

Kwangmin Kim

Data Science Portfolio

Data Scientist / Data Analyst

Email: kmink3225@gmail.com | Website: kmink3225.netlify.app



자기소개



- 데이터 사이언스 분야 7년 경력
- 통계 및 머신러닝
- R, Python, SQL, Apache 활용
- 학력
 - 학사: 생화학 & 수학
 - 석사: 생물통계학
- 리더십
 - Project Leading & Management 경험 보유
 - 목표 설정, 전략적 계획 수립, 자원 할당 및 일정 관리
 - 데이터 기반 의사결정
 - 분석적 사고 및 객관적 판단
 - 소통 능력
 - 팀 내/외 협업 및 비전문가와의 커뮤니케이션



기술 스킬

프로그래밍 언어

- R (숙련)
- Python (중급)
- SAS (초급)

도구 및 프레임워크

- R Shiny (중급)
- Apache Airflow (중급)

데이터베이스

- 데이터 거버넌스
- SQL

문서화 및 보고서

- Quarto
- R Markdown
- Jupyter

통계

- 회귀분석 및 생존분석
- 시계열 분석

머신러닝

- 예측, 분류 및 군집
- Elastic Net, Random Forest, Boosting, SVM, PLS, sparse PLS, PCA, DB SCAN, etc.



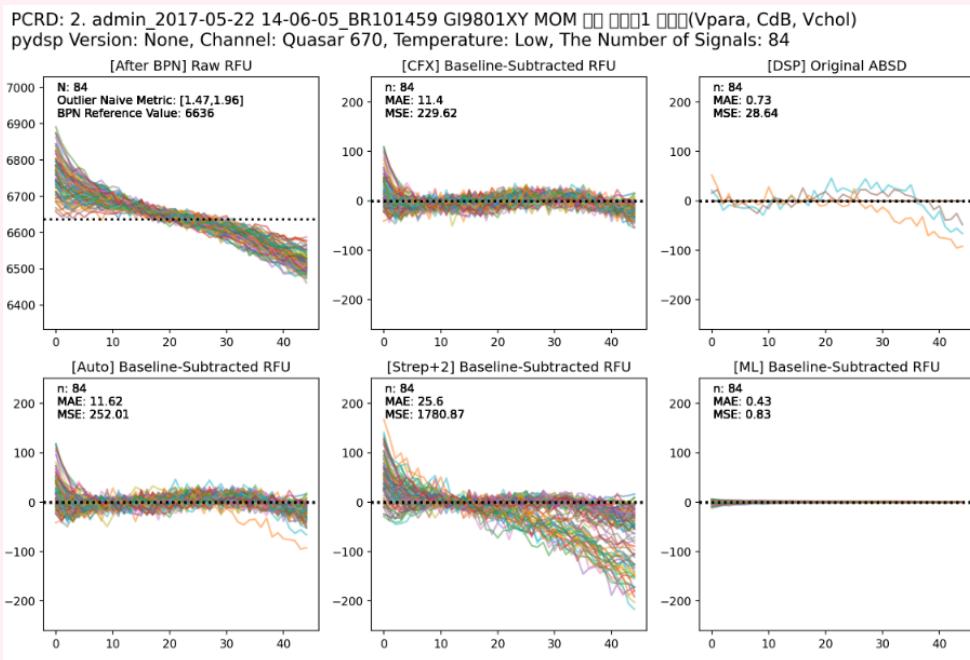
[주요 프로젝트 1] 진단 알고리즘 최적화

- **프로젝트:** Baseline Fitting Algorithm 최적화 (2024.01 - 2024.07)
- **역할**
 - 참여 인원: 데이터 사이언티스트 3명, 데이터 엔지니어 2명, 생물학자 2명
 - Project Manager: 데이터 파이프라인 및 분석 방법론 기획
 - Data Scientist: 데이터 전처리, 알고리즘 최적화 (저수준 코딩) 및 기획안 구현
- **주요 성과**
 - 시그모이드 곡선의 베이스라인 피팅 알고리즘 개선
 - 데이터 증강 및 다양한 베이스라인 피팅 방법 비교 분석
 - 알고리즘 개선: Rule-Based에서 Data-Driven으로의 전환
- **사용 기술**
 - Python, Quarto
 - 통계 분석, 특성 방정식, 신경망
- **영향**
 - 진단 정확도 향상 (위양성 감소)
 - 정량 지표의 이상치 발생 및 변동성 감소
- **주요 링크:** [프로젝트 세부 정보](#)



