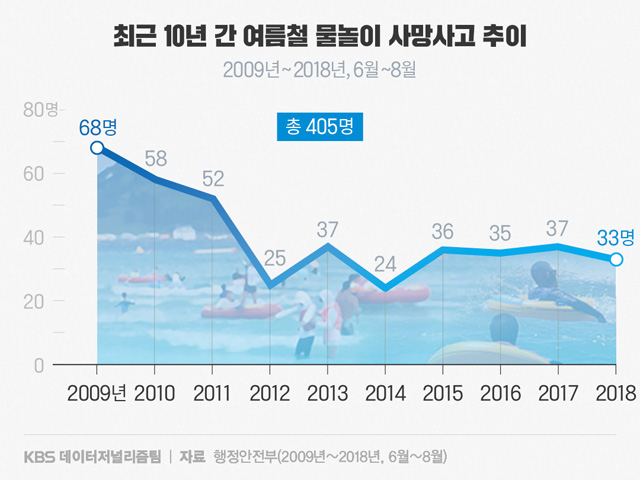
|  |
| --- |
| **Ⅰ. 배경 및 목적** |

수상레저는 스킨스쿠버, 수상스키, 레프팅을 포함하여 물에서 하는 모든 놀이 및 활동으로, 지난해 수상레저 이용객이 519만명에 이를 정도로 많은 국민들이 수상레저 활동에 참여하고 있다. 수상레저 활동을 하는데 있어서, 안전하고 쾌적한 수상 환경은 필수적이다. 따라서 수상레저활동자는 수상레저를 즐기기에 앞서서, 물이 수상레저를 즐기기에 적합한 환경인지에 대한 정보를 필요로 한다. 

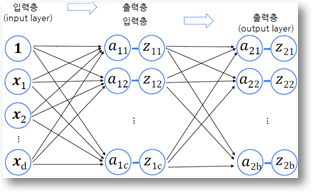
특히 근래의 여러 사건사고 이후 국민의 관심사가 ‘안전’으로 집중되고, 또한 매년 수상레저간에 사망사고가 안전사고로 인해 발생한다(물놀이에서만 수십명) .‘국민안전처’에 의하면 수온, 수심, 유속등이 수상레저 안전사고 예방에 중요하다고 한다, 이러한 안전과 관련이 큰 지표들을 사전에 알려준다면 국민의 수상레저 안전사고를 예방하는데 도움이 될 것이다.

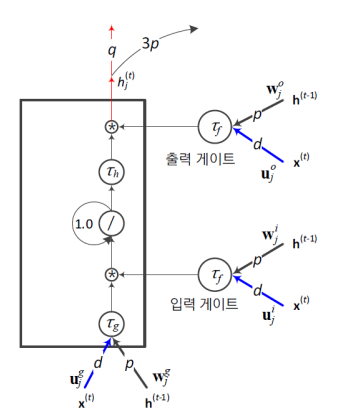
본 과제의 목표를 ‘데이터마이닝과 딥러닝을 통한 수상레저적합 요인 예측 및 지수개발’로 잡고 구체적인 목표는 수상레저와 관련된 여러 지표들의 예측 값을 제공하며, 국민들의 눈높이에 맞춰 한눈에 파악가능한 (그림과 같은) 지수 UI(User Interface)를 제공하는 것이다. 이를 통해 수상레저 계획자들이 미리 수상 환경을 체크할 수 있도록 정보를 제공할 수 있다. 특히, 수상레저별로 사용되는 주요변수들이 다른데, 이러한 부분도 고려하여 사용자가 선택한 각각의 수상레저가 적합한지에 대해 지표로 제공 할 수 있다.

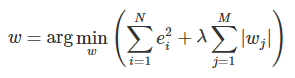
이근영

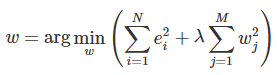
|  |
| --- |
| **Ⅱ. 분석 주요내용** |

수행 계획서에서 언급한 내용 간단히 리뷰 및 실제로 사용한 기법들을 기술한다. ; 기법소개 각 4 기법에 대해 Review 및 사진 도안 첨부

분석의 목적은 기상예보데이터를 이용하여 물관련지표들을 예측한다. 예측된 값을 국민들이 수상레저에 이용할 수 있도록 눈높이에 맞게 가공하여 제공하는 것이다. 따라서 분석의 핵심은 ‘물관련지표예측’과 ‘이용편의를 위한 가공’이 되며, 분석의 경우 크게 4가지 기법을 사용하였다.

첫째, DNN;(입력 계층(input layer)과 출력 계층(output layer) 사이에 복수개의 은닉 계층(hidden layer)들로 이뤄진 인공신경망(Artificial Neural Network, ANN)이다.)을 이용하여 시계열 특성을 반영하지 않고 각 지표들을 학습시켰다. 

둘째, LSTM(RNN 中);(기본원리는 DNN과 같으며, 시계열 데이터(time-series data)와 같이 시간의 흐름에 따라 변화하는 데이터를 학습하기 위한 딥 러닝 모델이다. 또한 vanishing gradient problem문제를 해결하기위해 장단기메모리를 도입한 LSTM기법이 RNN중 대표적이다)을 이용하였는데, 우리가 사용한 데이터는 대부분 시계열 데이터이므로 분석에 적합하다고 판단하였다.

