분류

- 분류는 캐글 플레이그라운드 시즌 3의 에피소드 2 뇌졸중 예측 관련 내용 진행

- 3대 사망 원인 중 하나로 데이터들의 변수들을 기반으로 뇌졸중에 걸릴 확률을 예측하는 내용

- 뇌졸중은 뇌의 일부분에 혈액을 공급하고 있 혈관이 막히거나(뇌경색 = 허혈성 뇌졸중) 터짐(뇌출혈 = 출혈성 뇌졸중)으로써 그 부분의 뇌가 손상되어 나타나는 신경학적 증상입니다.

-

- 데이터를 알아보기 쉽도록 컬럼의 타입, 고유값 등을 한번에 확인할 수 있도록 했다

- 해당 자료에 결측값으로 표현되고 있는 부분들은 현재 문자형으로 최소값, 최대값, 평균값에 결측치로 확인되고 있으나 결측치는 없습니다.

- 해당 데이터의 컬럼을 수치형과 범주형으로 나눠 진행

- 범주형의 경우는 고유값의 개수에 따라 세분화함

EDA 분석

**1. 수치형**

- 1. 나이

- 이 데이터에는 데이터에 40대 ~ 60대의 연령대가 많음 /

- 2. 평균 당 수치

- 정상 수치 : 공복혈당(100 미만) / 식후 2시간(140 미만)

- 대부분의 관측치가 평균 포도당 수준이 60-120에 분포하고 있습니다.

- 차트 상 파란색이 아닌 따로 떨어져 있는 부분들 : 당뇨 환자의 경우는 당수치가 높을 수 있으므로 높은 수치들을 이상치로 보기는 어려움

- BMI(비만도)

- 기본 정보 : 18.5 ~ 22.9(정상) / 23.0 ~ 24.9 (과체중) /  25.0 ~ (비만)

- 차트 상 20 ~ 40 사이에 많이 분포하고 있음 / 높은 수치에 위치하고 있는 부분들을 이상치로 판단하기 어려움

Stripplot

1. 나이 : 20 ~ 60대 연령대에 많이 분포하고 있다 / 이 차트를 보면 뇌졸중은 고령에서 더 많이 발병된다는 것을 확인할 수 있습니다.

2. 당수치 : 뇌졸중과 평균 당수치를 산점도로 표현한 내용 / 뇌졸중 환자들은 당수치가 낮은 사람들도 있지만 높은 쪽에도 많이 분포되어 있다

3. 비만도 : BMI의 분포를 보면 뇌졸중 환자의 수치도 20 ~ 40 구간에 많이 분포되어 있음

오히려 뇌졸중이 아닌 사람들의 BMI 수치가 더 높고 나타나기도 함 / BMI 변별력을 확인하기는 어렵다

1. 수치형 데이터들 간의 관계를 표현해보았다

2. 가장 왼쪽에 차트는 나이와 bmi의 관계로 20 ~ 60대 나이대의 구간을 보면 연령이 높아질수록 비만도가 점점 높아진다는 것을 확인할 수 있다

3. 중앙의 차트는 나이와 평균 당수치 간의 관계로 나이가 많아질수록 평균 당수치도 높아진다는 것을 확인 할 수 있다

3. 마지막 차트는 평균 당수치와 비만도의 관계를 표현한 것인데 해당 차트에서는 비만도는 낮지만 평균 당수치도 낮아진다는 것을 확인할 수 있다

**2-1. 범주형(이진형)**

- 거주지

- 도시 vs 시골 / 도시와 시골의 수치가 거의 동일하다는 것을 확인 가능

- 뇌졸중에 거주지 형태는 변별력이 없다는 것을 볼 수 있음

- 결혼

- 데이터 중 미혼보다 기혼의 데이터가 많으며 뇌졸중 환자가 기혼자에서 더 많다는 것을 확인할 수 있다

- 심장질환

- 1. 0 이 심장 질환 X / 1 이 심장 질환 O

- 데이터는 심장 질환이 없는 사람이 더 많으나 심장 질환이 뇌졸중에 영향을 미친다는 것은 알 수 있다

- 고혈압

- 0 이 고혈압 X / 1 이 고혈압 O

- 데이터는 고혈압이 없는 사람이 더 많으나 고혈압이 뇌졸중에 영향을 미친다는 것은 알 수 있다

**2-2. 범주형(명목형)**

- 성별

- 데이터 수는 남성보다 여성이 많은 것으로 확인되나 교차분석표를 통해서 뇌졸중 환자와 아닌 사람의 비율을 보았을 때는 현재 남성 뇌졸중 환자가 더 많다는 것으로 확인할 수 있음