[주요 프로젝트 1] 진단 알고리즘 최적화

- 주요 결과



Baseline Fitting Algorithm Comparison

- 1st Pannel [After BPN]: Normalized Raw Data
- 2nd Pannel [CFX]: (대조군1) 타사의 **Black Box** 알고리즘 결과
- 3rd Pannel [DSP]: (대조군2) DS팀의 **Legacy Rule-Based** 알고리즘의 결과
- 4th Pannel [Auto]: (대조군3) 생물 실험자들의 **Bio Legacy Rule-Based** 알고리즘의 결과
- 5th Pannel [Strep+2]: (실험군1) 팀원(물리학자)의 **Improved Rule-Based** 알고리즘의 결과
- 6th Pannel [ML]: (실험군2) 본인의 **Data Driven ML** 알고리즘의 결과



[주요 프로젝트 2] 알고리즘 안정성 검증 문서화

- **프로젝트:** FDA 알고리즘 Verification & Validation 문서 자동화 (2023.01 - 2023.11)
- **역할**
 - 참여 인원: 데이터 사이언티스트 5명, 데이터 엔지니어 3명 외 관련 실무자 29명
 - Project Manager: FDA 참고문헌 분석, 타부서 Needs 반영, 데이터 파이프라인 및 분석 방법론 기획
 - Data Scientist: 데이터 엔지니어링, 알고리즘 안정성 검증, 기획안 구현 및 문서 자동화
- **주요 성과:**
 - 알고리즘 안정성 검증 모델 디자인 및 개발
 - System Level의 통계적 테스팅 모델 기획 및 구현
 - 데이터에 따른 알고리즘 및 시약 안정성 검증 및 검증 보고서 자동화 시스템 구축 (반자동화)
- **사용 기술**
 - R, Quarto, Airflow
 - 통계 분석 (저수준 코딩)
 - χ^2 Test, the McNemar Test, the Breslow-Day Test, the Cochran-Mantel-Haenszel Test
- **영향**
 - 기존에 부재했던 알고리즘과 시약의 성능 평가 체계 확립
 - 미국 (FDA)이나 캐나다 (Health Canada)를 포함한 북미시장 진출 대비
 - 목표 시약 제품 실험 data에 대한 알고리즘의 성능 정확도 100% 입증
- **주요 링크:** [프로젝트 세부 정보](#)



[주요 프로젝트 2] 알고리즘 안정성 검증 문서화

- 주요 결과



산출물

5.3.1 Target Template S

In the target template S, the DSP algorithm showed 100% accuracy on average with 0% standard deviation across both the concentration and scenario levels (See Figure 5.2).

- For NC signals, the DSP algorithm scenarios showed 99.7753% accuracy with the standard deviation, 0.4552%.
- For PC signals, the DSP algorithm scenarios showed 99.8876% accuracy with the standard deviation, 0.3414%.
- scenario00 showed 100% accuracy on average across the different concentration levels with 0% standard deviation and .
- scenario99 showed 97.027% accuracy on average across the different concentration levels with 6.8398% standard deviation.
- scenario02 recorded the lowest accuracy, 97.027% with the standard deviation 6.8398% across the concentration levels.
- scenario02 recorded the highest standard deviation of accuracy, 6.8398%.
- the DSP algorithm scenarios showed the lowest accuracy 96.1202% with the standard deviation, 7.8584% in the negative concentration.
- The lowest accuracy of scenario02 in the negative concentration results from the false negatives, 0 cases and the false positives, 413 cases with the p-value, < 0.001 (See Table 5.4 and Table 5.5).
- A list of false negative and false positive cases related to concentration levels and scenarios is displayed in Table 5.6.

