학부생의 입장에서 바라본 2020 해커톤 수상작 감상문

“단기간 동안 공부를 진행한 학부생의 입장에서 바라본 해커톤의 수준은 어느 정도인가?”

# 개요

최근 FACTORYHACKKOREA에서 진행한 해커톤 우수 수상작들의 자료를 접할 수 있는 기회를 얻게 되었습니다. 수상작들 모두 매우 높은 수준의 머신 러닝 기법들을 사용하여, 주어진 문제를 적절하게 해결한 것을 확인할 수 있었습니다. 그런데 이번에, 단순히 이러한 자료들을 읽고 이해하는 것으로 넘어가는 것이 아니라, 현재 머신러닝을 8개월 남짓 공부한 학부생의 시선에서 이러한 산업체 데이터 기반 머신러닝 설루션을 바라보았을 때, 어느 범위까지 이해할 수 있었고, 어떤 부분은 이해하기 어려웠고 와 같은 경험을 감상문으로 작성하여, 향후 ‘AI 기반 머신러닝의 프로젝트에 참여할 수 있는 AI 엔지니어링 전문 가는 어느 정도 수준이 되어야 하는가?’ ‘AI 기반 머신러닝의 설루션 참여하는 사람이 일정 수준 이상 올라가기 위해선 어느 정도까지 공부와 경험을 병행해야 하는가?’에 대한 간단한 지표가 되었으면 합니다

## 글쓴이의 현 상태(status)

* 1.컴퓨터 공학과 4학년 재학중
* 2.IBM AI Engineering professional certificate 취득 [자격증 설명 주소](https://www.coursera.org/professional-certificates/ai-engineer#courses)
* 3.Pytorch 와 tensorflow 2.xx에 대한 기본적인 지식 습득 완료
* 4.현재 컴퓨터관련 공부를 진행한지 약 8개월

Javascript 라던지 Flutter 라든지 다른 여러가지 분야들 도 접해봤지만, 이 보고서는 AI 머신러닝에 관해 적은 감상문임으로, 그 분야와 관련된 내용들만 기입했습니다.

## 이번 해커톤이 독특했던 이유 그리고 그에 따른 수상작들의 방향성

이번 해커톤의 데이터는 Oil Cooler에 관한 데이터 셋을 바탕으로 냉각기 부품을 제조하는 업체에서 설계요소와 그에 따른 냉각 성능 및 압력 손실을 측정한 데이터를 이용하여, 제조된 냉각기의 냉각 성능과 압력 손실 결과를 예측하는 모델을 개발하고, 그것의 정확도를 측정하는 것을 주제로 가집니다.

이번 해커톤의 굉장히 특이한 점이 있는데, 바로 데이터가 일반적인 해커톤에서 사용되는 데이터와 달리 상당히 이질적이라는 것입니다. 일반적인 AI 해커톤에서 사용되는 데이터는 주로 ‘특정 서비스를 이용하는 사용자의 데이터’ ‘특정 조건을 파악하기 위해 모여진 이미지, 영상 데이터’ 와 같은 어느 정도 대중에게는 친숙한 데이터들을 가지고 서비스의 품질을 향상시키거나, 이미지를 분류하는 모델을 만들거나 혹은 그 둘 모두를 만족시키는 모델을 만드는 것을 목적으로 해커톤이 진행됩니다. 그것들과는 다르게 FACTORYHACKKOREA에서 진행한 이번 해커톤의 경우, 일반적인 AI 머신러닝을 공부하는 사용자들에게 전혀 친숙하지 않은 ‘냉각기’및 기타 제품의 성능 향상을 위한 모델을 제작하는 것을 목적으로 진행되었습니다. 실제로, 수상을 했던 팀들 모두 데이터의 분석에 매우 많은 시간을 사용한 것을 알 수 있었습니다. 또한, 해커톤 후기를 적은 내용에 모든 팀들이 ‘데이터가 익숙하지 않아 분석하는데 많은 시간이 걸렸다.’ ‘데이터가 복잡하여 해커톤을 진행하는데 어려움이 많았다’ 와 같은 반응들이 있었습니다. 실제로 저 또한 데이터를 살펴보았을 때, 하나의 모델마다 이렇게 많은 요소(Feature)들이 있는 데이터를 다뤄본 적이 없었습니다. 데이터 셋은 보통 많아봐야 10개의 요소(Feature)들로 각각의 인과관계가 상당히 명확해 데이터를 처리하는데 큰 문제가 없었습니다. 그러나 이번 해커톤에서 사용되었던 데이터 셋은 다릅니다. 여러 가지 모델마다 수많은 요소(Feature)들이 있었고, 냉각기라는 제품에 대한 기술적인 내용 또한 알지 못하기에 상당히 어려운 해커톤 데이터였다고 할 수 있습니다. 사실, 산업체에서 생산되는 대부분의 데이터는 매우 거칠고 잡음이 많은 데이터이기에 이는 당연한 결과였다고 생각합니다.

