2019 K-WATER 공모전| Hooray팀

데이터마이닝과 딥러닝을 통한 수상레저적합 요인 예측 및 지수 개발

일반 국민 전형

이근영, 박채영, 소은비

목차 |

배경 및 목적 | 선정배경

해양경찰청
KOREA COAST GUARD

해양경찰청

[경인방송 = 최상철 기자]
해양경찰청(청장 조현배)이 국민의 안전한 수상레저 활동을 위해 대비책 마련에 나섰습니다.

오늘(24일) 해양경찰청에 따르면 올해 강과 바다에서 수상레저를 즐긴 인구는 519만여 명으로, 지난해 431만명보다 약 20% 증가했습니다.

최근 수상레저를 즐기는 인구의 증가가 두드러짐과 동시에,

그에 따른 안전사고에 대한 문제도 꾸준히 제기되고 있다.

특히 여름철 물놀이에서만 매년 수십명의 사망자가 발생한다.

배경 및 목적 | 선정배경

[뚝뚝! 응급의료]물놀이 안전사고 막으려면...깊이·수온 먼저 확인, 준비운동은 필수

박효은 기자 anytoc@kyunghyang.com

일반적으로 수영하기에 알맞은 수온은 25~26℃ 정도이다. 물에 들어갈 때는 다음 사항을 꼭 지켜야 합니다.

야외 물놀이를 계획할 때에는 물이 깨끗하고, 자연 조건이 안전한 지역을 선택해야 합니다.

물놀이 이전에 확인하여 둘 것들

물의 깊이와 온도, 물 흐름의 빠르기를 먼저 확인한다.

물놀이를 하면 안되는 경우

1. 물이 오염되어 있나?

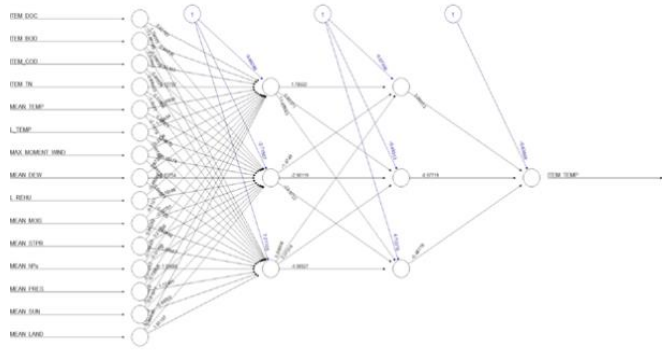
수상레저 안전에 중요한 것은 '수온','수심','유속','수질'이다.[국민안전처, 중앙의료원 출처]

BUT, 이러한 변수들은 예보가 제공되지 않음.

-> 위 변수들을 예보하고 각 수상레저를 즐기기에 적합한지 지표로 만들어 제공

배경 및 목적 | 목적

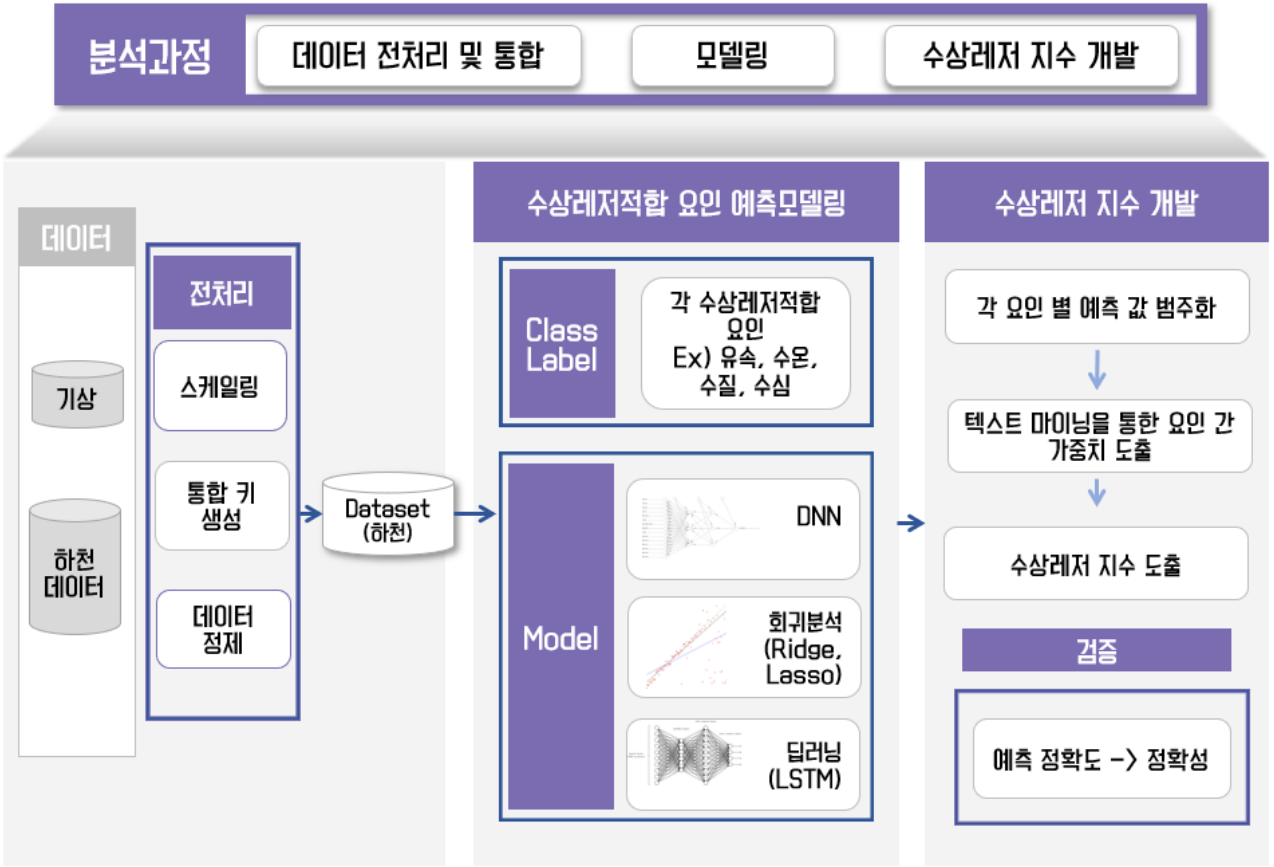
목적은 크게 두가지!
첫째, 수상레저와 밀접한 변수들을 여러 예측 기법을 통해 예측