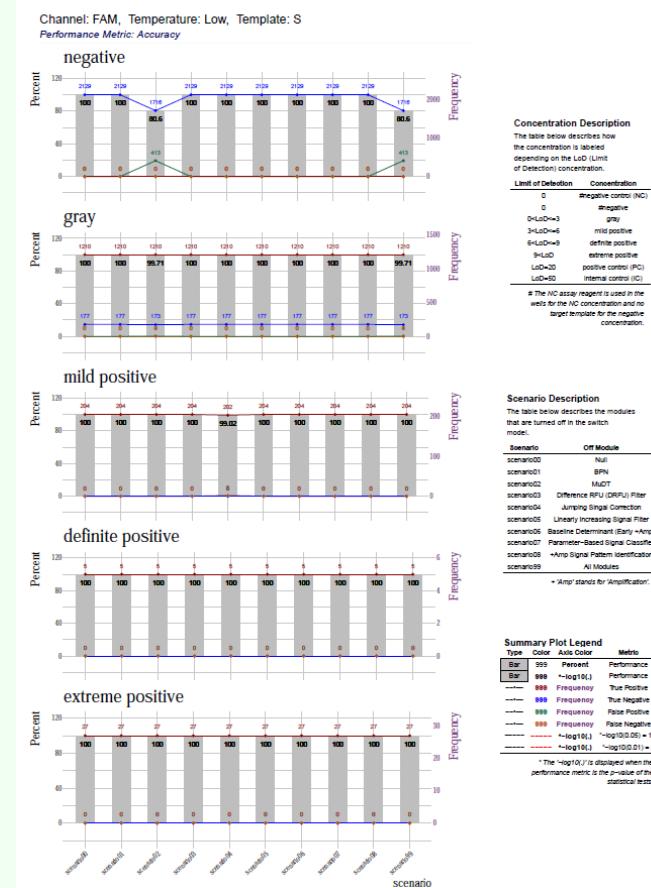
Table 5.4: Frequency Summary of DSP Performance for the Negative Concentration

template	concentration	scenario	false_negative	false_positive	true_negative	true_positive
S	negative	scenario00	0	0	2129	0
S	negative	scenario01	0	0	2129	0
S	negative	scenario02	0	413	1716	0
S	negative	scenario03	0	0	2129	0
S	negative	scenario04	0	0	2129	0
S	negative	scenario05	0	0	2129	0
S	negative	scenario06	0	0	2129	0
S	negative	scenario07	0	0	2129	0

59

Documentation Example

Figure 5.2: DSP Accuracy Performance in the Target Template S



Algorithm Performance Summary and Insights

Email: kmink3225@gmail.com | Website: kmink3225.netlify.app



[주요 프로젝트 3] 진단기기 QC Platform 구축

- **프로젝트:** 진단기기 QC Process 개선 및 QC Platform 구축 (2021.01 - 2021.09)
- **역할**
 - 참여 인원: 데이터 사이언티스트 1명, Full Stack 개발자 5명, 자문 교수 1명 외 관련 실무자 11명
 - Project Manager: 현업 부서 Needs 반영, 프로젝트 이슈 정리, 데이터 확보 및 분석 방법론 기획
 - Data Scientist: 데이터 엔지니어링, 진단 기기 QC 알고리즘 분석, 개선 및 구현
- **주요 성과:**
 - 웹 기반 QC 반자동화 플랫폼 구축
 - 진단 기기 평가 등급 도입
 - 기존 QC 프로세스 중 소요시간이 많은 불필요한 noise test 폐지
- **사용 기술**
 - R, python, 통계, 수리적 mechanistic modeling, Levenberg–Marquardt algorithm
- **영향**
 - President's Award, R&D 부문 우수상
 - 2개의 특허 발명
 - QC 프로세스에 소요되는 시간이 153배 이상 감소
 - 연간 6억 원(\$450,000)에 달하는 QC 비용을 13배 감소
- **주요 링크 :** [프로젝트 세부 정보](#)



