[Spring 2023] DATA MINING THEORY AND APPLICATION (IIE 4102)

Practice-HW7
Decision Tree

논문참조

강세정, 김규현, 신재욱, 장현우

01

Introduction

Research Objective

Data

02

Body

Concept of DT

Classification DT

Regression DT

03

Conclusion

Result

Limitation

Development

Introduction / Body

Contents

Decision Tree

01. Research Objective

- 날씨 요인에 따라 따릉이 이용이 활성화되었는지 파악해보자
- 날씨 요인에 따라 따릉이 이용객 수를 수치적으로 예측해보자

02. Data

X(Numerical Variable)

평균기온 (℃)	최고기온 (℃)	일강수량 (mm)	최대풍속 (m/s)	평균풍속 (m/s)	최소상대습도(%)	평균상대습도(%)	가조시간 (hr)
합계일조시간(hr)	1시간 최다 일사량 (MJ/m2)	합계 일사량 (MJ/m2)	일 최심적설 (cm)	평균전운량 (1/10)	9-9강수 (mm)	안개 계속시간(hr)	

X(Categorical Variable)

강수여부 (0, 1)	적설여부 (0, 1)	안개여부 (0, 1)	휴일여부 (0, 1)	미세먼지 기준 (1, 2, 3, 4)	초미세먼지 기준 (1, 2, 3, 4)

Data1_Y	Data2_Y
일일 서울시 공공자전거 (따릉이) 대여 건수	일일 서울시 공공자전거 (따릉이) 활성 정도(0 / 1)

「대기환경보전법」에 따른 미세먼지, 초미세먼지 기준

물질	단위	산정기준	등급			
			좋음(1)	보통(2)	나쁨(3)	매우나쁨(4)
미세먼지	㎍/㎥	24시간	0~30	31~80	81~150	151~
초미세먼지	㎍/㎥	24시간	0~15	16~35	36~75	76~

데이터는 서울 열린 데이터 광장, 공공데이터포털, 기상청 등에서 직접 수집했다

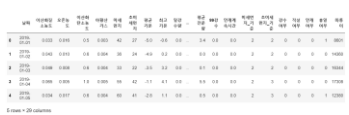
데이터 기간 : 2019.01.01 – 2019.12.31

Data2_Y 값은 Data1_Y 값에 대해 대여 건수 평균값 이상(1) / 이하(0)로 Binary 데이터로 변환했다

→ Data1_Y에 대해 Regression DT, Data2_Y에 대해 Classification DT 수행 (해석 가능)

03. Preprocessing & EDA

① `data.head()`



② `data.describe()`

③ `data.isnull.sum()`

```
날짜          0
이산화질소농도     0
오존농도        0
이산화탄소농도     0
아황산가스       0
미세먼지        0
초미세먼지       0
평균기온        0
최고기온        0
일강수량        0
최대풍속        0
평균풍속        0
최소상대습도      0
평균상대습도      0
가조시간        0
합계일조시간      0
1시간최다일사량    0
합계일사량       0
일최심적설       0
평균전운량       0
99강수        0
안개계속시간      0
미세먼지_기준     0
초미세먼지_기준    0
강수여부        0
적설여부        0
안개여부        0
휴일여부        0
따릉이         0
dtype: int64
```

④ `Variable Selection` : 회귀 모형에서 유의미한 변수들 선택

Data1_Y

```
selected_variables
['최고기온', '99강수', '합계일조시간', '초미세먼지', '일강수량']
```

Data2_Y

⑤ `VIF` 파악 : 모두 5를 크게 초과하고 있지 않음을 확인

	variables	VIF
0	평균기온	5.729010
1	일강수량	1.361139
2	초미세먼지	2.850703
3	평균전운량	4.070583
4	합계일사량	4.945388