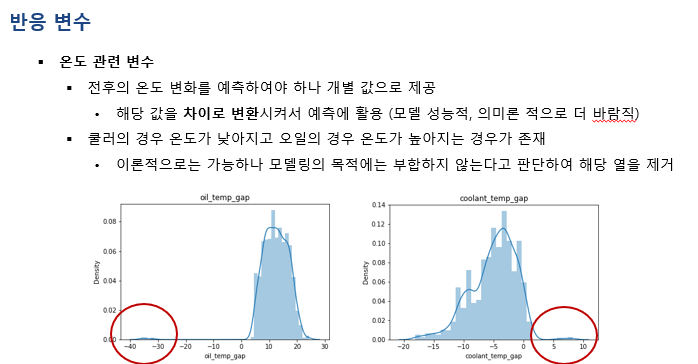
## 학부생의 시선에서 바라본 해커톤 솔루션

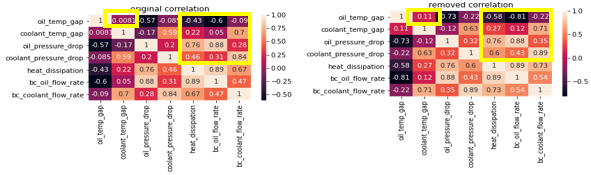
우선 이번 해커톤 설루션 중 대상을 받은 ’Mapokid’를 제외한 ‘SEA’ 와 ‘탑 바텀 정글’ 두 팀을 위주로 분석한 글을 적어 보았습니다. 특히 거기서 SEA를 좀 더 중점적으로 다루었습니다.(‘탑 바텀 정글’과 ‘SEA’모두 비슷한 방향성으로 접근을 시도하였습니다. 그러나 SEA가 조금 더 심도 있게 접근한 모델이기에 SEA를 위주로 아래를 설명하겠습니다.) ‘Mapokid’를 제외한 이유는 모델 자체가 상당히 복잡하고, 또한 ppt의 설루션 또한 한 차원 더 깊게 들어갔기에 이것을 이해하는데 매우 큰 어려움을 겪었기 때문입니다. (실제로 Mapokid만이 학부 이상 석사 재학 중인 대학원생들로 구성된 팀이었습니다.) 아래 설루션들에 대한 감상문은 필요한 부분의 PPT 화면을 첨부하여 그것에 대한 추가적인 설명과 생각을 적는 방식으로 작성하였습니다.

모든 설루션들은 아래와 같은 요구 사항을 반영하여 제작되었습니다.

* 예측 모델 개발
* 변수의 설명력 도출(XAI 방법론)
* 최적 설계 Requirement 도출

## SEA

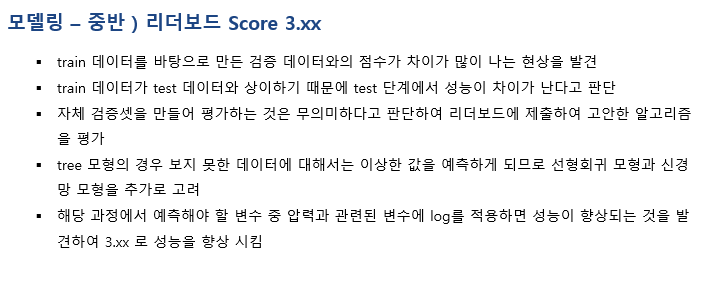


SEA 팀에서는 우선 데이터의 가공을 진행하였습니다. 데이터에 사용되는 변수들을 머신러닝에 적절한 값으로 바꾸어 사용하였습니다. 예를 들어, 위 온도 관련 변수를 개별적인 값들이 아닌, 온도의 차이로 변환시켜서 예측에 사용하였습니다. 또한, 극단적인 값들(아웃라이어에, 비정상적인 범위에 해당하는 값들)을 제거하여 데이터를 좀 더 다듬었습니다. 그리고 위에 데이터 가공의 결과를 Heat map을 통해 전반적인 변수들의 상관관계들이 더 강해진 것을 확인하는 과정을 거쳤습니다.(이는 실제로 데이터에 들어가는 실험 변수들과 결과를 보여주는 변수들 관의 상관관계가 더 강해져, 머신러닝 시 더 효율적인 학습과 결과를 가져온다는 것을 의미합니다.)