[주요 프로젝트 3] 진단기기 QC Platform 구축

- 주요 결과

Noise Test	As-Is	To-Be
QC 알고리즘 개발에 사용된 샘플 크기	n=100	Signals from 2552 experiments, n=61,248
QC 알고리즘 성능 비교에 사용된 샘플 크기	n=61,248	n=61,248
Evaluation Metrics	2 metrics	10 metrics (the existing 2 metrics + new 8 metrics)
Input Process	특정 프로그램에서 추출한 엑셀 파일의 데이터를 수동으로 복사하여 붙여넣기	웹 기반 자동화, 다수의 실험 파일 업로드
Output Process	Batch Evaluation method로서 장비의 신호 중 하나라도 부적합 판정되면 장비 자체가 QC 부적합 판정 (맹점: 휴면에러 신호가 1개라도 있으면 장비는 무조건 실격 처리)	Differential Evaluation method, 장비의 신호에 점수를 계산 후 평균값을 구하고 장비 등급을 A+, A, B, F로 지정. F인 경우 부적합. 오류 신호를 평가에서 제외하므로 오류 신호에 robust
Output 1	pass: 92.58%, fail: 7.42% (after excluding many human errors)	A+ (pass): 7.01%, A (pass): 12.91%, B (pass): 75.72%, F (fail): 4.36%
Output 2	NA	Visualized Plots and Tables.
Output 3	NA	Classification Results: Normal Signals, Human Errors, Device Errors, Manufacturing Errors
Time Consumed	About 30 minutes per 20 experiments	About 25 minutes per 2552 experiments
Data Management	Non-standard management method (작업자마다 다른 방식으로 Excel 파일로 다른 형태로 NAS 디렉토리에 저장)	RDB uploaded by a scheduler (장비 고장 주제 분석이 가능)

산출물

Email: kmink3225@gmail.com | Website: kmink3225.netlify.app



[주요 프로젝트 4] Long Life Family Study(LLFS)

- **프로젝트:** LLFS 치매와 유의한 관계가 있는 Metabolic Profiles 규명 (2018.12 - 2020.04)
- **역할**
 - 참여 인원: CUIMC Faculty 4명, 통계학자 2명, 역학자 1명, 생화학자 1명
 - Data Scientist: 데이터 QC, EDA, 데이터 마이닝, 분석 파이프라인 구축, 통계 분석 및 ML 기법 비교분석
- **주요 성과:**
 - 8개월 동안 연구소에서 파악하지 못한 강력한 교란자를 EDA와 데이터 마이닝을 통해 발견
 - Columbia University의 Mailman School of Public Health 연례 연구 발표회에서 포스터 발표
 - 연례 연구 경진대회에서 약 100명의 대학원생 중 상위 3명으로 선정되어 상금 \$1,000 및 학과장상 수상
 - Columbia University Irving Medical Center 신경외과 Job Offer
- **사용 기술**
 - R, EDA, Data Mining, 통계 분석 및 ML
- **영향**
 - 분류 정확도 약 20% 증가 (from 65% to 88%)
- **주요 링크 :** [프로젝트 세부 정보](#)
- **주요 결과**
 - 146개의 관측치와 약 3,000여개의 변수로 구성된 data에서 약 60개 내외의 대사물질이 질병과 5% 유의수준으로 유의한 관계가 있는 것으로 관찰됐고 partial least square 가 가장 성능이 좋은 것으로 관찰됐다.



[주요 프로젝트 5] 미분방정식을 이용한 찻잎의 중금속 흡착 과정 메카니스틱 모델링

- **프로젝트:** 중금속 오염수 처리를 위한 찻잎 흡착 동역학: 미분방정식 모델링 접근 (2015.01 - 2015.06)
- **역할**
 - 참여 인원: CUNY Faculty 2명, 수학자 2명, 생물학자 1명
 - Researcher: 메카니스틱 모델링 및 프로그래밍
- **주요 성과:**
 - mechanistic modeling 결과 파라미터의 유의성이 높은 것으로 관찰됐다.
- **사용 기술**
 - R, 미분 방정식
- **영향**
 - New York City College of Technology (CUNY) 포스터 발표
 - BMCC (CUNY) 포스터 발표
 - \$1,000 Stipend (CUNY)
 - Manhattan College에서 개최되는 2015 연례 모임 기고 논문 및 포스터 세션을 위한 발표회에서 포스터 발표
- **주요 링크 :** [프로젝트 세부 정보](#)



