보험 청구료 예측을 통한 보험금 책정

목차

1. 프로젝트 소개

➔ 보험 가입자의 특성에 따른 보험 청구료 데이터를 분석하여 보험 가입자의 주요 특성에 따라 보험금을 책정 한다

2. 문제 정의

➔ 선형 회귀 분석을 통해 보험 청구료에 유의미한 영향을 미치는 특성들을 찾아낸다

➔ 보험 가입자의 주요 특성에 따라 다른 보험금을 책정한다

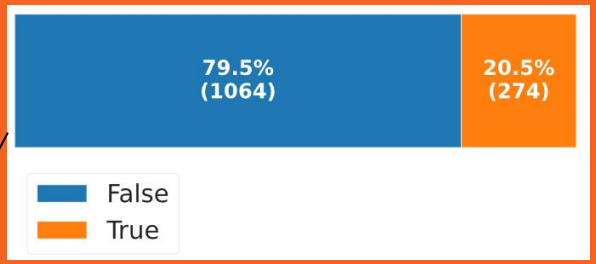
3. 데이터 소개

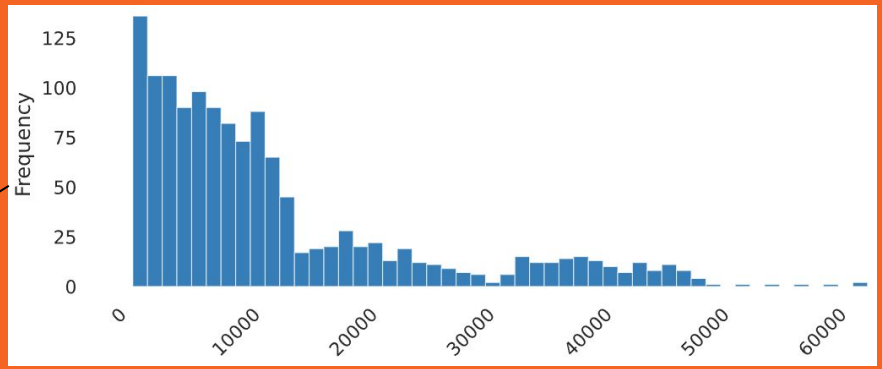
3.1) 컬럼별 설명

3.2) 전처리

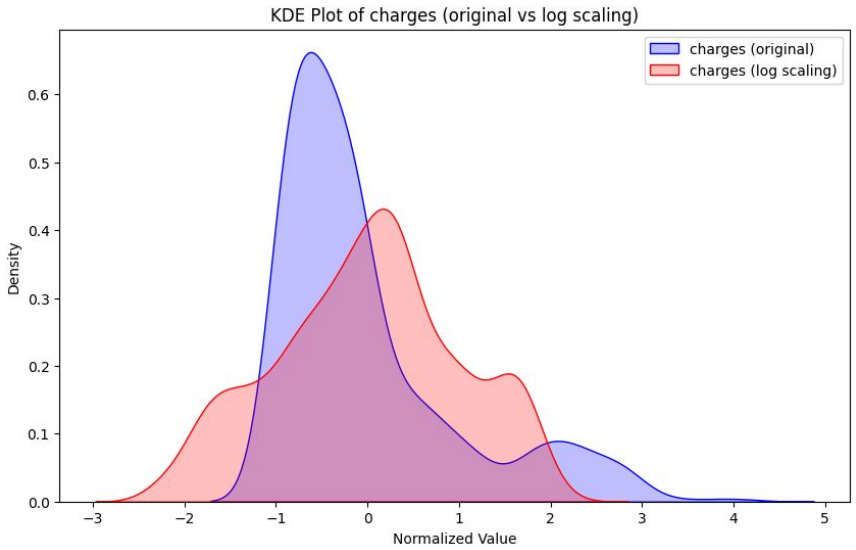
3.1) 컬럼별 설명

이름	타입	설명	특이사항
age	numeric	가입자 나이	균등분포
sex	categorical	가입자 성별	균등분포(2종류)
bmi	numeric	가입자 bmi	정규분포
children	categorical	부양 가족 수	X(6종류)
smoker	categorical	흡연 여부	**불균형 (2종류)**
region	categorical	거주 지역	균등분포(4종류)
charges	numeric	보험 청구료	**오른쪽 꼬리 분포**