셋째, 회귀모형을 통해 수치형 변수를 예측한다. 이 때, 선형회귀 계수에 대한 제약조건을 추가함으로써 모형이 과도하게 최적화되는 현상을 막는 방법인 '정규화 선형회귀 방법'을 이용한다. 정규화 선형회귀 방법에는 Ridge 회귀모형과 Lasso 회귀모형이 있다. Ridge 회귀모형은 가중치들의 제곱합을 최소화하는 것을 추가적인 제약조건으로 하며, Lasso(Least Absolute Shrinkage and Selection Operator) 회귀모형은 가중치의 절댓값 합을 최소화하는 것을 추가적인 제약조건으로 한다. 특히 Lasso회귀는 덜 중요한 특성의 가중치를 완전히 제거하려고 하는 특징이 있다. 회귀식의 차수를 늘린 다항회귀모형과 과적합을 피하기 위한 Ridge 및 Lasso 모형을 적용하여 모델의 설명력을 비교해볼 것이다.

경험하기 전에 더 잘 맞을 거라고 보장할 수 있는 모델은 없다는 NFL(No Free Lunch) 이론처럼, 어떤 모델이 최선인지 확실히 아는 유일한 방법은 모든 모델을 평가해 보는 것 뿐이다. 하지만 이것이 불가능하기 때문에, 나름대로의 가정을 통해, 적절하다고 생각되는 4가지의 모델을 선정하였다. 사용할 모든 변수에 대하여 대하여 선정된 모델을 적용해보고 각 변수에서 가장 예측효과가 좋은 기법을 채택하여 사용한다.

모델을 학습시키기 전 중요 사안은 예측 할 ‘종속변수’ 선정과 학습을 시킬 ‘독립변수’선정인데, 종속변수의 경우 수상레저와 밀접하게 관련이 있는 변수를 선정하였다. 중요한 것은 학습을 시킬 독립변수의 선택인데, 두 단계로 나누어 독립변수를 선택하였다. 우선, 도메인 지식을 사용하여 예측 지표와 관련 없는 변수를 솎아냈고, 이후 Random Forest;(‘의사결정나무’를 기본으로 하며 이러한 결정트리를 복원추출(배깅) 혹은 비복원추출(페이스팅)하여 여러번 시행하고(앙상블), 여러번 시행한 결과를 바탕으로 투표를 시행하여 가장 좋은 결과값을도출한다(과적합 방지) )를 이용하여 종속변수와 관련이 큰 변수들을 선정하였다.

요약하자면 ‘1. Random Forest를 통해 학습할 변수 선정 2. 선정된 변수를 이용하여 4가지 기법에 적용 3. 4가지 기법 中 예측 지표별로 효과가 가장 큰 기법 선택 4. 선택된 기법을 토대로 예측된 지표들을 국민들의 눈높이에 맞게 가공’의 순서이다.

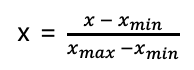
|  |
| --- |
| **Ⅲ. 활용 데이터** |

[데이터 수집] 활용한 데이터는 하천 및 해양 수질데이터와 기상데이터로, 크게 네 가지 데이터를 사용하였다. 하천데이터로는 공공데이터포털에서 제공하는 수질자동측정망 데이터를 사용하였고, 해양데이터로는 국가해양환경정보통합시스템(MEIS)에서 제공하는 해양수질자동측정망 데이터를 사용하였다. 또한 하천과 해양 각각의 데이터에 기상 일자료를 통합하기 위해, 기상자료개방포털에서 제공하는 종관기상관측 데이터와 해양기상부이 데이터를 사용하였다. 위 데이터 중 수질자동측정망 데이터를 제외한 세 개의 데이터는 csv파일로 제공받았으며, 수질자동측정망 데이터의 경우에는 openAPI를 통해 수집하였다. python에서 발급받은 service key를 통해 json형식으로 데이터를 가져온 후, 파싱과정을 거쳐 csv형식으로 변환하였다.

[데이터 전처리] 학습에 적합한 데이터를 만들기 위해, 데이터 전처리를 진행하였다. 먼저 이상치가 있으면 모델이 잘못 훈련될 수 있으므로, 사분범위에서 크게 벗어난 값을 이상치로 설정하여 제거하였다. 결과적으로 각 변수에 대해서 Q1-1.5\*IQR보다 작거나 Q3+1.5\*IQR보다 큰 값이 제거되었다. 또한 종속변수 및 독립변수의 값이 NA면 예측을 할 수 없으므로, 학습 및 예측에 필수적인 데이터에 대해 NA를 포함한 행을 제거하였다.

날짜 변수에서 값들의 형식이 다른 것이 파악되었다. 이 문제를 해결하기 위하여 날짜표현식을 정하여 이러한 형식에 어긋나는 날짜데이터는 제거하였다. 하천 및 해양데이터와 각각의 기상데이터를 통합하기 위해서, ‘날짜’와 ‘지역’ 변수에 대한 값을 통합하여 Composite Key를 생성하였다.

[데이터 통합] 하천과 해양에 대해 수질데이터와 기상데이터를 통합해야 하는데, 수질데이터와 기상데이터의 지역은 각각의 데이터를 측정하는 관측소의 위치가 달라 지역 값을 매칭 할 수 없다. 따라서 각 지역의 위도와 경도를 R의 ggmap을 통해 가져와 변수를 생성한 후, 수질데이터의 지역과 가장 가까운 기상데이터의 지역 위치를 유클리디안 거리로 계산하여 매칭하였다. 또한 아래와 같은 식을 적용하여 MinMax 방법의 정규화를 통해 모든 값을 0부터 1사이의 값으로 변환하여 줌으로써, 모델에 사용하기 위한 최종적인 데이터셋을 완성하였다.