[주요 프로젝트 5] 미분방정식을 이용한 찻잎의 중금속 흡착 과정 메카니스틱 모델링

- 주요 결과

$$\arg \min_k \sum_{i=1}^n (q(t_i, k) - \hat{q}(t_i))^2$$

- $S(t)$ is the number of heavy metal molecules adsorbed to the tea leaves at time t .
- $W(t)$ is the number of heavy metal molecules in the water (and not adsorbed) at time t .
- $W_0 = W(0) = S(t) + W(t)$
- $q(t)$ is the fraction of heavy metal molecules adsorbed out of the waste water at time t
- $q(t) = \frac{S(t)}{W_0}$
- $(t_i, \hat{q}(t_i))$ is the observed values,
- $q(t, k) = q_e \frac{1-e^{-kt}}{1-q_e e^{-kt}}$

수식

Adsorption of heavy metals onto tea leaves

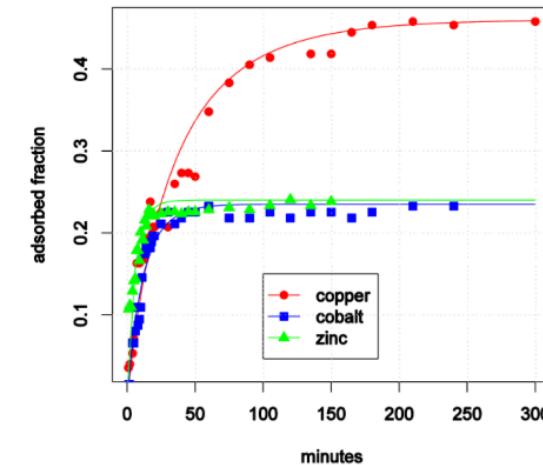


Figure : Graph of $q(t) = q_e \frac{1-e^{-kt}}{1-q_e e^{-kt}}$. The solid circles, squares, and triangles are data points. See table below for values for values of q_e and k .

Result

Metal	q_e	parameter	Estimate	Std. Error	t value	$Pr(> t)$	df	n
copper	.46	k	0.01281219	0.0008855	20.47	< 2e - 16	34	35
cobalt	.235	k	0.069941	0.003279	21.33	< 2e - 16	34	35
zinc	.24	k	0.150255	0.009133	16.45	< 2e - 16	32	33



교육 및 수상

교육

- Columbia University, CU (2017-2019)
 - 생물통계학 석사
- Baruch College, CUNY (2015-2017)
 - 수학 학사
- Rennert (2013-2015)
 - ESL Program, SIT TESOL
- 강원대학교, KNU (2006-2012)
 - 생화학 학사, 수석 졸업

수상 및 자격증

- 특허 출원 및 Know-How 발명 8건 (Seegene 2021 - 2023)
- Presient's Award, R&D 부문 우수상 (Seegene 2021)
- Chair's Award, 생물통계학 졸업 연구 경진대회 우승 (CU 2019)
- Stipend, \$1,000 대사체학을 위한 기계학습법 비교 연구 (CU 2019)
- Stipend, \$1,000 Mathematical Kinetic Modeling (CUNY 2015)
- SAS Certified Base Programmer (SAS 2018)
- SIT TESOL Instruction Certification (SIT 2014)
- Stipend, \$5,000 의료 융합 연구 (KNU 2012)
- 학장상, 성적 우수 수석 졸업 (KNU 2012)
- 전액 장학금, 성적 우수 장학금 (KNU 2010 - 2012)
- 사단장 표창, 리더쉽 경연 최우수상 (군대 2009)
- 중대장 표창, 사단 검열 우수상 (군대 2009)



연락처 및 링크

- Email: kmink3225@gmail.com
- Website: kmink3225.netlify.app
- LinkedIn: linkedin.com/in/kwangmin-kim-a5241b200
- GitHub: github.com/kmink3225

데이터를 통한 혁신과 가치 창출에 기여하고 싶습니다.

질문이나 협업 제안은 언제든 환영합니다!

[PDF로 다운로드](#)