이상치 또한 제거하였으며. 카테고리 변수 (간단하게 수치상으로 표현되는 것이 아닌 변수들, 예를 들어 남성과 여성을 표현하는 SEX 변수가 카테고리 변수에 해당합니다.)를 숫자 형 데이터로 변경, One hot encoding과 같은 방식으로 카테고리형 데이터를 머신러닝에 적절한 수취 데이터로 변경하여 데이터 가공을 추가로 더 진행하였습니다. 마지막으로, 길이와 직경 그리고 좌표와 같은 수취 데이터를 자신들이 생각한 기준으로 가공하여 데이터 가공을 마쳤습니다.

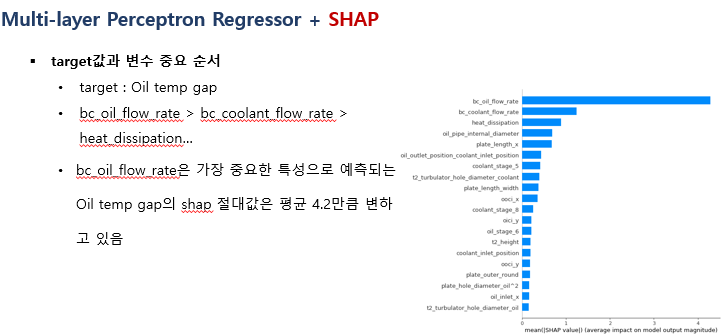
SEA 팀 같은 경우, 모델 개발에 관한 내용을 자세히 다루지는 않았습니다. 코드를 완벽하게 살펴보지는 않았지만, 아마도 데이터 가공을 완료했기에, 라이브러리의 도움을 받으면 모델의 제작은 매우 간단히 끝나기 때문입니다.

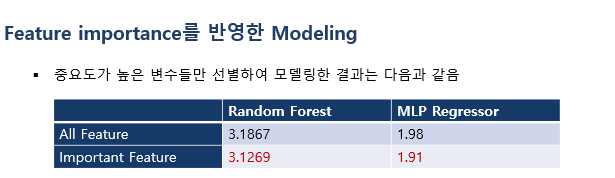
SEA팀과 다른 팀들 모두 여러가지 머신러닝 모델을 사용하여 정확도를 비교했습니다. 이것을 앙상블(Ensemble) 방식으로 적절한 머신러닝 모델을 찾을 때 주로 사용하는 방법입니다.

SEA 팀의 모델링 변화는 데이터 셋의 가공과 함께 진행되었습니다. 우선, 기본적으로 주어진 데이터를 tree 기반 모델에 적용하여 (Decision Tree를 사용하였습니다. 복잡하지 않아 초기 모델에는 매우 적합한 머신러닝 모델입니다.) 다음으로는 랜덤 포레스트(Random Forest – decision tree의 오버 피팅이라는 문제를 해결하기 위해 나온 모델입니다. 현재 주어진 데이터 셋은 매우 복잡하고 여러 가지 요소(Feature)들로 구성되어 있기에, 훈련된 모델의 예측 성능이 떨어지는 문제가 발생할 수 있습니다. 이것을 해결하기 위해 제시된 모델이 바로 Random Forest입니다.)

다음으로 SEA 팀은 정확도를 조금 더 올리기 위해, 학습에 사용하는 데이터와 검증에 사용하는 데이터를 다시 한번 살펴보는 과정을 거쳤습니다. 최종적으로는 신경망 모델을 사용하여 정확도를 초기보다 향상시켰습니다. 여기서 왜 초기부터 신경망 모델을 적용하지 않았는지 의문이 들 수도 있는데, 신경만 모델은 매우 복잡한 모델을 만드는데 매우 훌륭한 방법이지만, 학습시간이 길고 데이터 전처리에 매우 민감합니다. 또한, 데이터가 동질성이 높을수록 (예를 들어 Feature가 고정돼있고 데이터들이 정돈이 잘 된 경우) 사용하기 적절합니다. 위와 같은 조건들을 충족하지 못할 경우, 트리 기반 모델이 오히려 더 적합한 경우가 많기 때문에 이러한 과정을 거치지 않았나 생각됩니다. (그리고 일반적으로 대부분의 머신러닝 설루션에서 한 가지 모델만을 바로 사용하지는 않습니다.)

