실제 지질 및 운영 데이터가 확보될 때, 이를 그대로 확장하여 보다 정교한 예측 체계로 발전시킬 수 있는 구조를 제공한다.

결과적으로, 본 연구는 CCS 기술의 안정성 평가와 관련된 데이터 기반 접근의 가능성을 제시함과 동시에, 정책·제도적 지원이 기술적 안정성 확보에 미치는 영향을 경험적으로 검증하였다는 점에서 의의가 있다. 향후 연구에서는 다양한 지질·운영 데이터와 시계열 분석을 결합하여, 단기적 프로젝트 상태 예측을 넘어 장기적 저장 안정성 리스크를 평가하는 종합적 예측 모델을 개발하는 것이 필요하다.

참고 문헌

1.Global CCS Institute (2024). Global Status of CCS Report. Melbourne, Australia.

2.Arts, R., Eiken, O., Chadwick, A. et al. (2004). "Monitoring of CO_2 injected at Sleipner using time-lapse seismic data." Energy, 29(9–10), 1383–1392.

3.IPCC (2022). Carbon Dioxide Capture and Storage – Technical Summary. Geneva: Intergovernmental Panel on Climate Change.

4.Kaggle Dataset: Carbon Capture and Storage Projects (konradb/carbon-capture-and-storage), 2023.

5.Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." Annals of Statistics, 29(5), 1189–1232.

6.Bachu, S. (2008). "CO_2 storage in geological media: Role, means, status and barriers to deployment." Progress in Energy and Combustion Science, 34(2), 254–273.

7.Benson, S. M., & Cole, D. R. (2008). "CO_2 sequestration in deep sedimentary formations." Elements, 4(5), 325–331.