만들어진 데이터셋은 하천에 대한 데이터와 해양에 대한 데이터로 총 두 가지이며, 각각 수질 및 기상에 대한 데이터를 포함한다. 각각의 데이터셋의 변수 목록은 다음과 같다.

|  |
| --- |
| **Ⅳ. 분석 결과** |

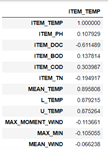
통합 데이터를 사용하여 각각의 변수에 대한 예측모델링과 분석 결과이다. 예측한 지표는 수온과 수질로, 수질과 관련된 변수는 BOD(Biochemical Oxygen Demand, 생화학적 산소요구량), COD(Chemical Oxygen Demand, 화학적 산소요구량), TP(Total Phosphors, 총 인), SS(Suspended Solid, 부유물질), PH(수소이온 농도지수), DOC(용존유기탄소)로 총 6가지이다. 수온을 기준으로 분석결과를 제시하였으며, 다른 변수의 분석 절차는 수온과 동일하다.

모델을 적용하기 전, 통합데이터는 많은 변수를 포함하므로 모델의 성능을 높이기 위해 변수선택(Feature Selection)을 진행하였다. Feature Selection은 다음과 같이 상관분석과 Random Forest를 사용하여 진행하였다.

[Feature Selection① - 상관분석] 상관분석이란 두 변수들 사이의 선형관계를 상관계수로 나타내는 것으로, 계수의 절대값이 클수록 변수 사이에 강한 관계가 있음을 나타낸다. Pearson 상관계수는 두 변수간의 관련성을 구하기 위해 보편적으로 이용되며, 아래와 같은 식에 의해 계산된다.

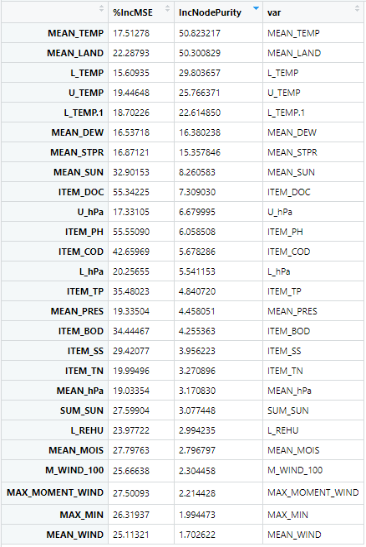


Pearson 상관계수의 경우, 절대값 1은 완전한 선형 관계를 나타내며 0에 가까운 값은 변수 사이에 선형 관계가 없음을 나타낸다. 또한 일반적으로 결정계수의 값이 -1과 1 사이이면, 거의 무시될 수 있는 선형관계라고 해석된다. 따라서 예측할 종속변수(y)를 선정한 후 그에 대한 상관분석을 진행하여, 상관계수의 절댓값이 0.1보다 큰 값만을 추출하여 변수로 사용하였다. 다음 사진은 하천 데이터의 수온 변수를 종속변수로 두고 상관분석을 진행한 결과 추출된 변수들의 일부를 나타낸 것이다.

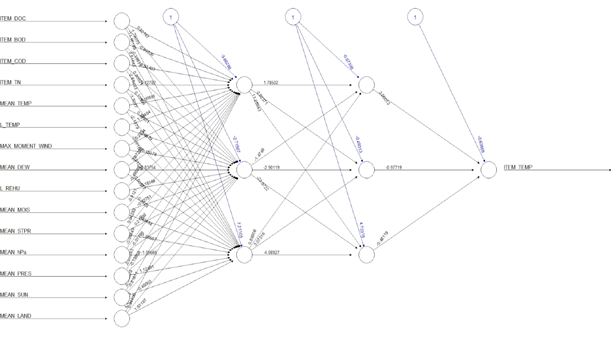


[Feature Selection② - Random Forest]

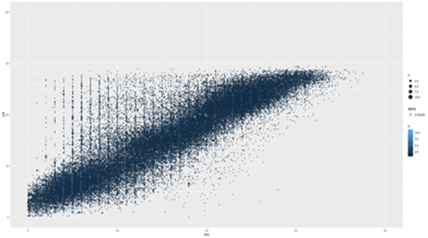
상관분석만으로 모델에 유의미한 변수를 선정하는 것에는 한계가 있다고 생각하여, 추가적으로 Random Forest 기법을 사용하였다. 랜덤포레스트 기법은 분류 알고리즘의 하나로, 여러 개의 의사결정나무를 임의적으로 학습하는 방식의 앙상블(여러 개의 모형을 학습한 후 예측 결과들을 종합하여 정확도를 높이는 방법)을 사용한 모델이다. 이는 보통 예측의 용도로 사용되지만, R의 ‘randomForest’ library를 이용하여 Purity를 기준으로 내림차순하여 사용할 변수를 선정하였다. MDA, MDG대신 갑자기 Purity쓴 이유가 궁금해유 incnodepurity면 mdg인가유? 이부분 조금 설명이나 근거 들어갔으면 좋겠어유. 보통 중요변수 찾을 때 mda 쓰는것같던데,, ITEM\_TEMP(수온)을 종속변수로, 그 외 변수들을 독립변수로 설정한 뒤, 랜덤포레스트를 실행하여 선정된 변수의 일부는 다음과 같다. 상위의 변수들을 살펴보면, ‘평균 기온’, ‘평균 지온’, ‘최저 기온’ 등이 있다.



