**데이터마이닝**

**프로젝트**

주제 : 인천공공시설 예약자 수 예측

정보통신공학과

12181785 서혜빈

**목차**

1. 주제
   1. 선정 배경
   2. Problem formulation
2. 사용 데이터 리스트 및 데이터 설명
   1. 인천공공데이터 포털
   2. 기상청
3. Data preparation
4. Data mining process
   1. Sampling
   2. Exploration
   3. Modification
   4. Modeling
   5. Assessment
5. 결론 및 적용방안
6. **주제**
   1. **선정 배경**

인천광역시에는 여러 공공시설물이 존재한다. 축구장, 테니스장 등이 그 예이다. 이러한 시설물들은 사전 예약을 통해 이용을 하게 된다. 기상 예보가 맑음이면 해당 날짜에는 예약률이 높을 것이다. 반대로 비가 오거나 눈이 온다면 예약을 하지 않을 것이다. 이렇듯 공공시설물 예약률에는 날씨가 영향을 미친다고 생각한다. 그러나 기상 예보가 맑음이어도 날짜에 따라 예약률이 달라질 수 있다. 이는 곧 날짜도 예약률에 영향을 미친다는 것이다.

그렇다면 날씨와 날짜를 통해 공공시설 예약률을 예측한다면 어떤 점이 좋을까?

공공시설물을 이용할 때 예약자가 적다면 이용객들은 더 쾌적하게 시설을 이용할 수 있다. 따라서 예약률 예측을 통해 날씨는 좋지만, 예약률이 적을 것으로 예상되는 날짜에 대한 인사이트를 추출할 수 있다. 그리고 이를 이용객에게 제공할 수 있다는 장점이 있다.

뿐만 아니라 이용률이 꾸준히 증가할 것으로 예측되는 시설물의 경우, 시설을 증축하는 등의 방안을 검토할 수 있다. 즉, 공공시설물 관계자에게 의사결정에 대한 인사이트를 제공하는 것이다.

마지막으로 해당 데이터 분석을 통해 어떤 공공시설물의 이용률이 높은지, 그 이유는 무엇인지에 대해 예상해 볼 수 있다는 것이다. 이는 인천의 지역홍보 등 포괄적으로 이용될 수 있다고 생각한다.

위와 같은 이유로 ‘날씨 및 날짜에 따른 인천공공시설 예약률 예측’ 주제를 선정하게 되었다.

* 1. **Problem formulation**

1. 데이터 수집

인천공공데이터포털에서 인천 시설 이용 데이터, 기상청에서 날씨 데이터를 수집한다.

1. 데이터 전처리

데이터의 결측치를 확인하고, 처리한다. 또한 새로운 칼럼을 만들 필요가 있으면 만들어준다. 그리고 필요 없는 칼럼을 삭제한다.

1. EDA

히스토그램, 왜도 등 시각화를 통해 데이터를 분석한다. 또한 정규성을 만족하지 못한다면 로그변환 등을 해준다.

1. Modeling

회귀분석, 라쏘회귀, 릿지회귀, 랜덤포레스트 등 여러 모델링을 진행한다. RMSLE를 구해 가장 적은 RMSLE를 보이는 모델을 선택한다. 그리고 앙상블을 해준다.

*반응변수 Y : 예약자 수*

*설명변수 X : 풍속, 강수량, 눈, 기온, 년, 월, 일 등*

분석을 진행할 때, 변수는 위와 같다.

1. **사용 데이터 리스트 및 데이터 설명**

‘인천공공데이터포털’에서 인천 시설 이용 데이터를, ‘기상청’에서 날씨 데이터를 가져온다.

인천 시설 이용 데이터에는 예약일, 단체명, 이용시작일, 이용종료일 등의 Column이 있다. 날씨 데이터에는 평균기온, 일강수량, 평균풍속, 최심신적설, 최심적설 등의 Column이 있다. 이 중 반응변수(Y)는 예약자 수이고, 설명변수(X)는 년, 월, 일, 평균기온, 일강수량, 평균풍속, 최신심적설, 최심적설 등이 될 예정이다.