Data1_Y

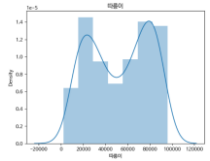
	variables	VIF
0	최고기온	3.608792
1	99강수	2.099618
2	합계일조시간	3.671065
3	초미세먼지	1.987403
4	일강수량	2.030607

Data2_Y

⑥ `Data2_Y RandomUndersampling`

```
따릉이
0    143
1    143
dtype: int64
```

⑦ `Data1_Y EDA`



타겟값인 따릉이 이용량의 분포는 그림과 같이, 이상적인 정규분포의 형태는 아님이 확인됨

Body. Concept of DT

01. CART, CHAID, C4.5

(1) CART

- 지니지수(이산형 목표변수) 또는 분산감소량(연속형 연속변수)을 이용하여 이진 분할을 수행
- C4.5와 매우 유사하지만 연속형 목표변수를 지원하고 규칙 집합을 계산하지 않는다는 점에서 다름

(2) CHAID

- 카이제곱-검정(이산형 목표변수) 또는 F-검정(연속형 목표변수)을 이용하여 다지 분리를 수행
- CART와 다른 점은 CHAID는 데이터를 overfitting 하기 전에 나무 형성을 멈춘다는 것

(3) C4.5

- 엔트로피 지수를 활용하여 범주형 변수뿐만 아니라 연속형 변수를 사용 가능
- 결측치가 포함된 데이터도 사용 가능하며 과적합을 방지하기 위한 가지치기
- CART와 다른 점은 CART는 이진분리를 하지만 C4.5는 가지의 수를 다양화할 수 있음

02. Classification, Regression DT

(1) Classification DT

- 목표변수가 이산형인 분류나무의 경우 상위노드에서 가지분할을 수행할 때, 분류(기준)변수와 분류 기준값의 선택 방법으로 카이제곱 통계량의 p-값, 지니 지수, 엔트로피 지수 등이 사용

(2) Regression DT

- 목표변수가 연속형인 회귀나무의 경우에는 분류변수와 분류 기준값의 선택방법으로 F-통계량의 F-값, 분산의 감소량 등이 사용

Body

Contents

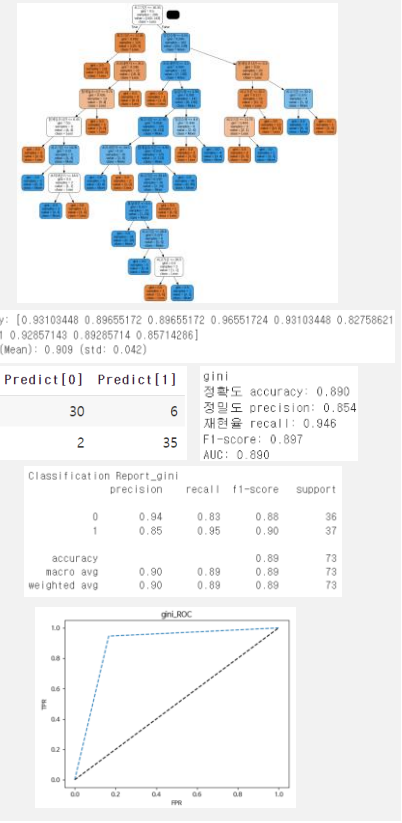
Decision Tree

Classification DT

01. Classification DT Modeling

Gini / Entropy 두 가지 방법으로 각각 DT를 그린 후 비교하였다.

① Gini



```
CV accuracy: [0.93103448 0.89655172 0.89655172 0.96551724 0.93103448 0.82758621
 0.96428571 0.92857143 0.89285714 0.85714286]
CV accuracy(Mean): 0.909 (std: 0.042)
```