[Modeling① - DNN(Deep Neural Network)] 통합 데이터를 DNN모델에 적용하기 전, 예측 성능을 파악하기 위하여 train set과 test set을 각각 7:3으로 랜덤 추출하였다. 수온 변수(ITEM\_TEMP)를 예측변수로 설정하여 DNN모델에 학습시킨 Neuralnet 모형은 다음과 같다.



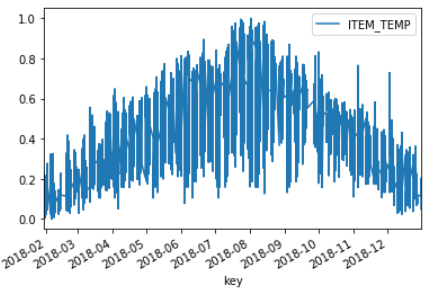
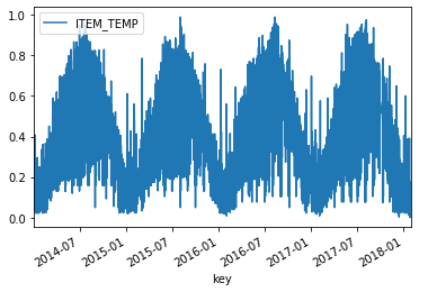
좌측의 변수들은 독립변수(input)이며, 우측의 변수는 예측하고자 하는 종속변수(output)이다. 가운데의 각 Layer에 포함되는 Node들은 서로 연결되어 있으며, 연결선에 기록된 w(가중치)값을 통해 예측하고자 하는 종속변수의 값을 예측하게 된다. DNN 모델링 시에 설정한 하이퍼파라미터는 다음과 같다. 하이퍼파라미터에 의해 모델의 성능이 개선 및 저하될 수 있다. Activation Function(활성화 함수)는 sigmoid를 사용하였고, Error function(오차 함수)로 회귀예측에 주로 사용하는 SSE(Sum of Squared of Errors, 잔차변동)를 설정하였다. 또한 최대 계산량을 설정하는 stepmax는 입력변수 및 record가 많아 기본값보다 더 여유를 주어서 1e7로 설정하였으며, hidden layer와 node의 수는 2~3개에서 그리드 탐색을 한 결과 (3,3)에서 결과가 가장 좋았다. 하이퍼파라미터에 대한 설명과 근거가 있었으면 좋겠어유. 예를들어 활성화함수에는 sigmoid, relu 등등의 종류가 있는데 ~~한 경우에는 sigmoid를 주로 쓰므로 sigmoid를 선택했다. 이런식으로. stepmax같은 것도 의미하는게 무엇인지 들어가면 좋을것같구먼유. 혹시 하이퍼파라미터를 그리드탐색처럼 이러한 조합으로 사용했을 때, 성능이 이정도 나왔고 이런거 기록한 표 있으면 좋을거같은데 힘들겠쥬..? 위의 모델의 경우 3개의 hidden layer와 각각 3개의 node를 사용하였다. 다음은 모델링 후, 수온에 대한 예측의 결과를 그래프로 나타낸 것이다. R square값은 0.788, RMSE값은 0.061의 결과를 얻었다.

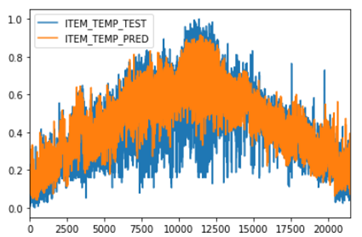


[Modeling② - LSTM(Long Short-Term Memory)]

두번째로 사용한 모델은 LSTM(Long Short-Term Memory)이다. LSTM을 간단히 설명하자면 가장 발전된 형태의 RNN구조이다. RNN을 이용하면 시계열 데이터를 효과적으로 모델링 할 수 있는데, 기존 RNN은 시간에 따라 학습이 계속 진행되면서 앞에서 받은 정보가 학습에 미치는 영향이 점점 감소하다가 사라져버리는 Vanishing Gradient Problem을 가진다. 하지만 LSTM은 RNN의 일종으로 시계열 데이터를 효과적으로 모델링 할 수 있으면서도, RNN의 문제점을 해결하여 은닉층의 메모리 셀에 입력 게이터, 망각 게이트, 출력 게이트를 추가하여 불필요한 기억을 지우고, 기억해야 할 것들을 정하는 방식이다. 따라서 시계열 데이터이고, 5개년의 긴 데이터를 사용할 본 과제에 적합하다고 판단하여 사용하게 되었다.

통합 데이터를 LSTM 모델에 적용하기 전, 다른 모델과는 다르게 LSTM의 특징을 살리기 위해 날짜 변수를 사용하였다. 그 후 학습을 시키기 위해 트레이닝셋과 테스트 데이터 셋을 분리하였는데, 각각 8:2의 비율로 분리하였다. 보통 7:3의 비율을 많이 사용하는데, 8:2의 비율을 사용한 이유는 사용할 데이터의 양이 2014년부터 2018년까지 총 5개년의 데이터이기 때문이다. 데이터에는 계절별 특성등이 드러나 있으므로 대략 2018년 데이터를 테스트 셋으로 놓는 식으로 진행하였다. 수온에 대한 트레이닝셋과 테스트 셋을 그림으로 그려보면 다음 사진과 같다.(좌 : 트레이닝 셋, 우 : z테스트 셋) 수온이므로 여름에는 높고 겨울에는 낮은 경향을 잘 드러내고 있음을 볼 수 있다. 이 둘을 나눌 때 주의할 점은 시계열 데이터이므로 섞으면 안된다는 점과, 시간순으로 정렬을 잘 해주어야 한다는 점이다.