인천 시설 이용 데이터, 기상청 데이터 모두 예약자 수는 존재하지 않는다. 이는 데이터 가공 과정에서 새로 Column을 만들어 줄 예정이다.

* 1. **인천데이터포털**

인천데이터포털에서 인천\_시설\_이용\_데이터.csv를 수집한다. 해당 데이터는 2010년부터 2019년까지 존재한다. 해당 데이터는 reservation.csv로 저장한다.

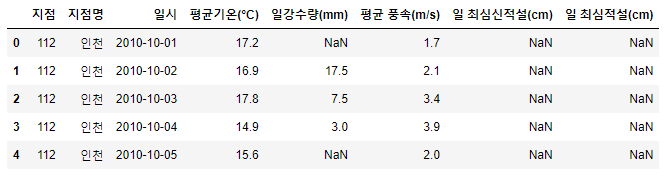
* 1. **기상청**

기상청에서 2010년 10월부터 2019년 12월까지의 날씨 데이터를 수집한다. 해당 데이터는 weather.csv로 저장한다.

결론적으로 예약 데이터에서는 이용시작일을 기준으로 예약자수를 구하여 새로운 칼럼을 추가할 예정이다. 따라서 예약자수와 이용시작일을 바탕으로 분석을 진행할 것이다. 날씨 데이터에서는 지점 등을 제외한 모든 데이터를 사용할 예정이다.

1. **Data preparation**

* Weather



Weather 데이터는 위와 같다. 지점, 지점명은 필요하지 않으므로 삭제해준다.

* Reservation

테이블이(가) 표시된 사진

자동 생성된 설명

Reservation 데이터는 위와 같다.

* 1. **결측치 확인 및 처리**
* Weather

텍스트이(가) 표시된 사진

자동 생성된 설명

결측치의 수를 count 한 결과, 일강수량, 평균 풍속, 일 최심신적설, 일 최심적설에서 결측치가 발견되었다. 하지만 이는 비가 내리지 않아서, 바람이 불지 않아서, 눈이 내리지 않아서와 같은 이유로 인한 것이다. 따라서 해당 결측치는 0을 넣어주었다.

* Reservation

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

Reservation 데이터의 경우 단체명의 결측치가 굉장히 많고, 인원에서 1개가 발견되었다. 단체명의 경우 이번 분석에서는 필요하지 않은 Column이므로 임의로 0을 채운다.

인원의 결측치는 발생한 row를 삭제해주었다. 임의로 0을 채우면 결과에 영향을 줄 것이라 예상했기 때문이다.

* 1. **Data merge**

Weather의 ‘일시’ 열과 reservation의 ‘이용시작일’ 열을 기준으로 두 데이터를 합쳐 주었다. 합쳐준 데이터는 merged라고 명명했다.

텍스트이(가) 표시된 사진

자동 생성된 설명

그런데 또 다시 결측치가 발생하였다. 이용시작일의 경우 10이라는 노이즈도 있다는 것을 알 수 있다.

<결측치 수> <결측치가 발생한 열>

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명 테이블이(가) 표시된 사진

자동 생성된 설명

해당 결측치들은 이용종료일이 노이즈이거나 2011-10-16의 결측치이다. 따라서 해당 행은 모두 삭제해준다. 노이즈의 경우 대체할 값이 없기 때문이고, 2011-10-16 하루의 데이터가 결과에 큰 영향을 주지 않는다고 판단하였기 때문이다.

Merged 데이터는 불필요한 자료를 많이 포함하고 있다. 더불어 예약 인원이 날짜별로 count 되어 있지 않아 반응 변수에 적합하지 않다. 따라서 새로운 데이터 프레임을 생성했다.

테이블이(가) 표시된 사진

자동 생성된 설명

이용시작일을 기준으로 날짜별로 이용한 인원을 count 해주었다. 그리고 불필요한 columns을 삭제했다. 사업부서, 대관시설, 예약접수일 등이 그 예이다.