3.2) 전처리




```
KS Test Statistic for charges: 0.0367
P-value for charges: 0.0536
The distribution of 'charges' is likely normal (fail to reject H0).
```

* α = 0.05

1. 결측치 X
2. 중복데이터 (2 rows)는 전산상 오류로 가정하여 삭제
3. 종속변수 로그 스케일링
4. 이상치 확인 (Z-score)

```
Z-score가 3을 넘는 값의 개수: 4
        bmi  bmi_zscore
116   49.06     3.016724
847   50.38     3.233182
1047  52.58     3.593945
1317  53.13     3.684136
```

```
Z-score가 3을 넘는 값의 개수: 0
Empty DataFrame
Columns: [charges, log_charge_zscore]
Index: []
```

3.2) 전처리

	age	bmi	children	charges	sex_male	smoker_yes	south	east
0	19	27.900	0	9.734236	False	True	1	0
1	18	33.770	1	7.453882	True	False	1	1
2	28	33.000	3	8.400763	True	False	1	1
3	33	22.705	0	9.998137	True	False	0	0
4	32	28.880	0	8.260455	True	False	0	0

5. 범주형 변수 인코딩 방식

이름	인코딩	추가 컬럼
sex	one-hot	-
children	label(=original)	-
smoker	one-hot	-
region	binary	south/east