모델을 돌릴 때에는 layer는 20개로 설정하고, loss의 기준을 ‘MSE’로, optimizer는 ‘adam’을 사용했을 때 가장 성능이 좋았다. 또한 epoch를 100으로 크게 설정하고 early stop을 적용하여 더이상 모델의 성능이 좋아지지 않고 머무르면 자동으로 멈추게 하였다. 여기서 한 번의 epoch는 인공 신경망에서 전체 데이터 셋에 대해 forward pass/backward pass 과정을 거친 것을 말한다. 즉, 전체 데이터 셋에 대해 한 번 학습을 완료한 상태이다. 너무 많은 epoch는 overfitting을 일으키고, 너무 적은 epoch는 underfitting을 일으키기 쉽다. 이런 상황에서는 epoch를 많이 돌린 후 early stopping을 이용해서 특정 시점에서 멈추는 기술이 필요하다. 이 특정 시점을 정하는 것의 기준은 예측값의 성능이 더이상 증가하지 않을 때 학습을 중지시킨다. 여기서는 loss의 기준을 ‘MSE’로 설정하였으므로 MSE가 더이상 줄어들 지 않으면 학습을 중단시키고 가장 좋은 모델의 값으로 돌아가게 된다. 

다음은 모델의 결과를 시각화 한 사진이다. 테스트 셋(ITEM\_TEMP\_TEST : 파랑)과 트레이닝 셋을 이용하여 예측한 예측 값(ITEM\_TEMP\_PRED : 주황)을 겹쳐 그렸는데, 증가하고 감소하는 추이 등이 잘 맞는 것을 볼 수 있다. lstm을 평가할 때 가장 많이 사용하는 RMSE를 이용하여 평가를 진행하였고, 그 외에도 자주 사용되는 R\_square 값도 구해 보았다. 그 결과는 아래 표에 나타내었다. 대표적으로 평균 수온에 대해 말해보자면, RMSE 값은 0.086, R\_square 값은 0.827으로 좋은 성능을 보였다.

위에랑 비슷하게 쓰면 될거 같아유. -쓰는중(채영)

1) Train set, test set 나눈 것

2) Lstm에 대한 간단한 설명. 이러한 특성이 있어서 ~~ 해주었다 같은거 (다른 모델에서는 사용하지 않은 변수이지만 이 모델의 경우에는 시간 개념이 반영되므로 key값으로 시간을 사용하였다. 나 trainset, testset 나눌 때 시간순으로 정렬해서 시간적 특성을 반영했다 라던가) 가 들어갔으면 좋겠구먼유

3) 하이퍼파라미터에 대한 설명과 설정한 것. 혹시 있다면 조정해본 결과

4) Item temp에 대한 예측결과. 거 실제값이랑 예측값이랑 같이 그래프그린거 예쁘던데유!

5) 외에 언급하고 싶은거, 모델 관련 캡쳐한거 있으면 넣어유!

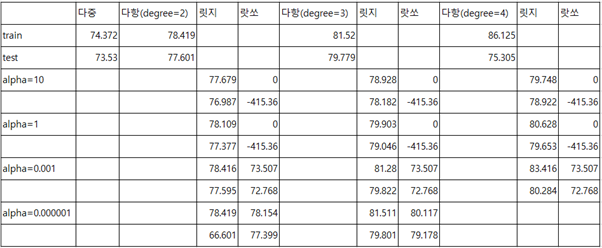
[Modeling③ - Regression(Polynomial, Ridge)]

모델링에 앞서, 모델의 성능을 평가하기 위하여 전체 통합데이터를 train set과 test set을 7:3 비율로 나누었다. 이와 같이 모델성능 평가 방법 중 하나로 전체 dataset을 train set과 test set을 나누는 방법을 Hold-out이라고 하며, 일반적으로 7:3 비율이 가장 자주 쓰인다. 회귀분석에 경우에는 LSTM과 달리 시간적 개념이 포함되지 않으므로, 수온이나 수질과 같은 예측변수가 시간의 영향을 받지 않는다는 가정 하에, 랜덤으로 추출하였다. 즉, 동일한 일자의 독립변수(x)만이 종속변수(y)에 영향을 미친다고 가정한 것이다.

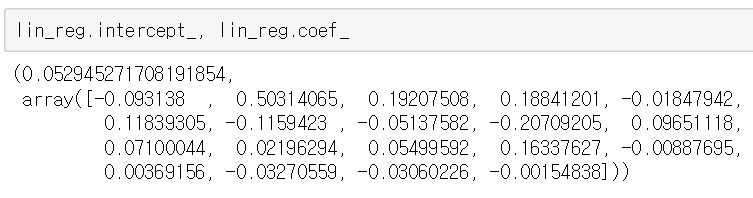
회귀분석은 관찰된 연속형 변수에 대해 독립변수(x)와 종속변수(y) 사이의 선형식을 구하고, 그 식을 이용하여 독립변수가 주어졌을 때 종속변수를 예측하는 분석방법이다. 잔차의 제곱합을 최소화시키는 최소제곱법을 통해 회귀식에서 x의 계수와 상수항을 변화시켜가며, 최적의 회귀식을 찾을 수 있다.