테이블이(가) 표시된 사진

자동 생성된 설명

그 후, weather와 다시 이용시작일과 일시를 기준으로 merge 해주었다. 그리고 각 열들의 이름을 재명명했다.

<각 변수 설명>

*\* date : 이용일*

*\* reserved\_num : 예약단체 수 (Number of reserved organization*

*\* avg\_temp : 평균기온*

*\* daily\_precipitation : 일강수량*

*\* avg\_wind : 평균풍속*

*\* latest\_snow : 최신심적설 (새로 와서 쌓인 눈의 두께)*

*\* most\_snow : 최심적설 (최대로 깊었던 적설량. 며칠씩 쌓인 최대량)*

1. **Data mining process**
   1. **Sampling**

해당 사항 없음.

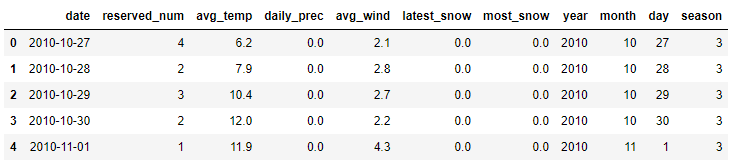
* 1. **Exploration**
     1. **데이터 가공**
* **date를 년, 월, 일로 분리하기**

테이블이(가) 표시된 사진

자동 생성된 설명

2010-10-27과 같이 입력되어 있는 열을 year, month, day로 분리해주었다.

* **범주화**



Month 열을 기준으로, 각 계절을 범주화 해주었다. 범주 기준은 아래와 같다.

*봄 : 3월, 4월, 5월 -> 1*

*여름 : 6월, 7월, 8월 -> 2*

*가을 : 9월, 10월, 11월 -> 3*

*겨울 : 12월, 1월, 2월 -> 4*

테이블이(가) 표시된 사진

자동 생성된 설명

다음으로는 avg\_temp를 범주화 해주었다. 범주 기준은 아래와 같다.

*-5도 이하 -> 0*

*-5도 초과, 0도 이하 -> 1*

*0도 초과, 5도 이하 -> 2*

*5도 초과, 10도 이하 -> 3*

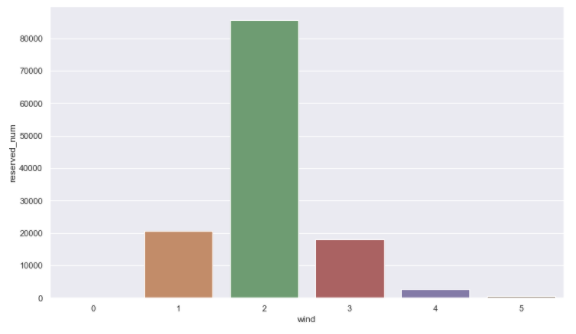
*10도 초과, 15도 이하 -> 4*

*15도 초과, 20도 이하 -> 5*

*20도 초과, 25도 이하 -> 6*

*25도 초과, 30도 이하 -> 7*

*30도 초과 -> 8*



다음으로는 avg\_wind를 범주화 해주었다.