둘째, 예측한 변수를 이용하여 국민의 안전과 즐거운
수상레저를 즐길 수 있도록 가공하여 UI형태로 제공

UI에서는 수상레저 종류, 장소, 시간 등을 선택 할 수 있고,
이에 대한 맞춤형 정보를 우측의 APP과 같이 국민의 눈높이
에 맞게 제공

분석 주요내용 | 방법론

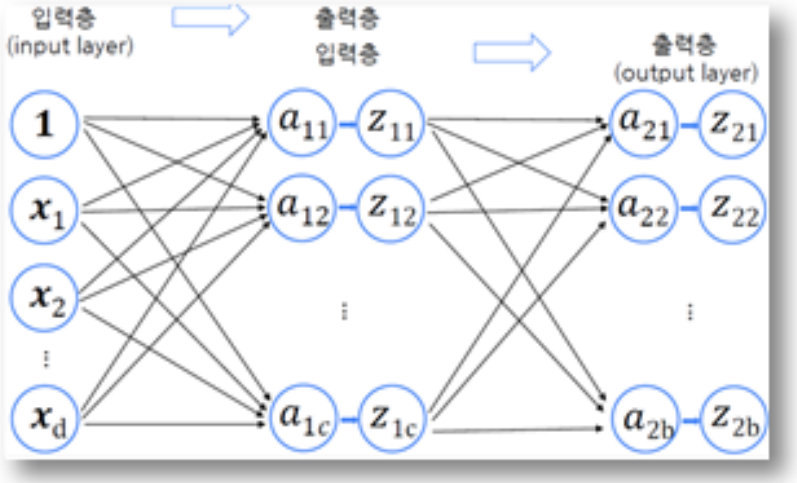


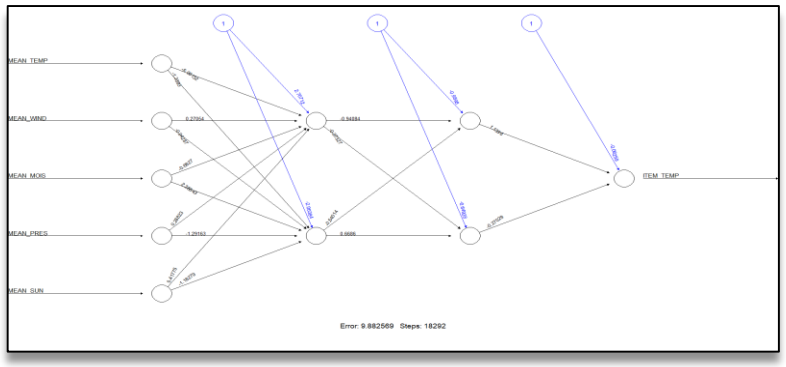
데이터 전 처리후,
랜덤포레스트와 상관분석을 통해
딥러닝과 회귀분석에 사용할 변수 선정

선정된 변수를 이용하여,
여러 기법 적용하여
종속변수예측

예측변수와 기법 별로 도출된
R^2/RMSE 테이블을 작성하여,
변수마다 가장 성능이 좋은 기법 선정

분석 주요내용 | 기법 소개 [DNN]





DNN이란?

- 딥러닝의 기본적인 형태
- 입력층과 출력층 사이에 '은닉계층'과 '노드'로 이뤄진 인공신경망
- Back- Propagation을 통해 학습됨
- 하이퍼파라미터 : 활성함수, 노드 수, 은닉계층 수, 학습률 등

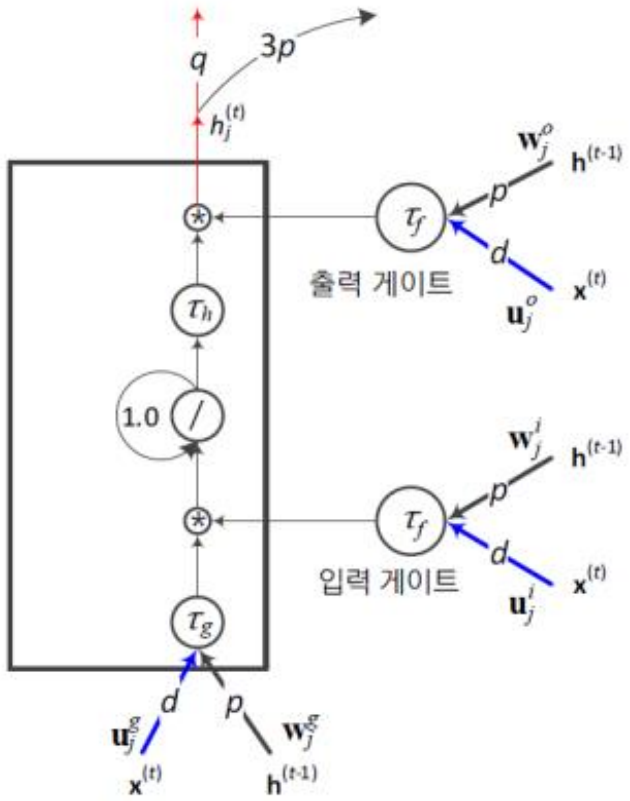
사용 언어 및 라이브러리

R언어 - 'neuralnet' library

하이퍼파라미터

'layer', 'node', 'stepmax', 'threshold', 'learningrate', 'algorithm'을 조절하여 종속변수를 예측함

분석 주요내용 | 분석 기법[LSTM]



LSTM이란?