4. 문제 해결

4.1) 상관계수 확인

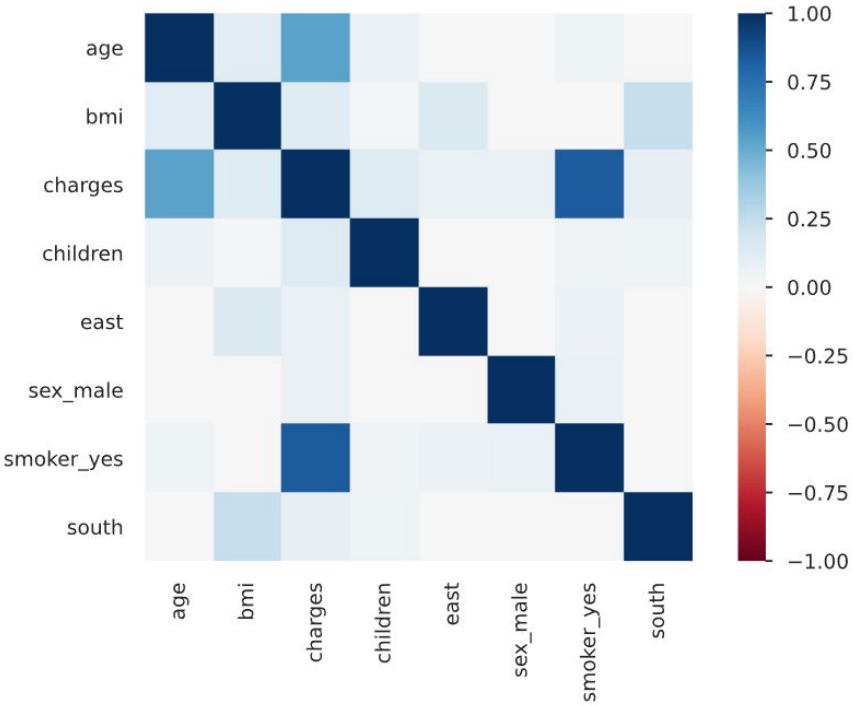
4.2) 범주에 따른 청구비용의 중앙값 확인

4.3) 파생변수의 범주 별 중앙값 확인

4.4) 나이에 따른 청구 비용의 변화

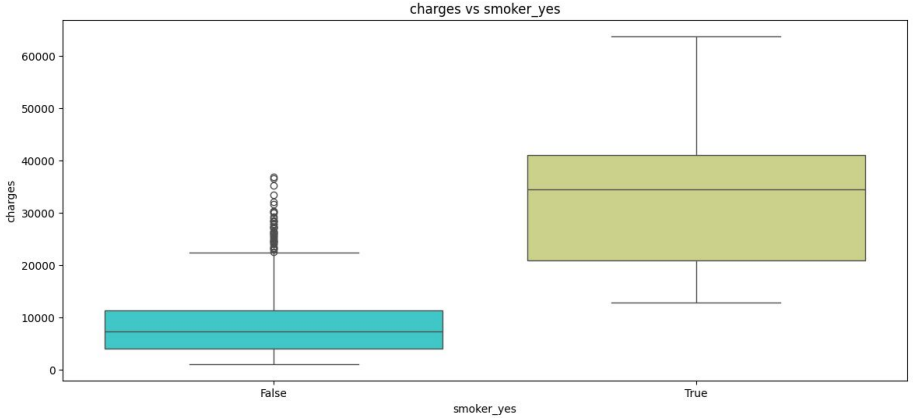
4.5) 군집화

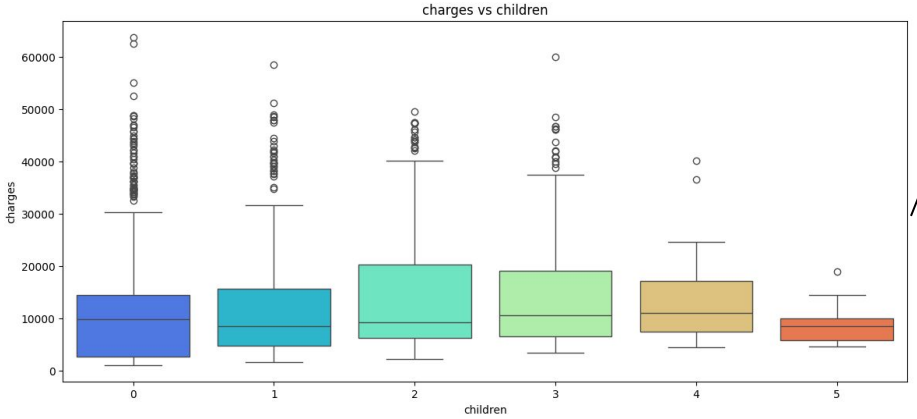
4.1) 상관계수 확인



1. 독립변수 사이에 강한 상관관계는 없음
2. 종속변수와 강한 상관관계를 갖는 독립변수는 age와 smoker_yes

4.2) 범주에 따른 청구비용의 중앙값 확인





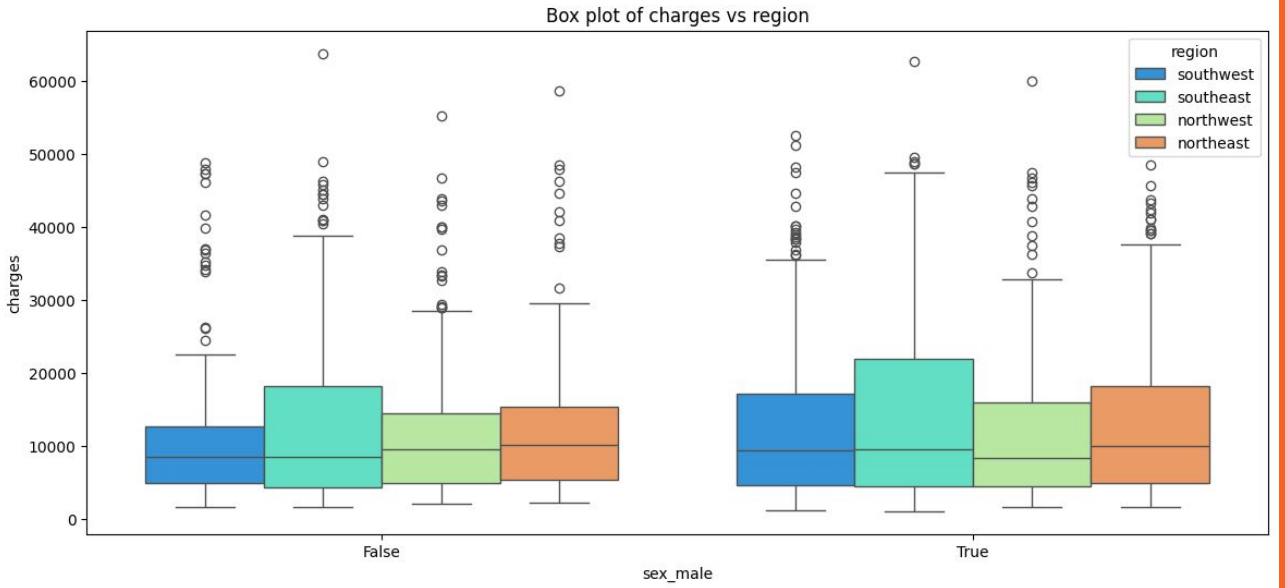
var	test	statistic	p-value	결과
sex	Mann-Whitney U	226198	0.6945	통계적으로 유의미하지 않음
smoker	Mann-Whitney U	283859	0	통계적으로 유의미
children	Kruskal-Wallis H	29.1207	0	통계적으로 유의미
region	Kruskal-Wallis H	4.6225	0.2016	통계적으로 유의미하지 않음