*0.0 이하 -> 0*

*0.0도 초과, 2도 이하 -> 1*

*2도 초과, 4도 이하 -> 2*

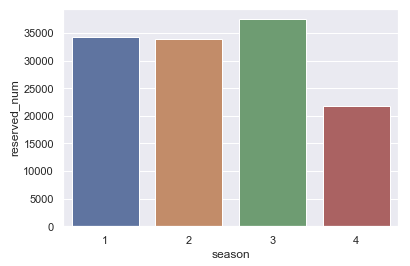
*4도 초과, 6도 이하 -> 3*

*도 초과, 8도 이하 -> 4*

*8도 초과 -> 8*

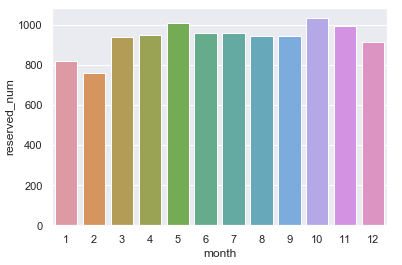
**4.2.2 시각화**

1. Season별 예약자 수



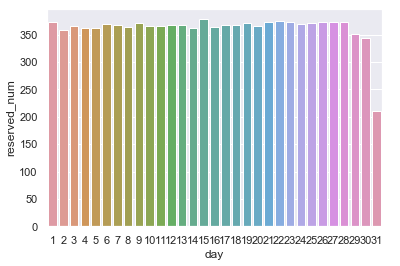
계절간 예약자 수가 큰 차이 없는 것으로 보인다. 다만 겨울인 4 범위에서 예약자 수가 가장 없다는 것을 알 수 있다.

1. Month별 예약자 수



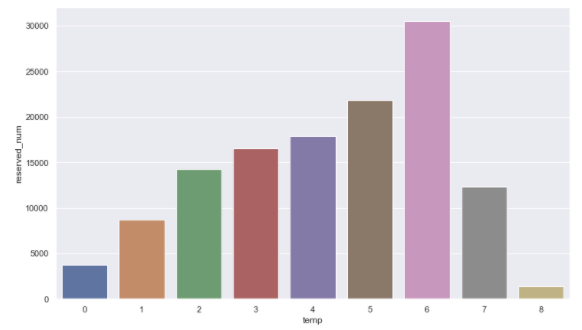
월별 예약자 수가 큰 차이 없어 보이지만, 겨울인 1월과 2월에 가장 적다는 것을 확인할 수 있다.

1. Day별 예약자 수



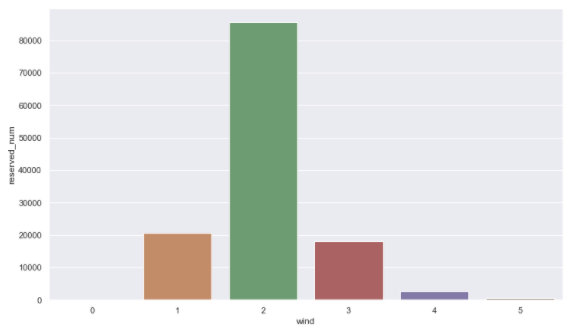
일별 예약자 수도 큰 차이 없다. 다만 31일이 유독 적은데, 이는 월마다 31일이 없는 월도 있기 때문으로 보인다.

1. Avg\_temp를 범주화한 temp별 예약자 수



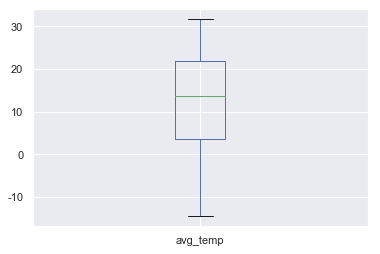
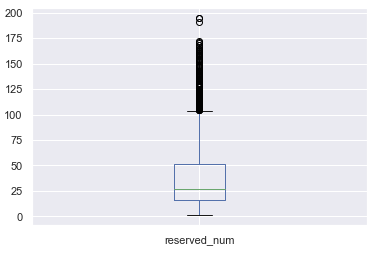
온도별 예약자 수는 큰 차이를 보인다. 온도가 30도 초과일 때에는 예약자 수가 가장 적었다는 것을 알 수 있다. 그리고 온도가 -5도 이하일 때에도 예약자 수가 적었다. 이는 기온이 예약자 수에 영향을 줌을 의미한다고 생각한다.

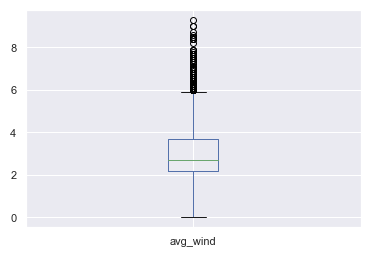
1. Avg\_wind를 범주화한 wind별 예약자 수



풍속에 따라서도 예약자 수가 큰 차이를 보였다. 바람이 가장 많이 부는 범위에서 예약자 수가 급격히 감소하였다. 이를 통해 풍속도 예약에 영향을 준다고 생각된다.