- 시계열 데이터와 같이 시간의 흐름에 따라 변화하는 데이터를 학습하기 위한 딥 러닝 모델
- vanishing gradient problem문제를 해결하기 위해 장단기메모리를 도입

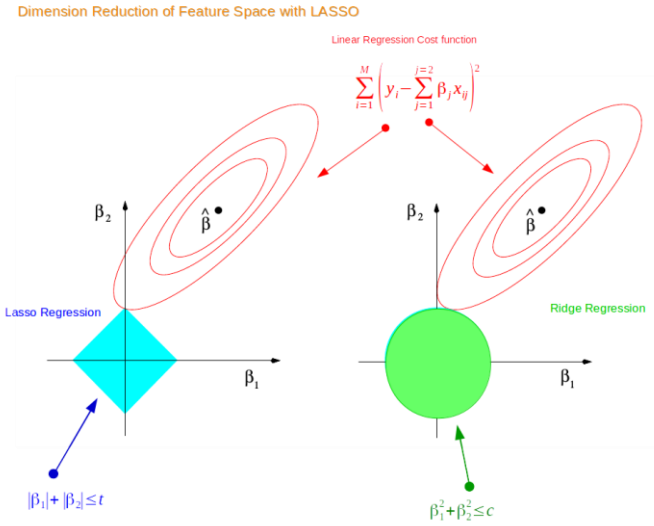
사용 언어 및 라이브러리

Python - keras

하이퍼파라미터

'layer', 'loss', 'optimizer', 'epoch' 를 조절하여 종속변수를 예측함

분석 주요내용 | 분석 기법[LASSO, RIDGE 회귀]



$$w = \arg\min_{w} \left(\sum_{i=1}^{N} e_i^2 + \lambda \sum_{j=1}^{M} |w_j| \right)$$

$$w = \arg\min_{w} \left(\sum_{i=1}^{N} e_i^2 + \lambda \sum_{j=1}^{M} w_j^2 \right)$$

Ridge, Lasso 란?

- 선형회귀에서 제약조건을 추가하여 과적합을 방지
- [Ridge 회귀모형] : 가중치들의 제곱합을 최소화하는 것을 추가적인 제약조건으로 함
- [Lasso 회귀모형] : 가중치의 절댓값 합을 최소화하는 것을 추가적인 제약조건으로 함

사용 언어 및 라이브러리

Python - sklearn

하이퍼파라미터

'degree', 'alpha'값을 조정하여 종속변수를 예측함

활용 데이터 | 데이터 소개 및 수집

데이터 소개

데이터 :
1. 하천 데이터
-> [공공 데이터 포털 제공] 수질자동측정망 데이터

2. 기상 데이터
-> [기상자료개방포털 제공] 종관기상관측 데이터

데이터 수집

1. 하천 데이터
-> openAPI를 통해 수집
-> Python을 통해 json 형식으로 데이터를 가져온 후 파싱하여 csv로 변환하여 사용

2. 기상 데이터
-> csv 파일로 제공받음

활용 데이터 | 데이터 전처리 및 통합

데이터 전처리

1. 이상치 제거
-> 사분범위에서 크게 벗어난 값을 이상치로 설정하여 제거

2. NA를 포함한 행 제거

3. 날짜 변수 전처리
-> 정규표현식을 이용하여 어긋나는 데이터 제거

4. Composite Key 생성
-> 하천 데이터와 기상데이터를 통합하기 위해 생성
-> 날짜+지역

데이터 통합

1. 하천데이터와 기상 데이터의 통합
-> 수질 데이터와 기상데이터의 데이터 측정소의 위치가 달라 지역 값 매칭이 어려움

-> R의 ggmap을 통해 각 지역의 위도, 경도 변수 생성

-> 수질 데이터의 지역과 가장 가까운 기상데이터의 지역 위치를 유클리디안 거리로 계산하여 매칭

2. 변수 정규화
-> MinMax 방법을 통해 모든 값을 0과 1사이의 값으로 변환

분석 결과 | 상관분석

	ITEM_TEMP
ITEM_TEMP	1.000000
ITEM_PH	0.107857
ITEM_DOC	-0.611506
ITEM_BOD	0.137920
ITEM_COD	0.303958
ITEM_TN	-0.194949
MEAN_TEMP	0.895827
L_TEMP	0.879232
U_TEMP	0.875277
MAX_MOMENT_WIND	-0.113684
MAX_MIN	-0.105122
MEAN_WIND	-0.066327

[ITEM_TEMP(수온)를 종속변수로 상관분석을 실행한 결과]

