Practice-HW
Logistic Regression / Model Assessment

논문참조

강세정, 김규현, 신재욱, 장현우

Contents

Logistic Regression
/ Model Assessment

01

Introduction

Research Objective

Data

02

Body

Preprocessing & EDA

Logistic Regression

Evaluation

Result

03

Conclusion

Result

Limitaion

Development

Introduction

Contents

Logistic Regression
/ Model Assessment

01. Research Question

날씨 요인에 따라 따릉이 이용객 수가 평소보다 많은 지 여부를 예측해보자

02. Data

(1) RQ1

X(Numerical Variable)

평균기온 (°C)	최고기온 (°C)	일강수량 (mm)	최대풍속 (m/s)	평균풍속 (m/s)	최소상대 습도(%)	평균상대 습도(%)	가조시간 (hr)
합계일조 시간(hr)	1시간 최다일사량 (MJ/m2)	합계 일사량(MJ/m2)	일 최심적설(cm)	평균전운량(1/10)	9-9강수 (mm)	안개 계속 시간(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)

Y

일일 서울시 공공자전거 (따릉이) 대여 건수

「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~

(2) RQ2

X(Categorical Variable)		Y
미세먼지 기준 (1, 2, 3, 4)	초미세먼지 기준 (1, 2, 3, 4)	강수여부 (0, 1)

(3) 추가 설명

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

데이터 기간 : 2019.01.01 – 2019.12.31

Y 값은 대여 건수 평균값 이상 / 이하로 Binary 데이터로 변환함

→Binary Logistic Regression 수행 (해석 가능)

Body

Contents

Logistic Regression
/ Model Assessment

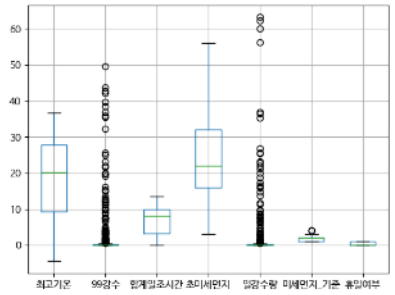
1. Preprocessing & EDA – Assumptions 1, 2, 4, 5 확인

1. 적절한 Outcome Type

Binary Logistic Regression에서, Y값이 binary 형태 : 적절

2. 관측치들 간 독립성: 데이터의 샘플들은 랜덤하게 얻어져야 함.

사용한 데이터는 시계열 데이터로, 해당 가정 만족 X



4. 이상치가 존재하지 않아야 함

Box-plot을 통해 이상치 제거.

단, 강수 관련 변수와 미세먼지 기준은 제거 X

5. 독립변수간 다중공선성 없어야 함

너무 많은 columns은 왜곡된 logistic의 결과 야기

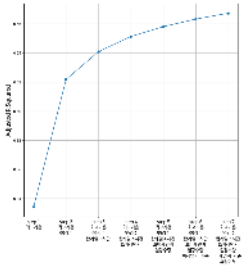
→forward variable selection을 통해 변수 일부 선정

[Numerical Variable 중 최종 사용 변수]

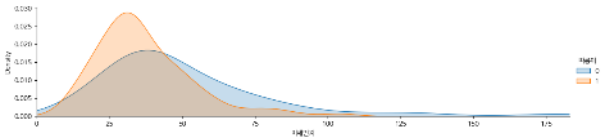
초미세먼지(㎍/㎥) / 최고기온(℃) /

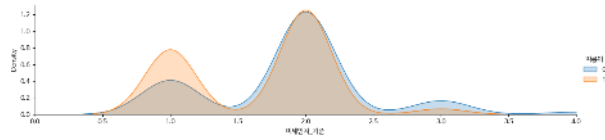
일강수량(mm) / 9-9강수(mm) / 합계일조시간(hr)

VIF는 모두 3을 크게 초과하지 않음 : 다중공선성 X



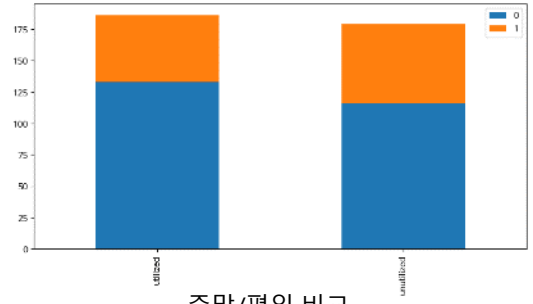
	variables	VIF
0	최고기온	3.608792
1	99강수	2.099618
2	합계일조시간	3.671065
3	초미세먼지	1.987403
4	일강수량	2.030607