SEA는 모델을 완성한 뒤에 다양한 XAI(설명이 가능한 인공지능) 방법론을 적용하여 각각의 예측한 변수들의 영향력과 그것의 결과 등을 평가하는 방법으로 설루션을 더 진행하였습니다. SEA의 경우 LIME 과 SHAP라는 방법론을 추가로 사용하였는데, 이 중에서 SHAP가 좀 더 이론상으로는 적절한 접근법으로 알려져 있습니다. SHAP를 간단히 설명하자면 어떤 예측에 대한 각 특성(여기서는 Feature가 될 것입니다.)의 영향력(이하 기여도)를 계산하여 예측된 결과를 해석하는 방식입니다. LIME은 SHAP보단 간단하며 빠릅니다. 특정 상황에서는 LIME 방식이 SHAP보다 더 좋을 수도 있습니다.

여기서 하나 더 설명해야 하는 것이 있는데, target과 요소(Feature)의 관계입니다. XAI의 관점에서 이를 쉽게 설명하자면 이와 같습니다. SHAP의 방법으로 예를 들어 Oil temp gap이라는 target의 값에 영향을 가장 많이 주는 요소(Feature)를 찾아보자고 해봅시다. 그 결과는 아래와 같습니다.

이렇게 SHAP 등의 다양한 방식으로 실행하여 영향력 있는 요소(Features)만을 골라서 모델에 적용시켜 모델을 향상시킬 수 있습니다. SEA도 이러한 방식으로 상위 10개의 변수들만을 적용하여 모델의 정확도를 올리는데 사용하였습니다.

이렇게 SEA는 XAI를 통해 머신러닝 모델에 필요한 요소들만을 채택하여 학습의 효율을 올린 모델을 제시하였습니다.

마지막으로 SEA는 최적 설계 requirement를 위에 XAI를 통해 ‘특정 target의 성능에 영향을 주는 요소는 X라는 요소이다.’라는 데이터를 가지고 최적의 requirement를 도출해내었습니다.

## SEA 의 솔루션에 대한 소감

데이터의 전처리 => 모델 향상 => 결과 검증 이 3단계를 다각도로 다루어 이해하기 쉽게 모델을 설명하였습니다. 특히, 중간에 특정 수취 데이터를 로 그 값으로 변환하여 성능을 비약적으로 상승시켰다는 점이 상당히 인상 깊었습니다. 머신러닝에서 이러한 특수한 전처리는 어떤 결과를 가져올지 모르지만, 결과적으로는 모델 성능 향상에 큰 도움을 주었다는 것은 매우 흥미로운 결과입니다.

다만, SEA의 모든 설루션에 대한 내용을 제가 완벽하게 이해한 것은 아닙니다. SHAP 와 LIME 과 같은 XAI 방법론은 제대로 배워본 적은 없었습니다. 머신러닝 교재에서 보통 후반부에 잠깐 설명하고 넘어갔던 부분인데, 이런 설명 가능한 AI(XAI) 방법론은 매우 흥미롭고 재미있었습니다. 아마 대부분의 일반적인 AI 머신러닝 과정을 밟은 학부생의 경우 이러한 자세한 방법론을 실제 프로젝트에 적용시키는 것은 무리가 있을 수도 있으나, XAI 개념 자체는 사실상 매우 단순한 것임으로 일정 기간 공부를 거친다면 단기간에 적용하는 것에는 전혀 무리가 없을 것으로 생각됩니다,(실제로 현재 AI 동아리에서 이에 관련된 공부를 진행 중에 있습니다.) 성능에 어떠한 영향을 미칠지는 알 수 없지만, SEA 팀에서 행하지 않은 것 중에서 시도해볼 만한 추가적인 작업은 요소(Feature)들을 서로 합쳐 복합적인 Feature로 새롭게 데이터를 만들어 내는 것입니다. 이러한 방법은 보통 어떤 Feature가 다른 Feature와 연관이 있다고 판단될 때 사용이 가능한데,(예를 들어 선박에서 발생한 사고의 생존자의 나이와 생존자의 위치 사이에 연관성이 있어 둘을 묶어 Pclass-Ages라는 새로운 Feature로 만들어 사용하여 정확도를 높입니다.) 이러한 합성 Feature는 우선 그 데이터에 대한 경험적인 예측과 해안이 필요합니다. 따라서, 일반적으로 익숙하지 않은 산업 데이터에 적용하기에는 무리가 있어 이러한 시도를 하지 않은 것으로 판단됩니다.

코드를 제가 조금 더 깊게 판단하고 이해하는 것도 필요하겠지만, 일반적으로 머신러닝에서 정확도를 결정짓는 중요한 요소는 데이터 전처리와 분석입니다. 따라서 이번 감상문에서는 그 2가지를 위주로 제가 이해한 내용을 바탕으로 해커톤 설루션의 진행 방향과 결과를 나름의 설명을 덧붙여 작성하였습니다.