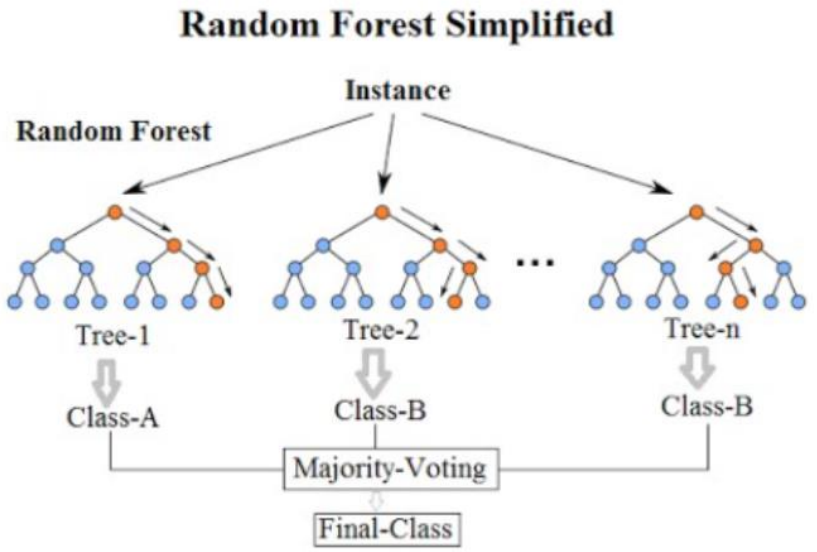
상관계수 : 두 변수들 사이의 선형관계를 나타냄.
Pearson 상관계수가 보편적으로 이용되며,
아래와 같은 식에 의해 계산됨.

$$r_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

분석 결과 | 랜덤포레스트

	%IncMSE	IncNodePurity	var
MEAN_TEMP	17.51278	50.823217	MEAN_TEMP
MEAN_LAND	22.28793	50.300829	MEAN_LAND
L_TEMP	15.60935	29.803657	L_TEMP
U_TEMP	19.44648	25.766371	U_TEMP
L_TEMP.1	18.70226	22.614850	L_TEMP.1
MEAN_DEW	16.53718	16.380238	MEAN_DEW
MEAN_STPR	16.87121	15.357646	MEAN_STPR
MEAN_SUN	32.90153	8.260583	MEAN_SUN
ITEM_DOC	55.34225	7.309030	ITEM_DOC
U_hPa	17.33105	6.679995	U_hPa
ITEM_PH	55.55090	6.058508	ITEM_PH
ITEM_COD	42.65969	5.678286	ITEM_COD
L_hPa	20.25655	5.541153	L_hPa
ITEM_TP	35.48023	4.840720	ITEM_TP
MEAN_PRES	19.33504	4.458051	MEAN_PRES
ITEM_BOD	34.44467	4.255363	ITEM_BOD
ITEM_SS	29.42077	3.956223	ITEM_SS
ITEM_TN	19.99496	3.270896	ITEM_TN
MEAN_hPa	19.03354	3.170830	MEAN_hPa
SUM_SUN	27.59904	3.077448	SUM_SUN
L_REHU	23.97722	2.994235	L_REHU
MEAN_MOIS	27.79763	2.796797	MEAN_MOIS
M_WIND_100	25.66638	2.304458	M_WIND_100
MAX_MOMENT_WIND	27.50093	2.214428	MAX_MOMENT_WIND
MAX_MIN	26.31937	1.994473	MAX_MIN
MEAN_WIND	25.11321	1.702622	MEAN_WIND

다수의 결정트리를 학습하는 앙상블 방법으로,
모델링 시 도출되는 MDG(MeanDecreaseGini)
값을 통해 중요변수 선정



[ITEM_TEMP(수온)를 종속변수로
랜덤포레스트를 실행한 결과]

분석결과 | DNN

1. Train, Test set

- train set : test set = 7 : 3 의 비율로 랜덤 추출(Hold out)

2. 하이퍼파라미터 조절

Layer/Node	3/3
Loss	SSE
Optimizer	Sigmoid
stepmax	1e7

* hidden layer와 node의 수는 2~3개에서 그리드 탐색을 한 결과 (3,3)에서 결과가 가장 좋았다.

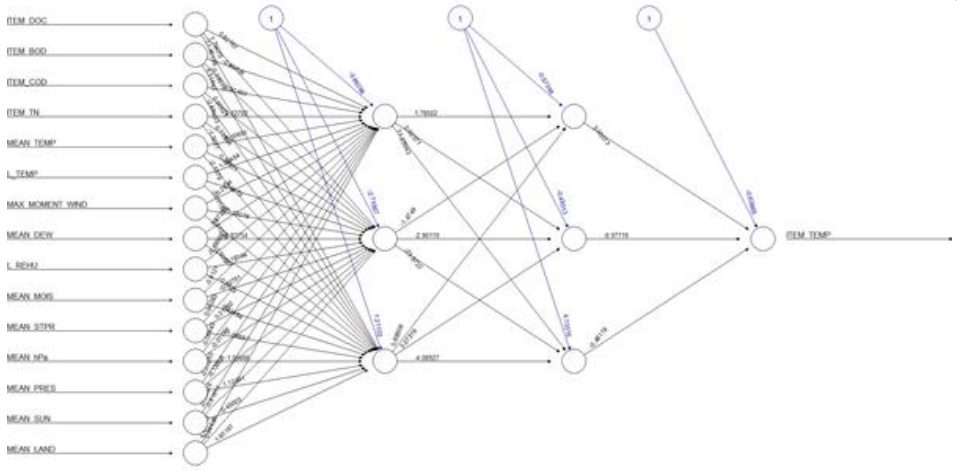
- 최대 계산량을 설정하는 stepmax는 입력변수 및 record가 많아 기본값보다 더 여유를 주어서 1e7로 설정하였다.