1. 이상치

테이블이(가) 표시된 사진

자동 생성된 설명

테이블이(가) 표시된 사진

자동 생성된 설명

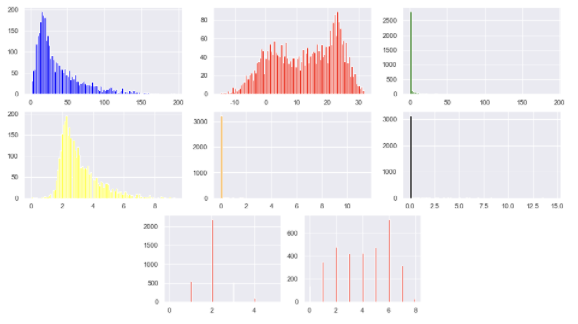
일부 박스 플롯을 확인한 결과, 이상치가 존재하는 변수들이 있었다. 하지만 눈, 비, 바람의 경우 날씨이기 때문에 당연한 결과라고 생각된다.

예약자 수가 굉장히 많은 변수가 발견되었는데, 정규성을 만족하지 못할 것으로 예상된다.

* 1. **Modification**

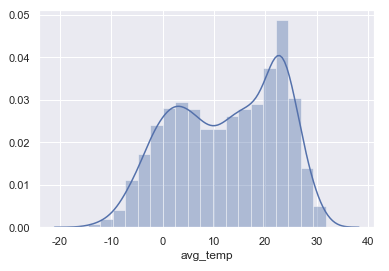
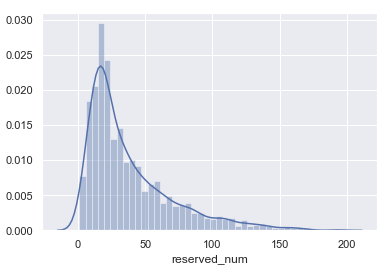
**4.3.1 정규성**

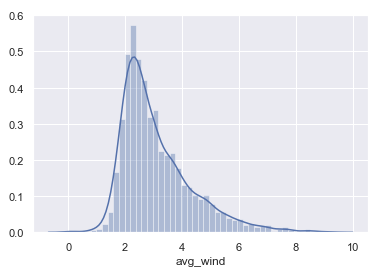
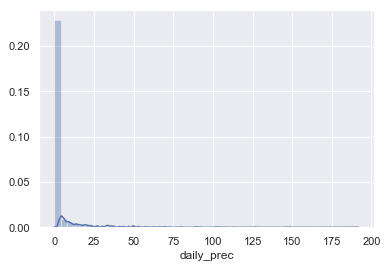
* **히스토그램**



Reserved\_num, avg\_temp, daily\_prec, avg\_wind, latest\_snow, most\_snow, wind, temp 순의 히스토그램이다. 대다수 좌우대칭이 아니고, 한쪽으로 쏠려있는 분포를 보여준다. 따라서 정규성을 만족하지 못하고 있다.

* **왜도**





시각화를 통해 한쪽으로 쏠려 있음을 더욱 명백하게 확인할 수 있다.

테이블이(가) 표시된 사진

자동 생성된 설명

왜도를 계산했을 때, 0.7 이상인 변수는 latest\_snow, most\_snow, daily\_prec, reserved\_num, avg\_wind, wind이다. 왜도를 줄이고, 정규분포와 비슷하게 만들어주기 위해 로그변환을 해보자.

테이블이(가) 표시된 사진

자동 생성된 설명

로그변환 후에도 아직 왜도가 남아 있는 변수가 있다. 하지만 이전보다는 나아졌다고 할 수 있다. 혹시 몰라 제곱, 세제곱 변환도 해보았지만 왜도가 더욱 심해졌다. 따라서 로그변환한 변수를 가지고 간다.