* α = 0.05

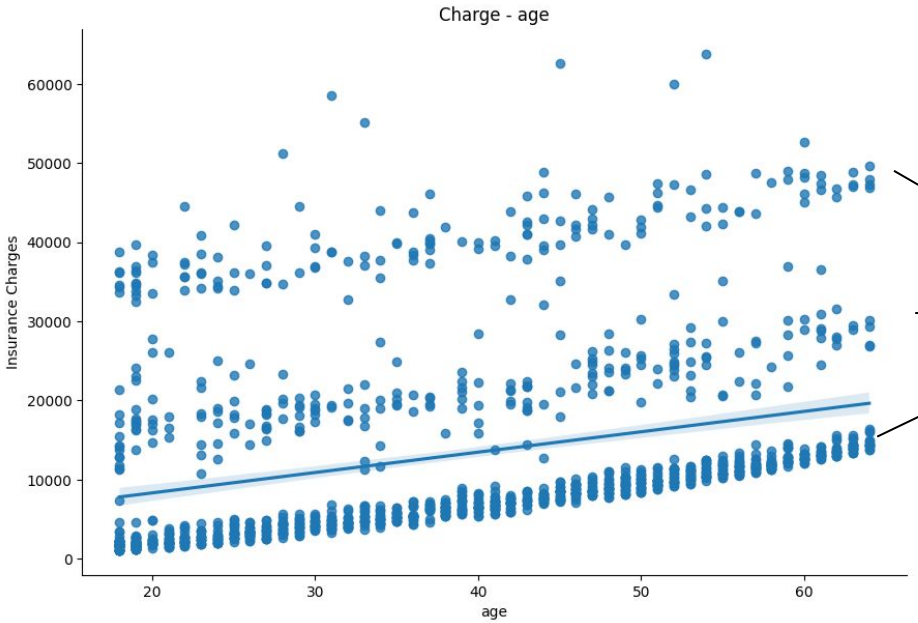
4.3) 파생변수의 범주 별 중앙값 확인

성별과 지역을 묶어 파생변수를 만들기
->
기각

```
H-statistic: 6.7290, p-value: 0.4576
카테고리(sex & region) 간의 의료비 차이는 통계적으로 유의미하지 않습니다. (귀무가설 채택)
```