회귀식에는 독립변수의 개수 및 차수에 따라 Simple Regression(단순회귀분석), Multiple Regression(다중회귀분석), Polynomial Regression(다항회귀분석)으로 나뉜다. 단순회귀분석은 독립변수(x)의 개수가 하나인 경우이며, 다중회귀분석은 독립변수(x)의 개수가 두 개 이상인 경우에 사용한다. 다항회귀분석은 독립변수(x)와 종속변수(y)가 선형적인 관계가 아닌 곡선 형태로 되어있는 경우를 말하며, x의 차수가 2차 이상인 항이 포함되어 있는 경우이다. 이러한 다항회귀는 매우 강력해서, 모델이 데이터에 과적합하게 되기도 한다. 이를 방지하기 위해 일종의 정규화 기법을 사용할 수 있으며, Ridge Regression, Lasso Regression 등이 있다.

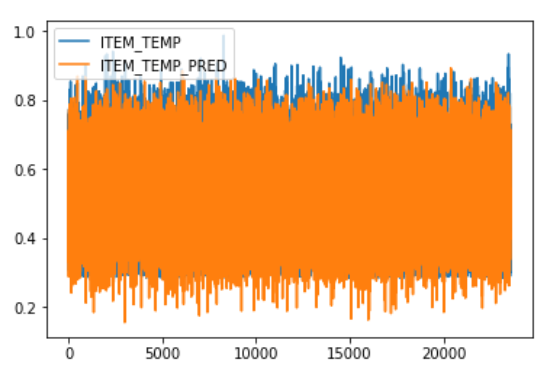
우리의 경우, 두 개 이상의 독립변수(x)로 종속변수(y)를 예측하므로 기본적으로 다중회귀분석을 이용하였고, 독립변수의 차수를 증가시켜 다항회귀와 Ridge회귀, Lasso회귀로도 모델링을 진행하였다. 다항회귀분석을 위해, python의 PolynomialFeatures함수를 이용하여 주어진 차수까지 특성간의 모든 교차항을 추가하였다. 예를 들어 두 개의 특성 a,b가 있을 때, degree=3으로 PolynomialFeatures 함수를 적용하면 , , , 뿐만 아니라 , 등의 항도 특성으로 추가된다. 또한 Ridge 회귀와 Lasso 회귀에서 하이퍼파라미터인 alpha를 통해 규제의 정도를 조정할 수 있는데, 다항회귀의 degree와 Ridge 및 Lasso 회귀의 alpha값을 변화시키며 R-squared값을 살펴본 결과는 다음과 같다. R-squared값은 모델의 설명력을 나타내며, 높을수록 모델의 설명력이 높다고 본다.



앞서 수온에 대해 RandomForest를 통해 도출된 중요변수를 독립변수로 하고, 수온을 종속변수로 설정하여 다양한 모델을 파라미터를 변경시켰을 때의 R-squared 값을 표로 나타낸 것이다. 다중회귀에서 차수를 늘려 보다 복잡한 모델인 다항회귀를 만들었을 때, 성능이 더 좋아진 것을 알 수 있다. 또한 차수가 더 클수록 train set과 test set의 차이가 커지는 것을 통해, 모델이 복잡해질수록 train set에 Overfitting(과적합) 되었음을 알 수 있다. 각각의 차수를 증가시킨 train set과 test set에 대해 Ridge회귀와 Lasso회귀를 적용하여 규제의 정도인 alpha값을10, 1, 0.001, 0.000001로 변경시켜보았다. Lasso회귀의 경우 덜 중요한 특성의 가중치를 완전히 제거하려는 특징 때문에 대체로 Ridge보다 더 작은 alpha값을 사용한다. 따라서 Lasso의 경우 alpha가 10일 때와 1일 때 정상적인 값이 나오지 않은 것을 확인할 수 있다. 또한 alpha의 값이 작아지면 규제가 작아 원래의 회귀식과 유사해지므로 각각 다항회귀식의 R-squared값과 비슷해지는 것을 알 수 있다. 차수가 2, 3일때는 과적합이 크게 나타나지 않아 Ridge회귀와 Lasso회귀를 적용했을 때 크게 개선이 없다. 하지만 차수가 4인 경우, 적절한 alpha값으로 규제를 가하면 train set과 test set의 차이를 줄일 수 있다. 즉 과적합을 줄여 test set에 대한 설명력을 높일 수 있는 것이다. Degree=4일 때의 Ridge와 Lasso는 실행속도의 한계로 생략하였다. 위와 같이 하이퍼파라미터 조정을 거쳐, Ridge회귀와 Lasso회귀의 alpha값은 0.005로 설정하여 모델링하였다. 또한 차수는 3으로 설정하였으며, 다항회귀와 Lasso보다 더 좋은 성능을 보였던 Ridge회귀만을 사용하였다. 수온 변수를 예측하였을 때, 다항회귀의 R-squared값은 80.2, Ridge회귀의 R-squared값은 80.3이었다. 다음과 같이 회귀식의 상수항과 가중치를 확인할 수 있다.



아래는 수온에 대한 예측값과 실제값을 그래프로 나타낸 것이다. 데이터가 많아서 잘 보이지 않지만, 두 값이 비슷한 추이를 보이는 것을 알 수 있다. LSTM은 시간이 반영되어 전체적인 흐름이 보이는 반면, 이 경우 데이터셋을 랜덤으로 추출하여 특정한 흐름이 보이지 않는다. 이 모델의 RMSE값은 0.07로 나타났다.