- 여성 중 약 3.89%, 남성 중 약 4.50%가 뇌졸중 환자로 확인됨

- 성별에서는 현재 other이라는 이상치가 발견되어 이상치 처리 요망

- 흡연 상태

- 1. 해당 차트 상 흡연이 뇌졸중에 미치는 영향을 확인하기에 명확하게 보이지 않으나 교차분석표를 확인해보면 흡연중과 이전에 흡연을 했던 사람들의 뇌졸중 비율이 높은 것으로 확인할 수 있다

==> 흡연이 뇌졸중에 영향이 미친다는 것을 확인해볼 수 있다

- 하지만, 흡연과 비흡연의 더 명확한 확인을 위해서 smokes와 never smoked로 재 분류하여 흡연과의 상관 관계 확인해볼 수 있을 것 같다

- unknown을 never\_smoked로 합친 이유 -> 데이터의 개수는 많지만 뇌졸중 환자의 비율은 적어 담배를 피지 않는 never\_smoked로 가정하여 진행한다

never\_smoked : never\_smoked, unknown

smokes : smokes, formerly\_smoked

- 직업

-  뇌졸중 환자 중 일을 해본 사람(Private, Govt\_job, Self-employed)의 비중이 높음

- 직장인(Private, Govt\_job) vs 자영업(Self\_employed) vs 일하지 않는 사람 (Never\_worked) 으로 특증을 재 분류해서 검토해볼 수 있을 것 같음

**3. 히트맵**

**-** Age, hypertension, avg\_glucose\_level, heart\_disease, bmi 변수는 대체적으로 상관관계가 있음

**4. 피처엔지니어링**

- EDA 분석을 통해서 거주지와 비만도는 뇌졸중을 예측하는데 변별력이 많이 떨어진다는 것을 확인했고 예측 정확도를 높이기 위해 해당 두 부분을 삭제 후, 진행

- 성별

- other은 뇌졸중 환자가 아님을 확인 / 데이터가 여성의 수가 많기 때문에 이상치 값을 남성으로 대체

- 흡연

- 흡연이 뇌졸중에 미치는 영향을 확인하기 위해서 비흡연과 흡연의 이진형 형태로 변환

흡연자 중 뇌졸중 환자가 많다는 것을 확인 가능

- 직업

- 자영업 vs 직장인 vs 직업 無 의 형태로 변경

- 뇌졸중은 직업 유무에 따라서 발병 여부가 다르다는 것을 비교할 수 있다 => 직업이 뇌졸중의 발병에 영향을 미친다

- 자영업 - 직장인 - 직업 無의 순으로 뇌졸중의 여부나 높음

- 인코딩

- 문자형으로 이루어진 데이터들을 수치형으로 변환 / age와 avg\_glucose\_level은 float 형태로 int 형태로 변환

- 스케일링

- 수치형 데이터의 범위가 다르기 때문에 범위를 맞춰주기 위해서 minmaxscaler를 사용했다

(사용한 이유는 다른 스케일링을 사용했을 때보다 민맥스를 사용했을 때 결과가 가장 좋았음)

**5. 베이스라인구축**

- 베이스모델은 로지스틱 회귀와 그리드 서치를 이용

- 풀데이터를 이용하여 베이스 모델 훈련을 진행

- 결과는 0.89342의 결과가 나왔다

**6. 모델비교**

- 성능 높이기 위해서 여러 모델들을 비교 진행했음

- 결과 상 ada부스트가 가장 좋은 예측 결과를 냈지만 XGBoost를 사용함

- 이유는 파라미터를 변경했을 때 xgb가 가장 좋은 결과를 냈다

- 이렇게 진행해서 이런 결과를 도출

**7. XGB**

- 그리드서치를 이용하여 xgb 최적의 파라미터를 찾았고 이를 이용하여 데이터를 훈련한 결과가 roc 곡선으로 시각화 진행

- 이걸로 풀데이터 훈련을 진행해서

**6. 제출**