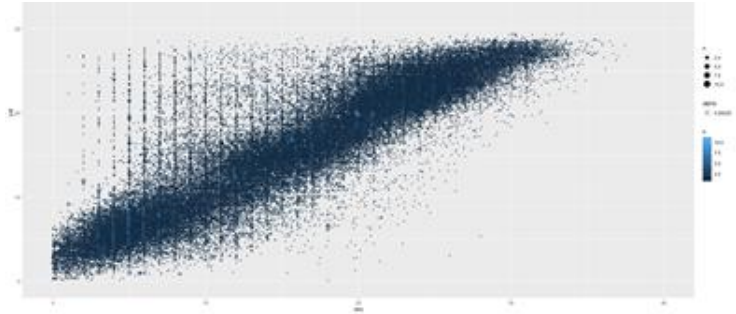
* 회귀예측에 주로 사용하는 SSE loss fuction을 사용하였다.

분석결과 | DNN

2. 모델 결과

- 증가하고 감소하는 추이 등이 잘 맞는 것을 볼 수 있다.





Y축 : 모델을 통해 예측한 값
X축 : 실제 값

< 상관분석 변수 결과>

- RMSE : 0.066
- R-squared : 0.762

< 랜덤포레스트 변수 결과>

- RMSE : 0.0614
- R-squared : 0.788

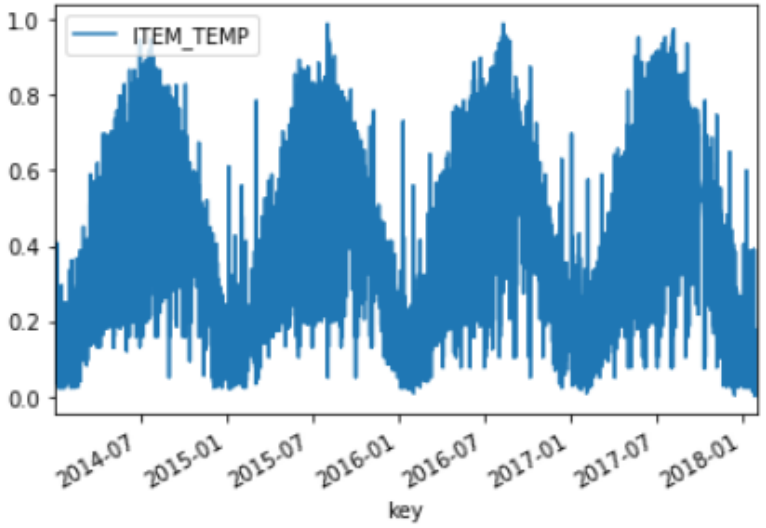
사용!

분석결과 | LSTM(Long Short-Term Memory)

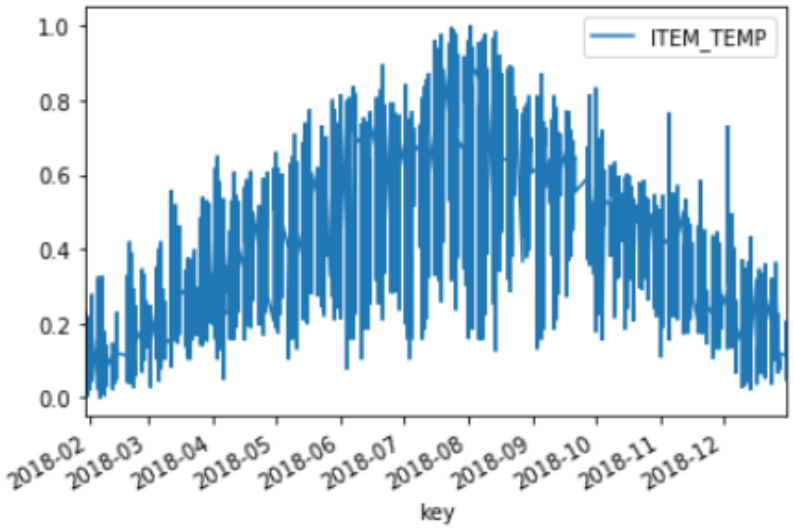
1. Train set, Test set 분리

- train set, test set의 비율을 8:2로 분리 (시간순으로 정렬하여)

  > 8:2 비율을 사용한 이유는 2014년 ~ 2018년 까지 총 5개년의 데이터를 사용했기 때문이다.
  강물의 특성 상 날씨의 영향을 많이 받는데, 날씨는 계절별로 특성을 지닌다.
  대략 4개년도의 데이터로 모델을 학습시키고, 1개년도로 테스트하는 방식을 택했다.



Train set



Test set

분석결과 | LSTM(Long Short-Term Memory)

2. 하이퍼파라미터 조절

Layer	20
Loss	MSE
Optimizer	Adam
epoch	100

* epoch : 전체 데이터 셋에 대해 한번 학습을 완료한 상태
  > epoch가 너무 많으면 overfitting 이 일어나고,
  너무 적으면 underfitting이 일어나기 쉽다.
  > epoch를 많이 돌린 후 early stopping을 이용해서 특정 시점에서
  멈춰야 한다.