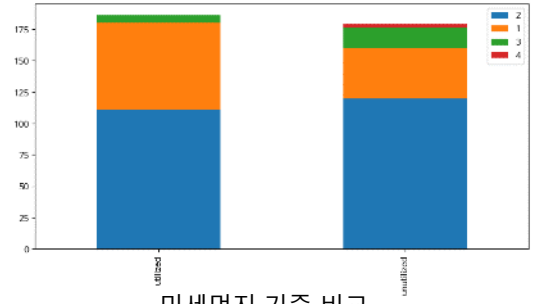


미세먼지 연속형 변수를 범주화하면, 성능이 다소 개선되는 것을 확인

추가로, 범주형 변수별 빈도 비교 결과 주말보다 평일에 따릉이 대여가 더 활발하며, 미세먼지는 1(매우 좋음)일 때 높은 대여, 3,4(나쁨, 매우 나쁨)일 때 낮은 대여 빈도가 확인됨. 2(좋음)일 때에는 별다른 차이 없음



주말/평일 비교



미세먼지 기준 비교

이 두 범주형 변수의 더미화를 통해 두 변수 간 correaltion =1이 되는 것을 방지

3. 독립 변수와 log odds 간의 linearity

6. 잔차간 등분산성이 존재해야 함

두 가정은 추후에 확인

02. Logistic Regression

Undersampling을 통해 비율 맞춤

0	112
1	112

```
Optimization terminated successfully.
         Current function value: 0.232170
         Iterations 33
```

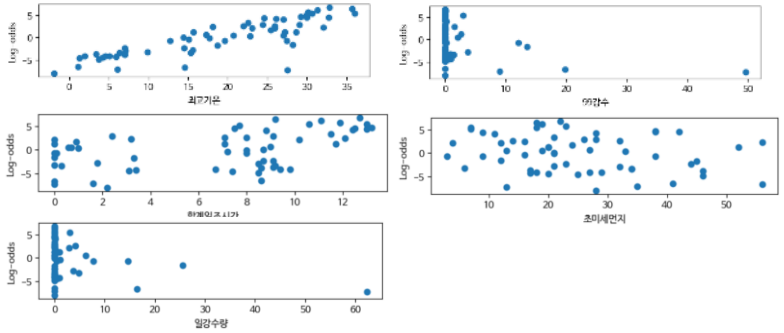
Body

Contents

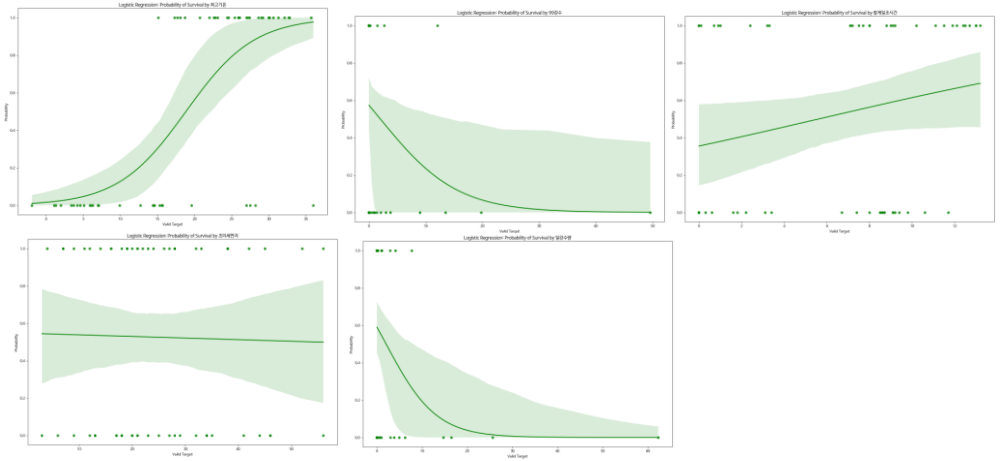
Logistic Regression
/ Model Assessment

03. Evaluation

(1) 가정 확인: **3. 독립 변수와 log odds 간의 linearity**



99강수, 초미세먼지, 일강수량 변수는 비선형 패턴을 보여주므로 로짓 선형성 가정이 위반됨



(2) 모델 평가

최적의 cut-off value 선정

0.1 ~ 0.9 비교 결과, 0.2 가 최적

```
0.2
[[22  6]
 [ 1 30]]
 Accuracy: 0.8813559322033898
Specificity: 0.7857142857142857
Precision: 0.8333333333333334
   Recall: 0.967741935483871
 F1 score: 0.8955223880597015
```