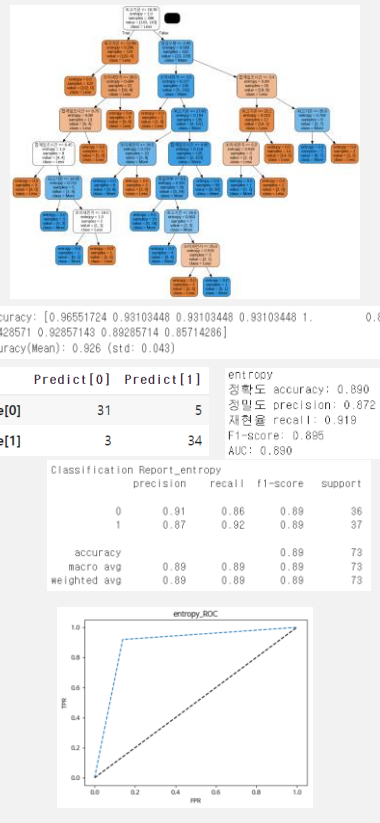
	Predict[0]	Predict[1]
True[0]	30	6
True[1]	2	35

```
gini
정확도 accuracy: 0.890
정밀도 precision: 0.854
재현율 recall: 0.946
F1-score: 0.897
AUC: 0.890
```

Classification Report_gini

	precision	recall	f1-score	support
0	0.94	0.83	0.88	36
1	0.85	0.95	0.90	37
accuracy			0.89	73
macro avg	0.90	0.89	0.89	73
weighted avg	0.90	0.89	0.89	73

② Entropy



```
CV accuracy: [0.96551724 0.93103448 0.93103448 0.93103448 1.         0.86206897
 0.96428571 0.92857143 0.89285714 0.85714286]
CV accuracy(Mean): 0.926 (std: 0.043)
```

	Predict[0]	Predict[1]
True[0]	31	5
True[1]	3	34

```
entropy
정확도 accuracy: 0.890
정밀도 precision: 0.872
재현율 recall: 0.919
F1-score: 0.895
AUC: 0.890
```

Classification Report_entropy

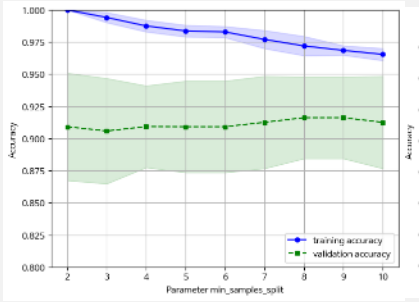
	precision	recall	f1-score	support
0	0.91	0.86	0.89	36
1	0.87	0.92	0.89	37
accuracy			0.89	73
macro avg	0.89	0.89	0.89	73
weighted avg	0.89	0.89	0.89	73

두 방법을 비교한 결과 정밀도는 entropy가, 재현율은 gini가 더 높게 나타났다. 그러나 Confusion Matrix에서 확인할 수 있듯이, 두 방법에서 틀린 예측의 수가 동일하여 AUC 값은 0.890으로 동일하게 나타났으며 F1 score도 0.895 / 0.897로 거의 차이가 없었다.

02. Optimization

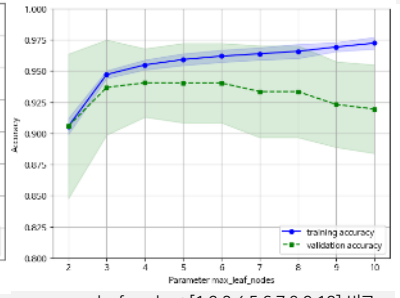
Gini / Entropy 두 가지 방법에 대해 각각 min_samples_split, max_leaf_nodes, max_depth 세 가지 hyperparameter tuning을 진행하였다.