[Modeling④ - 각 기법의 성능 비교표]

7가지 변수에 대한 각 모델의 예측 성능을 표로 나타낸 것이다. 첫 번째 표는 각 변수에 대한 상관분석을 통해 상관성이 높은 변수를 독립변수로 사용한 모델의 R-squared 값과 RMSE 값이다.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| R^2/RMSE | TEMP | BOD | COD | TP | SS | PH | DOC |
| DNN | 0.762/0.066 | 0.401/0.115 | / | 0.395/0.152 | / | / | 0.331/0.110 |
| LSTM | 0.827/0.086 | 0.308/0.122 | 0.410/0.120 | 0.421/0.150 | 0.308/0.122 | 0.188/0.170 | 0.356/0.233 |
| 다항회귀 | 0.780/0.064 | 0.415/0.115 | 0.600/0.103 | 0.472/0.145 | 0.478/0.157 | 0.244/0.170 | 0.377/0.111 |
| RIDGE | 0.782/0.064 | 0.416/0.115 | 0.600/0.103 | 0.473/0.145 | 0.478/0.157 | 0.244/0.169 | 0.384/0.110 |

두 번째 표는 각 변수에 대한 RandomForest를 통해 중요도가 높은 변수를 독립변수로 사용한 모델의 R-squared 값과 RMSE 값이다.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| R^2/RMSE | TEMP | BOD | COD | TP | SS | PH | DOC |
| DNN | 0.788/0.061 | 0.388/0.120 | / | 0.367/0.155 | / | / | 0.342/0.114 |
| LSTM | 0.789/0.096 | 0.344/0.119 | 0.509/0.110 | 0.346/0.160 | 0.425/0.162 | 0.200/0.169 | 0.378/0.211 |
| 다항회귀 | 0.802/0.061 | 0.412/0.115 | 0.526/0.113 | 0.502/0.140 | 0.477/0.156 | 0.305/0.162 | 0.384/0.110 |
| RIDGE | 0.803/0.061 | 0.438/0.112 | 0.527/0.113 | 0.505/0.140 | 0.482/0.155 | 0.311/0.161 | 0.402/0.109 |

|  |
| --- |
| **Ⅴ. 실용화 방안** |

[MyWater를 통한 수상레저지수 제공]

K-Water에서는 현재 국민들이 궁금해 할 만한 물에 대한 정보를 '물 정보 포털(MyWater)' 애플리케이션을 통해 제공하고 있다. MyWater 애플리케이션에서는 현재 물에 대한 실시간 통계 정보를 알 수 있다. 이를 통해서 원하는 위치와 날짜를 선택하여 댐/보 별로 수질 정보 등을 제공받을 수 있다. 우리나라 지난해 수상레저 이용객이 519만명에 이를 정도로 국민들이 큰 관심을 받고 있는 상황에서, 국민들이 쉽게 어플리케이션을 통해 수상레저 관련 정보를 알 수 있게 도울 수 있다. 뿐만 아니라, 실제 상용화 되어 있는 ‘물 정보 포털(MyWater)’ 애플리케이션에 기능을 추가하는 것을 목표로 하기 때문에 새로운 어플리케이션 또는 활용방안을 제시하는 것 보다 활용, 홍보적, 재정적 차원에서 큰 이점을 가질 것으로 보인다.

‘수상 레저 예측 지수’는 사용자 친화적으로 설계되었다. 수상 레저 예측 지수를 구성하고 있는 4개의 항목에는 수온, 유속, 수질, 수심 등이 있는데 이를 실제 예측 수치 값으로만 나타낸다면, 해당 수치가 수상레저를 하는데 적절한 것인지, 위험한 것인지를 판단하기 어려울 수 있다. 따라서 예측 수치와 함께 주의, 심각, 정상 등의 표시와 함께 직관적으로 빨리 파악할 수 있도록 각각을 노랑, 빨강, 초록 색으로 표시하여 국민들의 이해를 도울 수 있다.

[수상레저 계획 도우미 및 안전사고 예방]

먼저 수상레저 계획 시 도움을 받을 수 있다. 보통 여행을 계획할 때 날씨를 고려하여 세부 일정을 정하기 마련이다. 마찬가지로 해당 서비스를 이용하면 국민들은 수상레저를 계획하기 전, 미리 수상레저 지수를 확인할 수 있게 되어 레저를 즐길 적절한 장소, 날짜 등을 정하고 즐기는데 도움을 받을 수 있을 것이다.

또한 14년도 이전에는 수상레저 사고가 3~4건이었던 것에 반해, 17년도에는 54건에 이르는 등 수상레저의 안전에 대한 문제가 빈번하게 발생하고 있다. 이처럼 수상레저 사고의 빈도가 증가하고 있는 가운데, 국민들이 수상레저 지수 예보를 통해 수상환경 및 안정성에 대한 정보를 미리 얻는다면 수상레저 사고를 방지할 수 있을 것으로 예상된다.