* Early stopping 의 특정 시점의 기준
  > 예측값의 성능이 더 이상 개선되지 않을 때 학습을 중지
  > loss의 기준이 'MSE'이므로 MSE가 더 이상 줄어들지 않으면 학습을 중지

분석결과 | LSTM(Long Short-Term Memory)

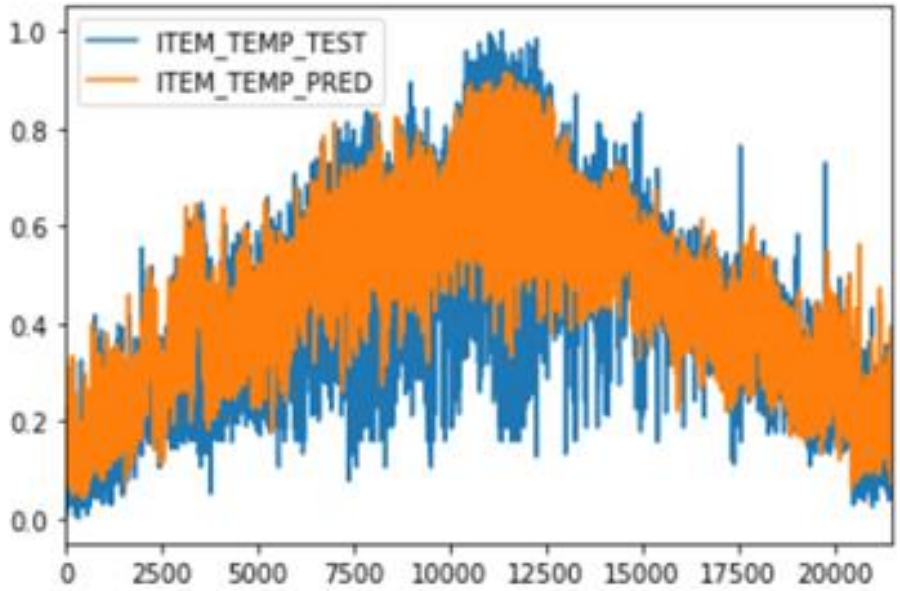
3. 모델 결과

- 증가하고 감소하는 추이 등이 잘 맞는 것을 볼 수 있다.

< 상관분석 변수 결과>
- RMSE : 0.086
- R-squared : 0.827

사용!

< 랜덤포레스트 변수 결과>
- RMSE : 0.096
- R-squared : 0.789



주황 : 모델을 통해 예측한 값
파랑 : 실제 값

분석결과 | Ridge, Polynomial Regression

1. Train, Test set

- train set : test set = 7 : 3 의 비율로 랜덤 추출(Hold out)

2. 하이퍼파라미터 조절

	다중	다항(degree=2)	릿지	랏쏘	다항(degree=3)	릿지	랏쏘	다항(degree=4)	릿지	랏쏘
train	74.372	78.419			81.52			86.125		
test	73.53	77.601			79.779			75.305		
alpha=10			77.679	0		78.928	0		79.748	0
			76.987	-415.36		78.182	-415.36		78.922	-415.36
alpha=1			78.109	0		79.903	0		80.628	0
			77.377	-415.36		79.046	-415.36		79.653	-415.36
alpha=0.001			78.416	73.507		81.28	73.507		83.416	73.507
			77.595	72.768		79.822	72.768		80.284	72.768
alpha=0.000001			78.419	78.154		81.511	80.117			
			66.601	77.399		79.801	79.178			

-> degree = 3
alpha = 0.005로 설정

수온에 대해 RandomForest를 통해 도출된 중요변수를 독립변수로, 수온을 종속변수로 하여 다양한 모델(다중, 다항, 릿지, 랏쏘)을 파라미터를 변경시켰을 때의 R-square값

분석결과 | Ridge, Polynomial Regression

3. 모델 결과

- 오른쪽 그림은 수온에 대한 예측값과 실제값을 그래프로 나타낸 것이다.

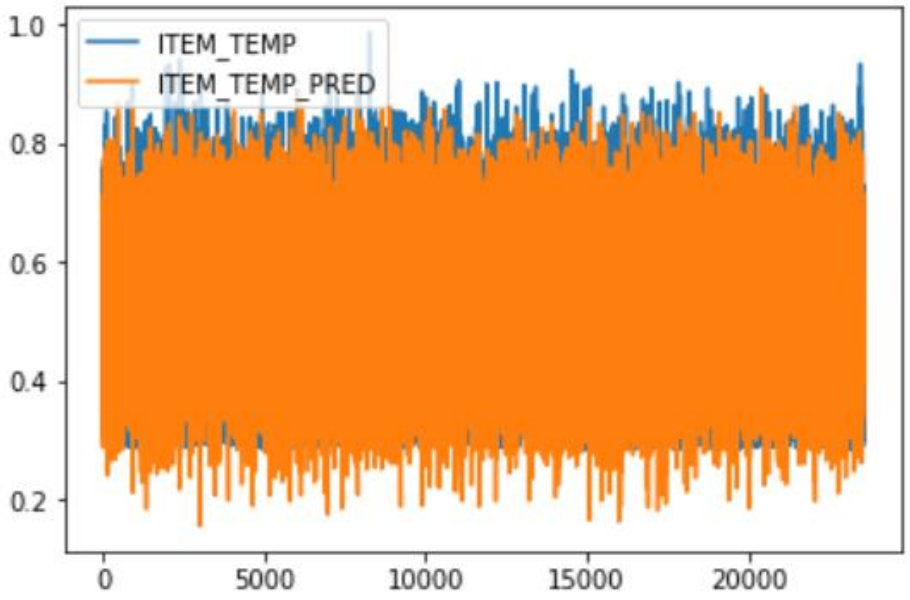
< 상관분석 변수 결과>

- RMSE : 0.086
- R-squared : 0.827

< 랜덤포레스트 변수 결과>

- RMSE : 0.096
- R-squared : 0.789

사용!