① Gini



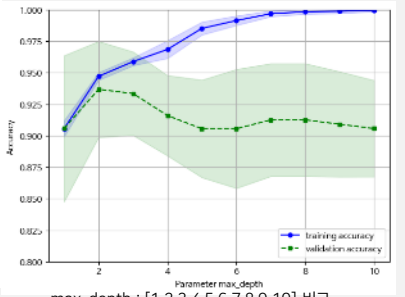
min_samples_split : [1,2,3,4,5,6,7,8,9,10] 비교

오버피팅 발생 전인 8, 9중에서 train accuracy가 더 높은 8 선택



max_leaf_nodes : [1,2,3,4,5,6,7,8,9,10] 비교

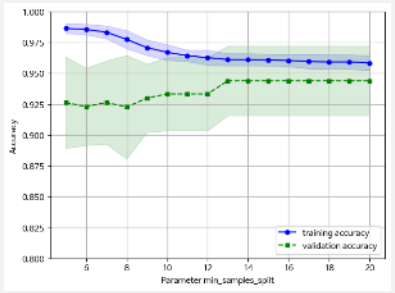
오버피팅 발생 전인 6 선택



max_depth : [1,2,3,4,5,6,7,8,9,10] 비교

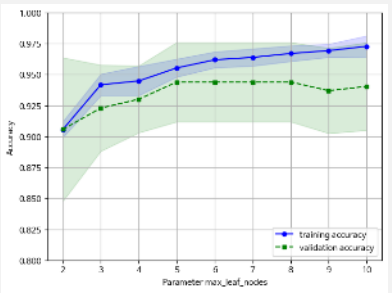
오버피팅 발생 전인 2 선택

② Entropy



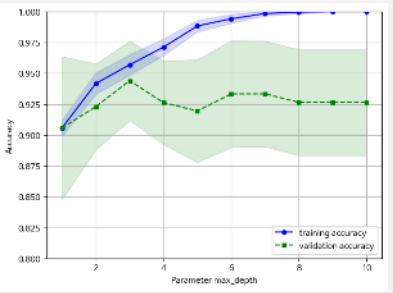
min_samples_split : 10까지 했을 때 결정되지 않아 20까지 확인

13부터 val acc이 높아지지 않고 train acc이 떨어지므로 13 선택



max_leaf_nodes : [1,2,3,4,5,6,7,8,9,10] 비교

오버피팅 발생 전인 8 선택



max_depth : [1,2,3,4,5,6,7,8,9,10] 비교

오버피팅 발생 전인 3 선택

Body

Contents

Decision Tree

03. Grid Search

① gini

param_range1 = [1,2,3,4,5]
param_range2 = [1,2,3,4,5]
param_range3 = ['gini']
param_range4 = [2,3,4,5,6,7]
param_range5 = [2,3,4,5,6,7]

결과

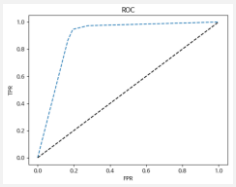
Best score 0.9439655172413793
'decisiontreeclassifier__max_depth': 3,
'decisiontreeclassifier__max_leaf_nodes': 4,
'decisiontreeclassifier__min_samples_leaf': 2,
'decisiontreeclassifier__min_samples_split': 2

	Predict[0]	Predict[1]
True[0]	30	6
True[1]	2	35

정확도 accuracy: 0.890
정밀도 precision: 0.854
재현율 recall: 0.946
F1-score: 0.897
AUC: 0.890

Classification Report

	precision	recall	f1-score	support
0	0.94	0.83	0.88	36
1	0.85	0.95	0.90	37
accuracy			0.89	73
macro avg	0.90	0.89	0.89	73
weighted avg	0.90	0.89	0.89	73



② entropy

param_range1 = [1,2,3,4,5]
param_range2 = [1,2,3,4,5]
param_range3 = ['entropy']
param_range4 = [2,3,4,5,6,7]
param_range5 = [2,3,4,5,6,7]

결과

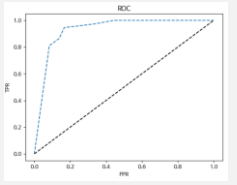
Best score 0.9474137931034482
'decisiontreeclassifier__max_depth': 4,
'decisiontreeclassifier__max_leaf_nodes': 6,
'decisiontreeclassifier__min_samples_leaf': 5,
'decisiontreeclassifier__min_samples_split': 2