**4.3.2 상관관계**

텍스트이(가) 표시된 사진

자동 생성된 설명

히트맵을 통해 상관관계를 시각화했다. Latest\_snow와 most\_snow, wind와 avg\_wind, temp와 avg\_temp 사이에서 약간의 상관성이 발견되었다. Wind와 avg\_wind, temp와 avg\_temp의 경우 avg 변수를 바탕으로 범주화한 것이기 때문이다. Latest\_snow와 most\_snow의 경우 뒤에서 더 살펴보고 어떻게 처리할지 결정해보자.

* + 1. **P-value & VIF**

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

해당 결과를 살펴보면 latest\_snow와 most\_snow 중 latest\_snow의 P값이 더 높게 나타났다. 따라서 앞으로 latest\_snow 변수를 제거한 데이터프레임을 가지고 분석을 진행한다.

텍스트, 전자기기, 스크린샷이(가) 표시된 사진

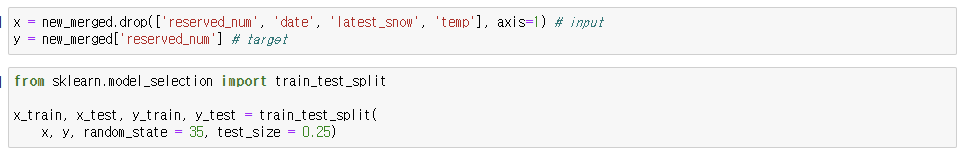
자동 생성된 설명

다중공선성 확인해 본 결과 avg\_temp와 temp가 10이 넘는 결과가 나왔다. 따라서 더 높은 avg\_temp 변수를 삭제해주었다.

* 1. **Modeling**

**4.4.1 RMSLE**

RMSE에 로그를 취하는 방식인 RMSLE를 구하여 모델을 채택하는 방식으로 모델링을 진행했다. 각 모델링 후 RMSLE를 출력하여 값이 가장 작은 모델 몇 개를 선택하여 앙상블을 진행할 예정이다.



우선 test set과 train set으로 나누어 주었다. 학습과 성능 확인을 위함이다.

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

Linear Regression, Lasso Regression, Elasic Net Regression, Kernel Ridge Regression, Gradient Boosting Regression, XGBoost, LightGBM, Random Forest Regression 모델을 사용했다.

* 1. **Assessment**

|  |  |
| --- | --- |
| 모델 | RMSLE |
| Linear Regression | 0.66057 |
| Lasso Regression | 0.66057 |
| Elastic Net Regression | 0.66057 |
| Kernel Ridge Regression | 0.65305 |
| Gradient Boosting Regression | 0.38031 |
| XGBoost | 0.44656 |
| LightGBM | 0.5567 |
| Random Forest Regression | 0.24062 |

따라서 RMSLE가 가장 적은 Random Forest Regression, Gradient Boosing Regression, XGBoost 모델을 선택했다.

텍스트이(가) 표시된 사진

자동 생성된 설명

마지막으로, 다음과 같이 앙상블을 진행했다.

1. **결론 및 적용방안**

씨와 날짜를 통해 인천의 공공시설물 예약률 예측을 진행해보았다. 여러 그래프로 시각화해보며 온도와 바람, 눈 등이 공공시설물을 예약하는 데에 있어 영향을 준다고 예상해 볼 수 있었다. 또한 월별, 일별, 계절별 시각화를 통해 겨울에 예약률이 상대적으로 낮다는 것을 알 수 있었다.

오차를 통한 모델링을 하고, 앙상블을 통해 예측을 해보았다. 이러한 예측을 통해 공공시설물을 예약할 때, 사람들은 보다 적은 예약이 있는 날을 결정할 기회를 얻을 수 있다. 또한 인천 관계자 분들은 예약률이 증가할 것이라 예측되는 시설물을 증축하는 방안을 검토할 인사이트를 제공받을 수 있다.

더불어 꾸준히 예약률이 증가할 것이라 예상되는 시설물의 경우, 왜 증가하는 지에 대해 분석할 수 있다. 이를 통해 인천이라는 지역의 홍보까지도 나아갈 수 있을 것이라 생각한다.