주황 : 모델을 통해 예측한 값
파랑 : 실제 값

분석결과 | 결과 정리

R^2/RMSE	수온(TEMP)	생물화학적 산소요구량 (BOD)	화학적 산소요구량 (COD)	총인(TP)	부유물질(SS)	수소이온농도 (PH)	용존 산소 (DOC)
DNN	0.762/0.066	0.401/0.115	0.523/0.137	0.395/0.152	0.216/0.255	0.452/0.105	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	0.598/0.103	0.473/0.145	0.478/0.157	0.244/0.169	0.384/0.110

상관분석을 통해 도출된 변수를 사용한 결과

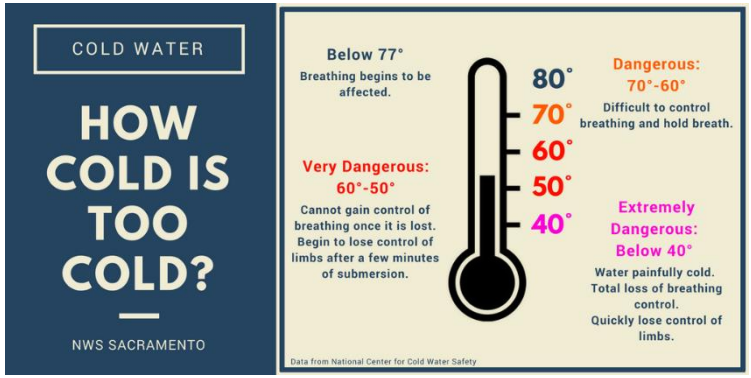
R^2/RMSE	수온(TEMP)	생물화학적 산소요구량 (BOD)	화학적 산소요구량 (COD)	총인(TP)	부유물질(SS)	수소이온농도 (PH)	용존 산소 (DOC)
DNN	0.788/0.061	0.388/0.120	0.428/0.127	0.367/0.155	0.518/0.142	0.258/0.151	0.342/0.114
LSTM	0.789/0.096	0.344/0.119	0.509/0.110	0.346/0.160	0.425/0.162	0.200/0.169	0.404/0.211
다항회귀	0.802/0.061	0.412/0.115	0.527/0.113	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	0.526/0.113	0.505/0.140	0.482/0.155	0.311/0.161	0.382/0.109

랜덤 포레스트로 도출된 변수를 사용한 결과

분석결과 | 예측 지표의 범주화

범주화 : 예측 지표로 레저 별 적합도를 판단하기 위함.

- 물놀이의 경우, 우측 자료처럼 '미국 기상청' 등이 기준을 제공.
- 기관 및 논문, 전문가 의견을 수렴하여 지표산출



[미국 기상청] 온도별 물놀이 기준

수온
26.5
4.1
17.4

기관, 논문, 전문가 의견 등을 이용해 변수별 적합도 산출

수온
위험
안전
보통

종합 레저지수 산출을 위해 Numeric type으로 변환한다

수온
3
1
2

분석결과| 레저별 지표산출을 위한 벡터가중치법 적용

[텍스트 마이닝]
'각 레저를 대표하는 네이버 카페'로 부터
각 예측 지표들의 언급빈도를 가져옴.

[벡터정규화]
각각의 가중치를 산출.

벡터정규화란?
유클리디안 거리를 평가하여
각 지표의 선호순위를 결정하는 기법

$$z_{ij} = \frac{b_{ij}}{\sqrt{\sum_{i=1}^{n} b_{ij}^2}}$$

	수온	수질	유속	수심
스쿠버다이빙	6034	3	1343	27
낚시	1274	522	93	5000
서핑	1683	22	10	1562

[텍스트 마이닝을 통한 각 지표 별 언급 빈도]

가중치	수온	수질	유속	수심
스쿠버다이빙	0.000157902	1.66261E-06	0.0007443	0.037037037
낚시	4.73532E-05	2.06478E-05	3.7187E-06	0.0002
서핑	0.000319178	9.01481E-06	4.0985E-06	0.000640205

[벡터 정규화를 이용하여 산출한 레저별, 변수별 가중치]

분석 결과 | 활용 예시

[예측 지표의 범주화] - 1:위험, 2 : 보통, 3 : 안전

다이빙	수온	수질	유속	수심
20.08.13	3	3	3	3
20.08.14	2	3	1	2
20.08.15	3	2	2	3

곱한다(가중치 적용)

가중치	수온	수질	유속	수심
스쿠버다이빙	0.000157902	1.66261E-06	0.0007443	0.037037037
낚시	4.73532E-05	2.06478E-05	3.7187E-06	0.0002
서핑	0.000319178	9.01481E-06	4.0985E-06	0.000640205

[벡터 정규화를 이용하여 산출한 레저별, 변수별 가중치]

레저지수	스쿠버다이빙	낚시	서핑
20.08.13	보통	보통	안전
20.08.14	위험	위험	위험
20.08.15	보통	위험	안전

[종합 레저 지수]

실용화 방안 | [MyWater]를 통한 수상레저 지수 제공

K-WATER 에서 현재 운영중인 MyWater 어플리케이션 사용

[기존] 원하는 위치와 날짜 별 댐/보의 실시간 수질 정보 등 다수 서비스

[추가] 원하는 날짜, 수상레저, 장소를 선택하여 종합 레저 지수 및
각 요인 별 예측 값 추가

* 사용자 친화적인 [수상 레저 예측 지수]
-> 사용자들이 한눈에 이해하기 좋게 수치 값 뿐만 아니라 위험, 보통, 안전 등을 의미하는 빨강, 노랑, 초록 색 표시를 사용하였다

* 기존의 어플리케이션에 기능을 추가하므로 홍보, 경제적 차원에서도 이점

[예상 어플리케이션 UI]

실용화 방안 | 수상레저 계획 도움 및 안전사고 예방

[수상레저 계획을 세우는 데 도움]

- 여행을 계획할 때 날씨를 고려하여 세부 일정을 정한다.
- 해당 서비스를 이용하면 미리 수상레저 지수를 확인할 수 있게 되어 레저를 즐길 적절한 장소, 날짜 등을 정하고 즐기는데 도움을 받을 수 있을 것이다.