	Predict[0]	Predict[1]
True[0]	31	5
True[1]	3	34

정확도 accuracy: 0.890
정밀도 precision: 0.872
재현율 recall: 0.919
F1-score: 0.895
AUC: 0.890

Classification Report

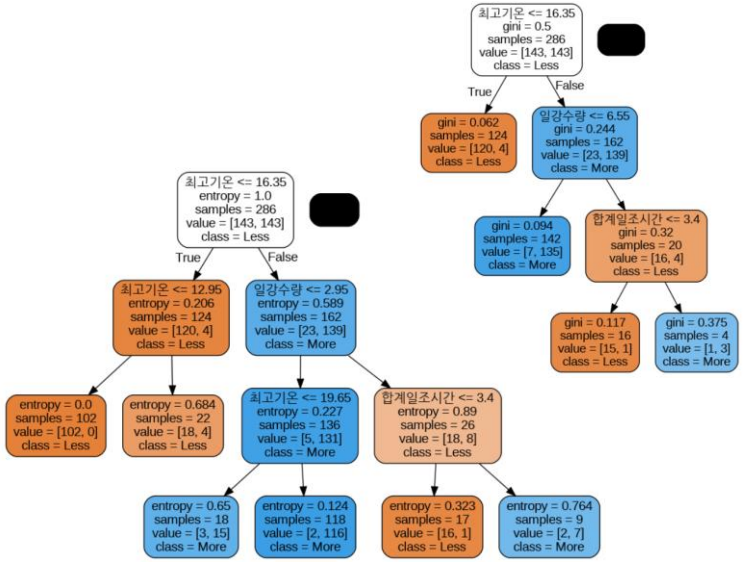
	precision	recall	f1-score	support
0	0.91	0.86	0.89	36
1	0.87	0.92	0.89	37
accuracy			0.89	73
macro avg	0.89	0.89	0.89	73
weighted avg	0.89	0.89	0.89	73



Grid search 결과 가장 성능이 좋은 hyperparameter를 찾긴 하였으나, Gini와 entropy 모두 Default model과 동일한 성능을 보이고 있다.

따라서 두 방법을 비교하면 마찬가지로 정밀도는 entropy가, 재현율은 gini가 더 높게 나타났다. Confusion Matrix에서 확인할 수 있듯이, 두 방법에서 틀린 예측의 수가 동일하여 AUC 값과 F1 score를 유사한 수준으로 얻었다.

04. Classification DT 규칙해석



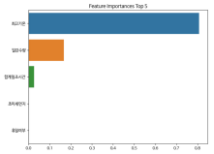
Criterion : Gini

가장 우선적으로 최고 기온이 16.35도를 기준으로 그 이하일 경우 전체적으로 따릉이 대여가 비활성화된다.
최고 기온이 16.35도보다 높을 때에는, 일 강수량이 6.55보다 낮을 때 활성화되며, 일 강수량이 6.55보다 높지만 합계 일조시간이 3.4시간 이상인 경우는 활성화된다.
→ 우선적으로 기온이 일정 수준 이상이며, 강수량이 적거나 강수량이 많더라도 합계 일조시간이 많은 경우 따릉이 이용이 활성화된다.

Criterion : Entropy

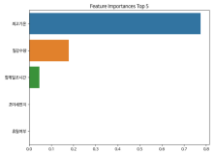
가장 우선적으로 최고 기온이 16.35도를 기준으로 그 이하일 경우 전체적으로 따릉이 대여가 비활성화되고, 이상일 경우 활성화되었다.
최고 기온이 16.35도보다 높을 때에는, 일 강수량이 2.95보다 낮을 때 활성화되며, 일 강수량이 2.95보다 높지만 합계 일조시간이 3.4시간 이상인 경우는 활성화된다.
→ 마찬가지로 우선적으로 기온이 일정 수준 이상이며, 강수량이 적거나 강수량이 많더라도 합계 일조시간이 많은 경우 따릉이 이용이 활성화된다.