4.4) 나이에 따른 청구 비용의 변화

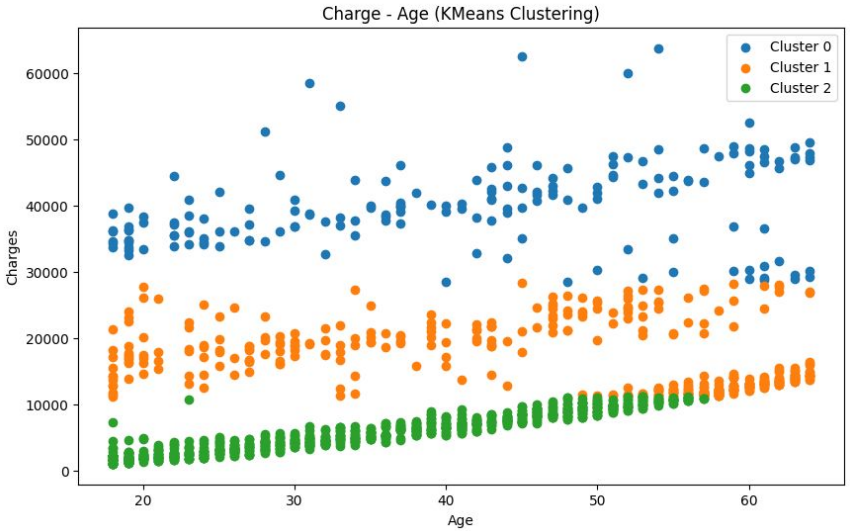


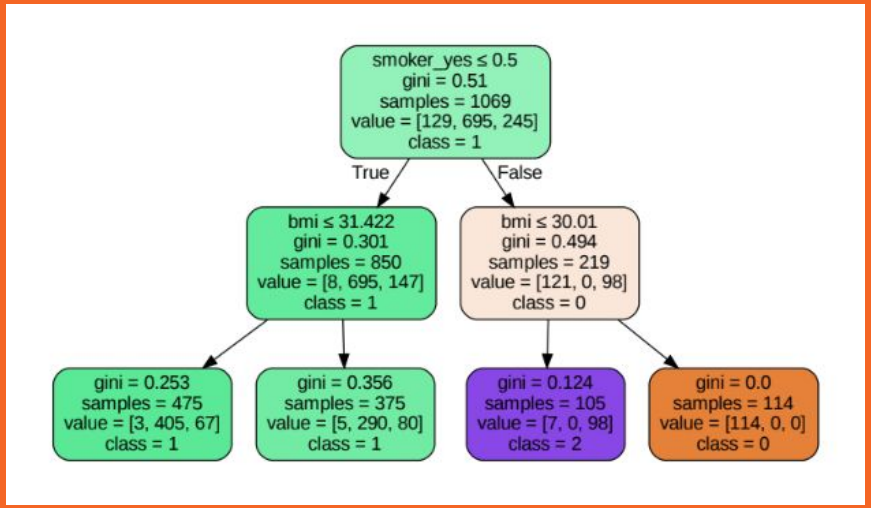
나이에 따라 청구 비용이 선형적으로 증가하는 세개의 군집을 확인 가능

4.5) 군집화

K-means Clustering
(Silhouette Score : 0.5933)

Decision Tree를 통해 군집 분류 조건 확인





4.5) 군집화

군집 분류 조건을 적용한 결과:

Cluster 0 : 흡연자, BMI > 30.01
Cluster 1 : 비흡연자
Cluster 2 : 흡연자, BMI <= 30.01

5. 모델 구축

5.1) 평가 방법 / 정규화 방식

5.2) 최종 회귀 직선

5.1) 평가 방법 / 정규화 방식

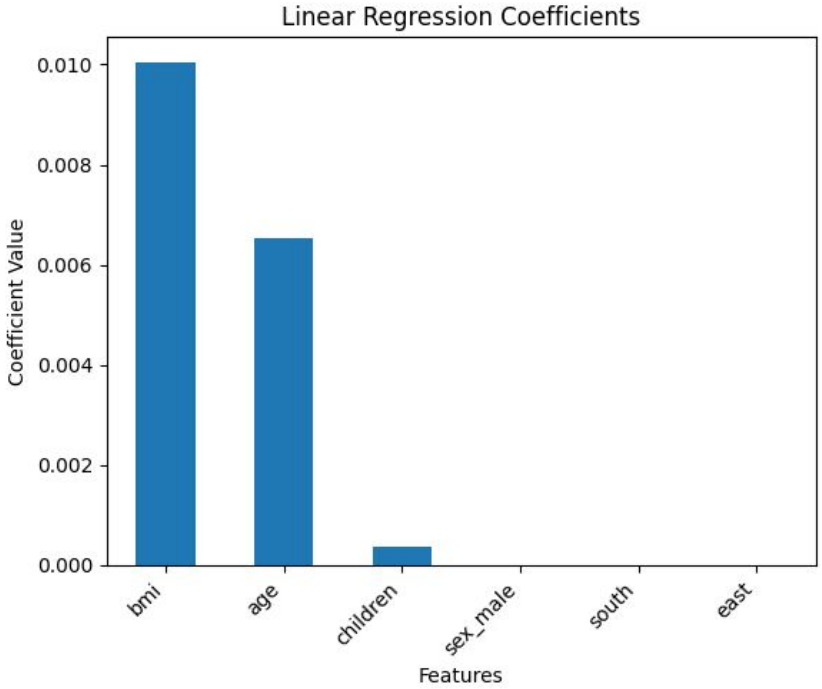
$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$\bar{X}_{all} = \frac{n_1X_1+n_2X_2+\ldots+n_kX_k}{n_1+n_2+\ldots+n_k} = \frac{\sum n_iX_i}{\sum n_i}$$

$$J(\theta) = MSE(\theta) + \alpha\sum_{i=1}^{n}|\theta_i|$$

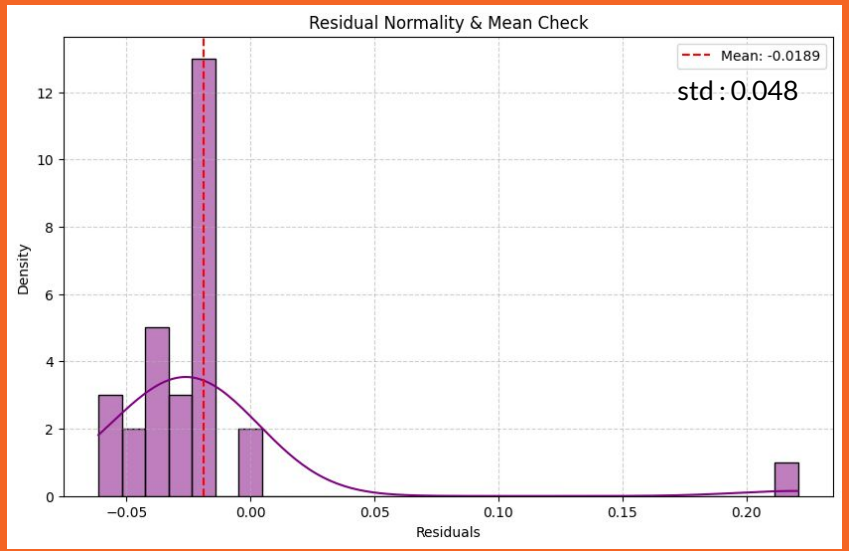
1. 군집별 회귀 직선에 대해서는 MSE를 사용
2. 가중 산술평균을 사용하여 전체를 평가
3. Lasso 회귀를 사용