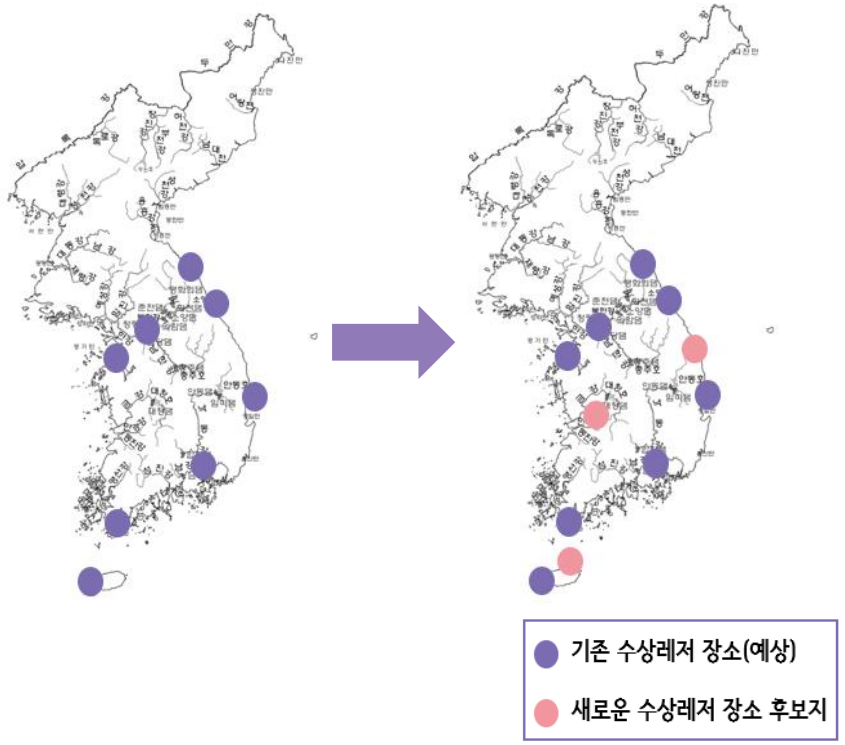
[안전사고 예방]

- 14년도 이전에는 수상레저 사고가 3~4건, 17년도에는 54건에 이르는 등 안전 문제가 빈번하게 발생
- 국민들이 수상레저 지수 예보를 통해 수상환경 및 안정성에 대한 정보를 미리 얻는다면 수상레저 사고를 방지할 수 있을 것으로 예상된다.

실용화 방안 | 안전지킴이 수 예측 및 새로운 레저 관광 지역 개발

효율적인 안전요원 배치 가능

- 예측된 수상환경에 따라,
  안전 지킴이 수를 유동적으로 조절할 수 있다.



종합레저 지수 상 레저에 적합하다고 생각 되는 곳 중
활성화 되지 않은 곳을 발굴 할 수 있다.

기대 효과 : 해당 지역 경제 활성화 + 국민들의 쾌적한
수상레저 환경 제공

결론 및 시사점 | 결론

날씨, 하천 정보 -> 하천의 수온 등의 수상환경 예측

- DNN, LSTM, 릿지 회귀, 라쏘 회귀, 단순 선형 회귀 등의 기법 사용
- 변수 별로 RSME, R-squared 값이 가장 우수한 모델을 사용하여 예측 값 도출
- 논문 또는 전문가의 의견을 수렴하여 각 수상레저별 좋음, 보통, 나쁨 등을 결정할 지표 결정
- 텍스트 마이닝을 통해 변수들의 가중치 산출 후 최종 종합 레저 지수 산출

결론 및 시사점 | 한계점

API 활용가이드와는 달랐던 데이터

- 수위, 유량의 데이터가 실제 API에서는 누락되어 있음
- 따라서 분석 시에는 수온과 수질을 나타내는 몇가지 요소(BOD, COD, PH 등)를 사용함
- 수위, 유량 등의 데이터가 주어지면 동일한 방식으로 학습이 가능하도록 준비해 두었음

국립환경과학원 National Institute of Environmental Research

공공데이터 개방·공유·활용·체계 개발 OpenAPI 활용가이드

wmdep	수심 (단위 : m)	6, 1	0	0.5	수심 (단위 : m)
itemLvl	측정값(수위) (단위 : m)	14, 4	0	40.8	측정값(수위) (단위 : m)
itemAmnt	측정값(유량) (단위 : ㎥/sec)	14, 4	0	1.099	측정값(유량) (단위 : ㎥/sec)
itemTemp	측정값(수온) (단위 : ℃)	14, 4	0	12.4	측정값(수온) (단위 : ℃)
itemPh	측정값(수소이온농도(pH)) (단위 : -)	14, 4	0	7.1	측정값(수소이온농도(pH)) (단위 : -)
itemDoc	측정값(용존산소(DO)) (단위 : mg/L)	14, 4	0	11.5	측정값(용존산소(DO)) (단위 : mg/L)
itemBod	측정값(생물화학적산소요구량(BOD)) (단위 : mg/L)	14, 4	0	3.5	측정값(생물화학적산소요구량(BOD)) (단위 : mg/L)

누락

API를 통해 가져온 실제 데이터

ITEM_TEMP	ITEM_PH	ITEM_DOC	ITEM_BOD
5.3	8.9	13.4	0.3
3.4	8	10.4	0.5
3.8	7.9	13.8	0.3
5.4	8.2	10.1	1.3
4.1	8.3	14.2	0.3
5.9	8.2	10.8	0.3

API 활용가이드 에는 수심, 수위, 유량이 제공된다고 되어있지만, 실제 데이터에는 누락되어있다.

감사합니다