[기존의 해양예보시스템의 개선 및 확장]

현재 기상청 날씨누리 홈페이지에 따르면 12시간 전부터 12시간 후 까지의 지점별 풍향, 풍속, 기압, 습도, 기온, 수온, 파고 등의 정보를 확인할 수 있다. 하지만 12시간 후밖에 예측값이 없기 때문에 미리 여행 계획을 세우는 국민들이 사용하기에는 불편한 점이 존재한다. 본 과제는 하천을 대상으로 분석을 진행하였지만, 해양데이터에 동일 방법을 적용하여 짧은 미래까지만 예측 가능한 기존의 해양예보시스템을 개선할 수 있다. 또한 하천과 동일하게 종합레저지수로까지 확장할 수 있을 것이다.

[안전지킴이 수 예측]

한정된 자원과 인력으로 해양 및 하천에 안전지킴이를 얼마나 배치하는 것이 효율적인지 결정하는 것은 어려운 일이다. 하지만 본 서비스를 통해서라면 사람들이 많이 모일 것 같은 날, 그리고 날씨 등으로 통해 위험한 날 등 목적에 맞게 안전지킴이 수를 유동적으로 조절할 수 있다. 이와 같은 일 또한 미리 일정을 정할 수 있으므로 유용하게 사용될 것으로 예상된다.

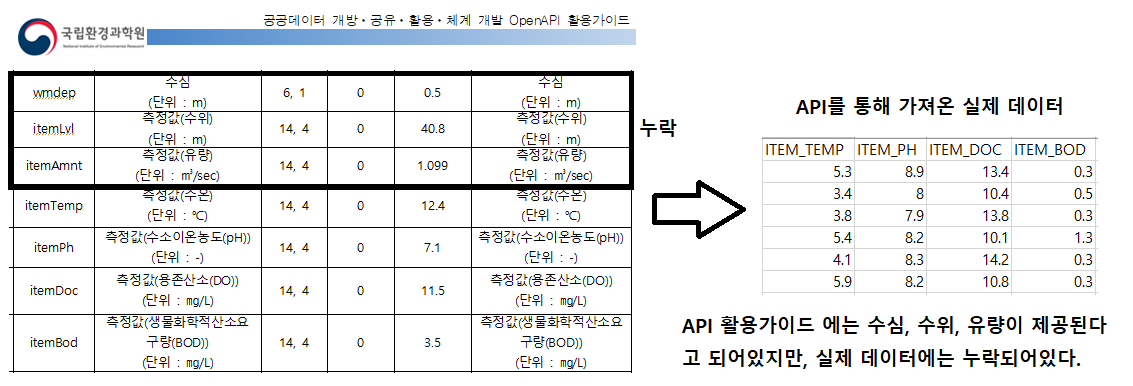
[새로운 레저 관광 지역 개발]

정부적 차원에서는 이러한 지수를 예보에 사용하는 것뿐만 아니라, 새로운 수상레저 관광지 개발에 활용할 수 있다. 높은 점수의 수상레저 지수를 가지는 지역 중 수상레저가 활성화되지 않은 지역에 수상레저를 활성화시키는 것이다. 이 경우 해당 지역경제를 활성화할 수 있는 동시에 국민들은 보다 쾌적한 수상환경에서의 수상레저를 즐길 수 있을 것이다.

|  |
| --- |
| **Ⅵ. 결론 및 시사점** |

본 과제는 제공되는 과거의 날씨, 하천의 정보를 통해서 현재 제공되고 있지 않은 미래의 하천의 수온과 같은 값을 직접 측정하지 않고 예측할 수 있다는 점에서 큰 의의를 가진다. 실제로, 미래의 일은 직접 측정하지 않으면 알 수 없는 경우가 대부분이다. 하지만 짧은 미래의 일을 예측해주는 곳이 있다. 바로 기상청이다. 과거 데이터를 바탕으로 날씨와 수온, 수심 등의 관련성에 대해 모델을 학습시킨 후, 기상청의 데이터를 토대로 날씨 데이터를 이용한다면 미래의 하천의 수온, 수심, 유속 등을 예측할 수 있겠다는 생각에서 본 과제는 시작되었다. DNN, LSTM, 릿지 라쏘 회귀, 단순 선형회귀 등의 기법을 사용하여 분석을 진행하였고 변수별로 가장 RMSE 및 R\_square 값이 가장 우수한 모델을 이용하는 방향을 택했다. 또한 종합 레저 지수를 생성하기 위해 실제 레저를 즐기는 사람들의 목소리를 들어보기로 하였다. 따라서 레저별 네이버 카페를 텍스트마이닝 하여 본 과제에서 추출한 변수들의 가중치를 정하여 최종 레저 지수를 완성할 수 있었다.

물론 본 과제에도 한계점이 존재한다. 아래 사진과 같이 사용한 API의 활용가이드에는 분명 수위, 유량 등이 표시되어 있지만, 실제 데이터에는 누락되어 있음을 볼 수 있다. 따라서 분석 시에는 수온과 수질을 알 수 있는 몇가지의 요소(BOD, COD, PH 등)들만 사용할 수밖에 없었다. 수위, 유량 등과 같은 데이터가 주어지기만 한다면, 동일한 방식으로 간단한 전처리만 거친 후 바로 모델 학습이 가능하도록 준비해 두었기 때문에, 추가 데이터를 이용할 수 있다면 더 유용한 서비스로 거듭날 수 있을 것으로 기대된다.



* MyWater를 통한 수상레저지수 제공
* 기존의 해양예보시스템의 개선 및 확장
* 해양 및 하천 안전지킴이 수를 예측에 따라서 유동적으로 조정하여 배치할 수 있다.
* 새로운 레저 관광 지역 개발