05. Feature Importance



Criterion : Gini

Feature Importance의 경우 최고기온, 일강수량, 합계일조시간, 초미세먼지, 휴일여부 순으로 높았다. 최고 기온이 따릉이 활성화 정도에 가장 큰 영향을 미치며, 그 정도가 다른 요인들에 비해 압도적임을 알 수 있다. 최고 기온이 가장 우선적으로 분리 기준으로 사용되었다는 사실을 다시 한 번 확인할 수 있다. 초미세먼지와 휴일 여부는 그 영향이 적으며, 실제로 Tree 그림을 보면 분리 기준으로 사용되지 않았다.



Criterion : Entropy

Gini와 아주 유사한 Feature Importance를 얻는다. Feature Importance의 경우 최고기온, 일강수량, 합계일조시간, 초미세먼지, 휴일여부 순으로 높았다. 최고 기온이 따릉이 활성화 정도에 가장 큰 영향을 미치며, 그 정도가 다른 요인들에 비해 압도적임을 알 수 있다. 최고 기온이 우선적이고 자주 분리 기준으로 사용되었다는 사실을 다시 한 번 확인할 수 있다. 초미세먼지와 휴일 여부는 그 영향이 적으며, 실제로 Tree 그림을 보면 분리 기준으로 사용되지 않았다.

Body

Contents

Decision Tree

Regression DT

01. Regression DT Modeling

Regression의 경우 undersampling을 적용하지 않는다. 365건의 데이터를 8:2로 학습, 테스트 데이터로 분리했다.

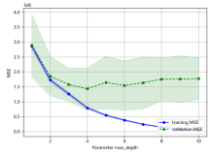
Regression DT의 Default Model 결과

R squared: 0.834
MSE: 133988984.685

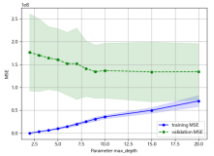
high R-squared value, high MSE인데, y값의 단위가 크기에 MSE가 크거나, default 모델은 데이터에 overfitting되었을 수 있다.

02. Optimization

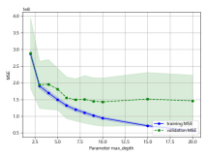
max_depth : [1,2,3,4,5,6,7,8,9,10] 비교
: train accuracy와 validation accuracy가 모두 적절한 성능 & 오버피팅 발생 직전인 max_dept = 4 선택



max_depth : [1,2,3,4,5,6,7,8,9,10, 15, 20]
: train accuracy와 validation accuracy가 모두 적절한 성능 & 오버피팅 발생 직전인 min_samples_split = 9 선택



max_leaf_nodes [1,2,3,4,5,6,7,8,9,10, 15, 20] 비교
: train accuracy와 validation accuracy가 모두 적절한 성능 & 오버피팅 발생 직전인 max_leaf_nodes = 6 선택



03. Grid Search

Scoring - neg_mean_squared_error

max_depth : [1,3,5,10,20,30]
min_samples_leaf : [3,5,10,20]
criterion : ['squared_error', 'friedman_mse', 'absolute_error', 'poisson']
min_samples_split : [3,5,10,20]
max_leaf_nodes : [3,5,10,20]

결과

Best score 0.8035760542003997
Best parameter {criterion': 'absolute_error', max_depth': 10, max_leaf_nodes': 20, min_samples_leaf': 3, min_samples_split': 3}