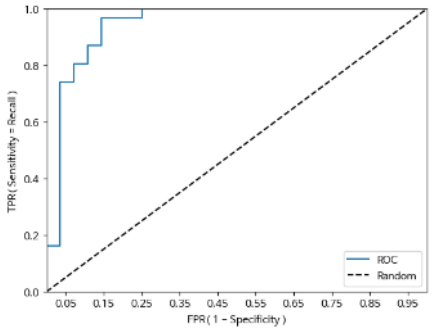
- 해석

22개의 샘플이 실제 Negative 클래스이고 예측도 Negative로 정확하게 예측
6개의 샘플이 실제 Negative 클래스이지만 예측은 Positive로 잘못 예측
1개의 샘플이 실제 Positive 클래스이지만 예측은 Negative로 잘못 예측
30개의 샘플이 실제 Positive 클래스이고 예측도 Positive로 정확하게 예측

이 모델의 정확도는 88.14%로, 대부분의 샘플을 정확하게 예측하고 있음
특이도는 78.57%로, 실제 Negative 클래스를 정확하게 예측한 비율이 높음
정밀도는 83.33%로, Positive로 예측한 샘플 중에서 실제 Positive 클래스인 비율이 높음
재현율은 96.77%로, 실제 Positive 클래스 중에서 정확하게 예측한 비율이 높음
F1 점수는 89.55%로, 정밀도와 재현율의 조화 평균이 높음

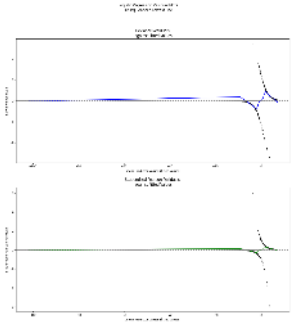
종합적으로, 전반적으로 높은 성능을 보임

ROC Curve 해석



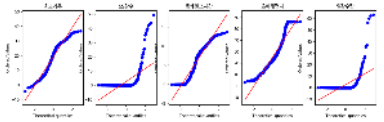
TPR (Sensitivity = Recall) 과 FPR (1- Specificity) 의 관계가
왼쪽과 같이 좌측 상단에 밀접한 그래프로 나옴
결과가 잘 나옴

(3) 가정 확인: **6. 잔차간 등분산성이 존재해야 함**



실선은 잔차의 추세를 나타내며, 실선이 y = 0 의 intercept를 가진 상태로 수평으로 근사되는 것으로 보아, 모델의 불완정성이 낮음
그러나, 잔차가 잔차가 무작위로 배치되지 않았으므로, 잔차의 등분산성이 만족하지 않음

추가로, qqplot과 bartlett 검정으로도 확인



bartlett 검정 결과,
p-value < 0.05 → 귀무가설 기각

Results

Contents

Logistic Regression / Model Assessment

Model.summary

Logit Regression Results

Dep. Variable:	따릉이	**No. Observations:**	224
Model:	Logit	**Df Residuals:**	214
Method:	MLE	**Df Model:**	9
Date:	Wed, 19 Apr 2023	**Pseudo R-squ.:**	0.6650
Time:	13:48:40	**Log-Likelihood:**	-52.006
converged:	True	**LL-Null:**	-155.26
Covariance Type:	nonrobust	**LLR p-value:**	1.423e-39

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.6014	1.238	-4.523	0.000	-8.029	-3.174
최고기온	0.3148	0.044	7.145	0.000	0.228	0.401
99강수	-0.1144	0.038	-3.030	0.002	-0.188	-0.040
합계일조시간	0.2207	0.079	2.800	0.005	0.066	0.375
초미세먼지	-0.0644	0.030	-2.124	0.034	-0.124	-0.005
일강수량	-0.0607	0.056	-1.090	0.276	-0.170	0.048
미세먼지_기준_2	0.6322	0.758	0.834	0.404	-0.854	2.118
미세먼지_기준_3	1.2346	1.473	0.838	0.402	-1.652	4.121
미세먼지_기준_4	-94.5226	9.32e+20	-1.01e-19	1.000	-1.83e+21	1.83e+21
휴일여부_1	-1.0285	0.581	-1.770	0.077	-2.168	0.111