5.2) 최종 회귀 직선

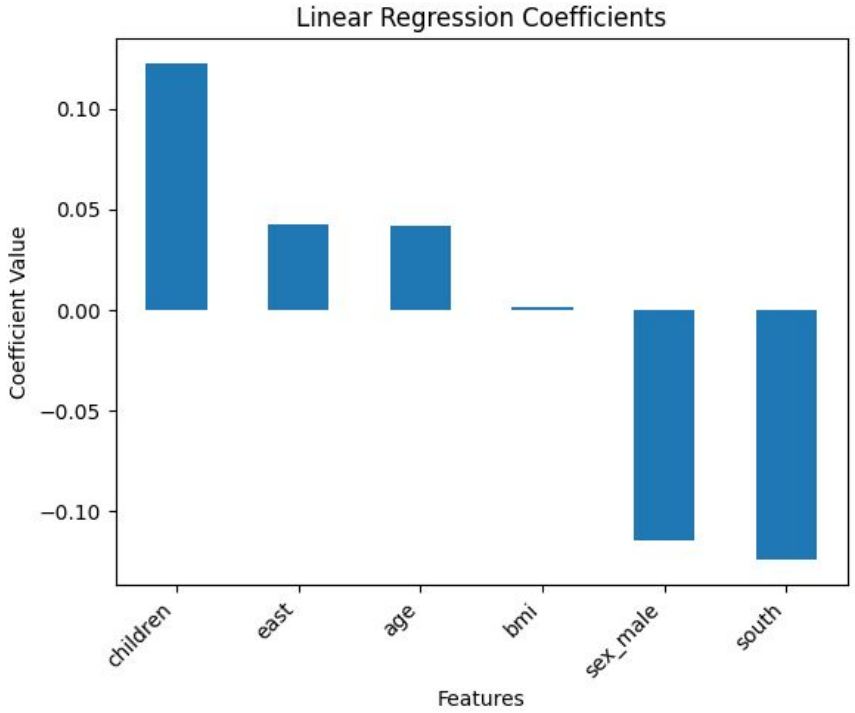


Cluster 0 : 흡연자, BMI > 30.01

- Count = 144
- R^2 = 0.8519
- MSE = 0.0027
- α = 0.01

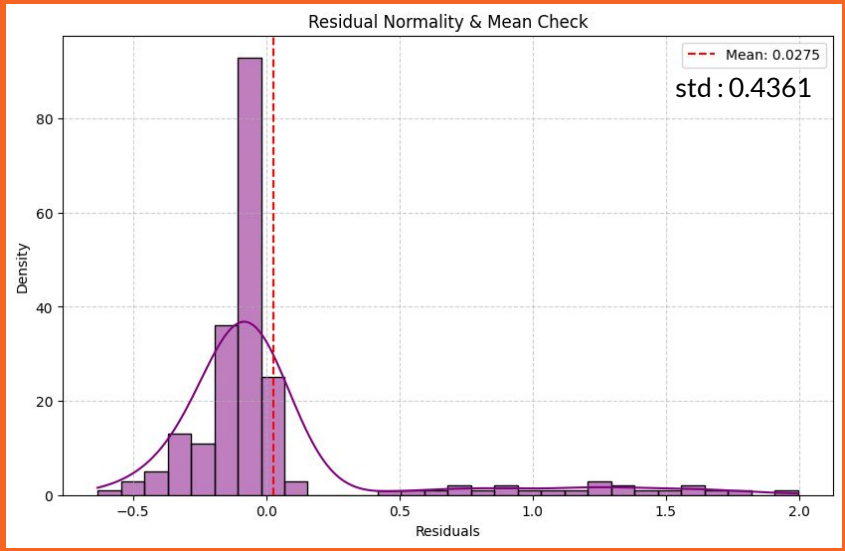


5.2) 최종 회귀 직선

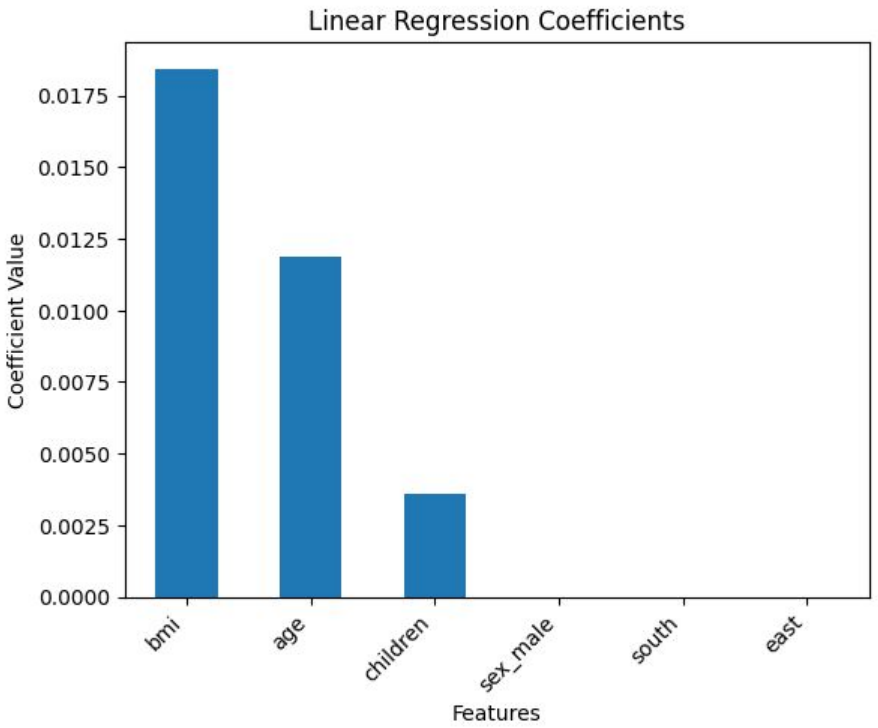


Cluster 1 : 비흡연자

- Count = 1063
- R^2 = 0.6560
- MSE = 0.1909
- α = 0.0

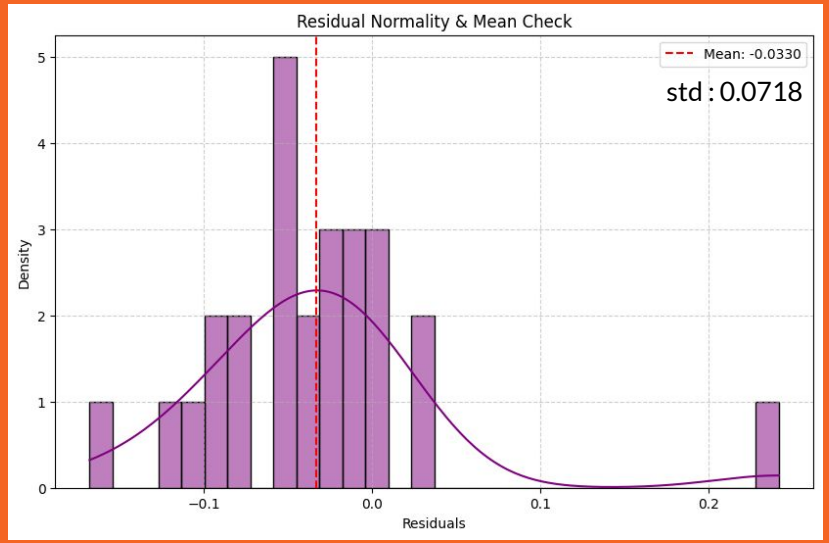


5.2) 최종 회귀 직선



Cluster 2 : 흡연자, BMI <= 30.01

- Count = 130
- R^2 = 0.8690
- MSE = 0.0062
- α = 0.01



5.2) 최종 회귀 직선

회귀 직선	r^2	mse	count
Cluster 0	0.8519	0.0027	144
Cluster 1	0.656	0.1909	1063
Cluster 2	0.869	0.0062	130
가중 평균	**0.698**	**0.153**	-

군집별 회귀 직선의 특이사항 정리

군집	주요 변수	예측 경향	예측 변동성
0	bmi, age	과대평가(소)	매우 작음
1	children, sex, region	과소평가(중)	매우 큼
2	bmi, age, children	과대평가(대)	작음

6. 보험금 책정

6.1) 보험 가입자 segment 유지

6.2) segment 별 보험금 책정

6.1) 보험 가입자 segment 유지

Cluster 0 : 흡연자, BMI > 30.01 → 건강 고위험군 (base : 22473$)

Cluster 2 : 흡연자, BMI <= 30.01 → 건강 위험군 (base : 8356$)

Cluster 1 : 비흡연자 → 일반 가입자 (base : 1172$)

6.2) segment 별 보험금 책정

건강 고위험군 (base : 22473$)

건강 위험군 (base : 8356$)

일반 가입자 (base : 1172$)

회귀식

base + (w1x1 + … + w1x3) + EBTI_1

base + (w1x1 + … + w1x3) + EBTI_2

base + (w1x1 + … w6x6) + EBTI_3

조건) EBTI_1 < EBTI_2 < EBTI_3

7. 한계점

7.1) 데이터 부족

7.2) 보험사의 입장을 충분히 반영하지 못한 평가 방법

감사합니다