Scoring – r2

비교 파라미터 동일

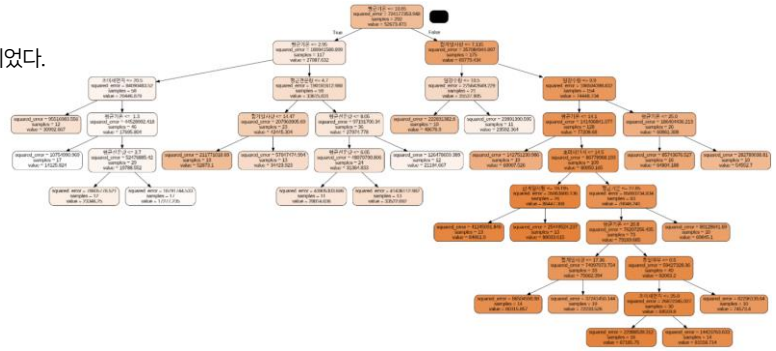
결과

Best score 127675667.03100717
Best parameter {criterion': 'squared_error', max_depth': 10, max_leaf_nodes': 20, min_samples_leaf': 5, min_samples_split': 3}

두 scoring 방식의 결과가 비슷함을 확인할 수 있다. 최종 모델에는 Mse 기준으로 한 최적의 파라미터를 활용했다.

최종 모델 : Default모델보다 r2, mse 모두 향상되었다.

R squared: 0.841
MSE: 128370013.228



Body / Conclusion

Contents

Decision Tree

Body. Regression DT

04. Regression DT 규칙 해석

가장 우선적으로 평균 기온이 10.65도를 기준으로 그 이하일 경우 전체적으로 따릉이 이용량이 적고, 이상일 경우 이용량이 많다.

특히, 평균기온이 2.95도보다도 낮을 때 이용량이 더 적은데, 초미세먼지 수치가 20.5 이하일 경우 약 31000건 정도로 그 중에서 그나마 이용량이 높다. 초미세먼지가 20.5보다 큰 경우, 평균기온이 -1.3도 이하일 때 이용량이 가장 적고, 평균 기온이 -1.3도보다는 높을 때 구름이 3.7보다 적으면 약 23000건, 크면 약 17000건이 발생했다.
→기온이 높고, 미세먼지가 적으며, 하늘이 맑은 날을 선호함을 알 수 있다.

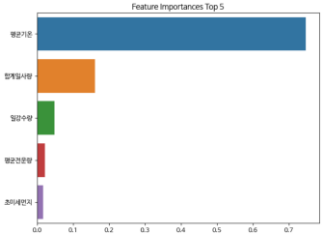
평균기온이 2.95도에서 10.65도 사이일 경우, 평균 전운량이 4.7보다 낮고 합계일사량이 14.47보다 낮을 때 약 52000건으로 이용량이 이 그룹에서 가장 많았다. 평균 전운량이 8.05보다 높은 경우 약 22000건으로 이 그룹에서 가장 적었다.
→구름이 너무 많거나, 합계 일사량이 너무 높은 경우는 선호하지 않음을 알 수 있다.

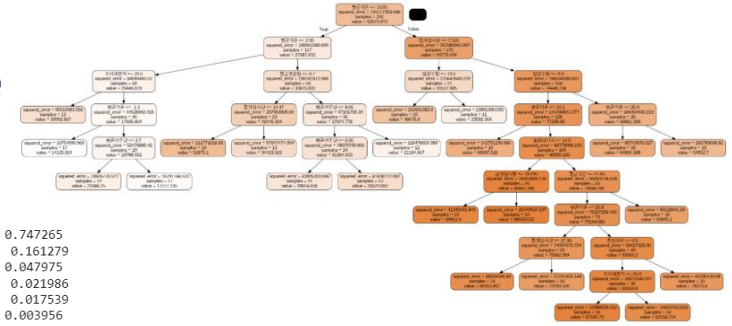
평균 기온이 10.65도보다 높은 경우, 합계 일사량이 7.115보다 낮은 경우는 일강수량에 따라 강수량이 10.5이하인 경우 이용량이 많고, 이상인 경우 적다. 합계 일사량이 7.115보다 높은 경우, 이 역시 일강수량이 0.9 이상인 경우보다 이하인 경우 이용량이 많다.