Model.summary2

Model:	Logit	Pseudo R-squared:	0.665
Dependent Variable:	따릉이	AIC:	124.0123
Date:	2023-04-19 13:48	BIC:	158.1288
No. Observations:	224	Log-Likelihood:	-52.006
Df Model:	9	LL-Null:	-155.26
Df Residuals:	214	LLR p-value:	1.4230e-39
Converged:	1.0000	Scale:	1.0000
No. Iterations:	33.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-5.6014	1.2384	-4.5232	0.0000	-8.0285	-3.1742
최고기온	0.3148	0.0441	7.1453	0.0000	0.2285	0.4012
99강수	-0.1144	0.0378	-3.0297	0.0024	-0.1884	-0.0404
합계일조시간	0.2207	0.0788	2.7995	0.0051	0.0662	0.3752
초미세먼지	-0.0644	0.0303	-2.1239	0.0337	-0.1239	-0.0050
일강수량	-0.0607	0.0557	-1.0897	0.2758	-0.1698	0.0485
미세먼지_기준_2	0.6322	0.7582	0.8338	0.4044	-0.8539	2.1183
미세먼지_기준_3	1.2346	1.4729	0.8382	0.4019	-1.6522	4.1214
미세먼지_기준_4	-94.5226	93159643064628434432.0000	-0.0000	1.0000	-182589545219306605772 8.0000	182589545219306605772 8.0000
휴일여부_1	-1.0285	0.5812	-1.7695	0.0768	-2.1676	0.1107

Odds ratio(승산비)

```
const          3.692859e-03
최고기온          1.369998e+00
99강수          8.918823e-01
합계일조시간        1.246932e+00
초미세먼지         9.375850e-01
일강수량          9.411214e-01
미세먼지_기준_2     1.881733e+00
미세먼지_기준_3     3.437082e+00
미세먼지_기준_4     8.899113e-42
휴일여부_1         3.575596e-01
dtype: float64
```

미세먼지_기준_2 : 미세먼지_기준_1보다 odds가 1.88배 증가
미세먼지_기준_3 : 미세먼지_기준_1보다 odds가 3.44배 증가

휴일여부_1 : 휴일여부_0보다 odds가 0.357배 증가 (64.3% 감소)

Null deviance : P > |Z| : 0.05 미만이면 기각, x가 y를 설명하는 변수임을 의미

최고기온 / 99강수 / 합계일조시간 / 초미세먼지 : 귀무가설 기각, 따릉이 이용객 수를 설명하는 유의미한 변수

일강수량, 미세먼지_기준 (2,3,4), 휴일여부 : 귀무가설 기각할 수 없음, 따릉이 이용객 수를 설명한다고 할 수 없음

Logit(따릉이 = 1) = -5.6014 + 0.3148 * 최고기온 – 0.1144 * 99강수 + 0.2207 * 합계일조시간 – 0.0644 * 초미세먼지
- 0.0607 * 일강수량 + 0.6322 * 미세먼지_기준_2 + 1.2346 * 미세먼지_기준_3 – 94.5226 * 미세먼지_기준_4 – 1.0285 * 휴일여부_1

Conclusion

Contents

Logistic Regression
/ Model Assessment

결론

따릉이 이용객 수를 평균값을 기준으로 하여 많다 / 적다로 구분한 다음, 날씨 데이터를 이용한 logistic regression 모델로 따릉이 이용객 수를 예측함.

전체 데이터 중 5개의 숫자형 변수(최고기온, 9-9강수, 합계일조시간, 초미세먼지, 일강수량)과 2개의 범주형 변수(미세먼지_기준, 휴일여부) 를 전진 선택법을 활용하여 선택한 다음 VIF로 다중공선성을 확인함. 이 과정에서 이상치의 존재 여부를 확인했고 outlier가 Q3+1.5IQR 안으로 들어오도록 보정 진행.

모델 성능 평가 결과,
정확도 = 88.14%, 특이도 = 78.57%, 정밀도 = 83.33%, 재현율 = 96.77%, F1 score = 89.55%로 종합적으로 높은 성능을 보임.

한계

Logistic Regression을 위한 6가지 가정 중 일부 조건은 만족하였으나 충족되지 못한 조건들이 존재함.

- 날씨 데이터와 따릉이 데이터를 사용했는데, 이는 날짜에 따른 데이터이기 때문에 시계열 데이터임. 따라서 관측치들 간에 독립성이 성립하지 않음.
- 독립 변수와 log odds 간의 linearity를 확인하였을 때 최고기온처럼 선형성을 보이는 데이터도 있었으나 대부분 선형성을 보이지 않았음.
- 데이터들 간의 등분산성이 성립되지 않았음.

개선점

Data 관련 개선점

관측치의 독립성 : 시계열 데이터가 아닌 반복적 측정 X, 시간 및 공간에 따른 상관관계 X 데이터 활용을 통해 관측치들 간 독립성 확보 가능

변수와 log odds의 선형성 : polynomial terms 등 고차 다항식 항식을 통합하거나 연속형을 범주형으로 변환 통해 독립 변수의 선형 패턴을 확보함으로써 로짓 선형성 가정을 만족할 수 있음

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

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

(4) 2016 ~ 19년 공휴일 데이터
- https://superkts.com/day/holiday/2019