일강수량이 0.9 이상인 경우, 평균 기온이 25도보다 높은 경우보다 낮을 때 이용량이 많다.
일 강수량이 0.9 이하인 경우, 평균기온이 14.1도 이하인 경우보다 이상인 경우 이용량이 많다.

평균기온이 14.1도 이상인 경우, 초미세먼지 수치가 14.5보다 낮은 경우 이용량이 더 높은데, 이 때 합계일사량이 18.195보다 낮을 때 이용량이 약 4000건 정도 더 많다.
초미세먼지 수치가 14.5보다 높은 경우, 평균 기온이 27.85도보다 높을 때 이용량이 다소 감소한다. 27.85도보다 낮을 경우에는 20.8도보다도 낮은지에 따라 나뉘는데, 그 경우 합계일사량이 17.36보다 낮을 때 이용량이 약 80000건으로 약 72000인 높을 때보다 많다. 20.8도보다는 높은 경우, 휴일여부에 따라 평일이 더 이용량이 많으며, 초미세먼지 수치가 25보다 적을 때 높을 때보다 이용량이 적다.

05. Feature Importance





```
평균기온        0.747265
합계일사량       0.161279
일강수량        0.047975
평균전운량       0.021986
초미세먼지       0.017539
휴일여부        0.003956
dtype: float64
```

Feature Importance의 경우 평균기온, 합계일사량, 일강수량, 평균전운량, 초미세먼지 순으로 높았다. 평균 기온이 따릉이 이용량에 가장 큰 영향을 미치며, 그 정도가 다른 요인들에 비해 압도적임을 알 수 있다. Tree 그림 해석에서도 평균 기온이 우선적이고 자주 분리 기준으로 사용되었다는 사실을 다시 한 번 확인할 수 있다.

Conclusion

01. Result

Decision Tree를 통해 따릉이 이용의 활성 유무, 그리고 따릉이 이용자 수를 예측할 수 있다.

Classification DT

우선적으로 기온이 일정 수준 이상이며, 강수량이 적거나 강수량이 많더라도 합계 일조시간이 많은 경우 따릉이 이용이 활성화된다. gini와 entropy Criterion 모두 동일한 결과를 얻을 수 있다.

Regression DT

마찬가지로 기온이 가장 중요한 요인이며, 평균 기온이 높고, 미세먼지가 적으며, 하늘이 맑은 날에 따릉이 이용 건수가 많다. 또한 구름이 너무 많거나, 합계 일사량이 너무 높은 경우는 선호하지 않음을 알 수 있다. 추가로, 특정 상황에서 평일에 주말보다 이용량이 더 높다.

02. Limitation

Classification DT

- 일부 파라미터에 대한 Greed Search 결과, default 모델과 비교하여 성능이 크게 개선되지 않았다.

Regression DT

- 타겟값인 따릉이 이용량의 분포가 정규분포의 형태가 아니다.

03. Development

변수 선택을 VIF 기준으로 진행했으나, 그 이후에 주관적인 판단을 일부 개입하여 최종 변수 선택을 진행했는데, 보다 객관적인 기준을 적용할 수 있다.

Classification DT

- Test data가 70개로 적은 것에서 기인한 문제일 수 있으므로, data를 더 많이 확보한다면 greed search를 통한 성능 개선이 극명해질 수 있다. 또한, 튜닝을 진행하는 파라미터의 종류와 탐색 후보를 늘린다면 성능 개선이 이루어질 수 있다.

Data Source

[데이터 출처]

(1) 서울 열린 데이터 광장

- 서울시 일별 평균 대기오염도 정보 http://data.seoul.go.kr/dataList/OA-2218/S/1/datasetView.do#
- 서울시 공공자전거 이용현황 https://data.seoul.go.kr/dataList/OA-14994/F/1/datasetView.do

(2) 기상청

- 일별 종관기상관측(ASOS) 자료 https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36

(3) 2016 ~ 2019년 공휴일 데이터 https://superkts.com/day/holiday/2019