<슬라이드 1>

안녕하십니까, 심장마비의 원인 분석 이란 주제로 발표를 하게 된 201조의 000, 000 입니다.

<슬라이드 2>

시작하기에 앞서, 저희가 사용한 데이터의 출처에 대해 말씀드리겠습니다.

저희는 Keggle에서 데이터를 가져왔습니다. 하지만, 다행히, 개시자께서 해당 데이터의 Raw데이터를 올려 둔 덕분에 데이터의 전처리를 수행하는 과정을 수행할 수 있었습니다.

<슬라이드 3>

데이터는 여러 요인들에 대한 값과, 심장마비의 유무를 포함한 2180x14크기의 csv파일입니다.

<슬라이드 4>

해당 데이터에 대한 설명을 하겠습니다.

우선 age는 대상의 연령을 나타냅니다.

Sex 성별, cp는 흉곽 통증의 유형이며, 0 = 전형적인 협심증, 1 = 비정형 협심증, 2 = 비각통, 3 = 무증상을 의미합니다. Trestbps는 휴식 상태 혈압, chol: 혈청 콜레스테롤 수치, fbs: 공복 혈당이 120 mg/dl 보다 큰지. restecg: 휴식 심전도 결과를 나타내며, 0 = 정상, 1 = ST-T파 이상, 2 = 좌심실 비대를 나타냅니다. 이때 ST-T파는 그림에서 알 수 있듯이, 보라색~T에 해당합니다. Thalach는 최대 심박수, exang는 운동으로 인한 협심증 유무, oldpeak는 운동으로 유도된 ST의 감소입니다. Slope는 최고 운동 ST 세그먼트의 기울기이며, 0 = 상향 기울기, 1 = 평평함, 2 = 하향 기울기를 의미합니다. ca값은 형광 투시를 통해 색이 칠해진 주요 혈관 수의 값이며 0, 1, 2, 3이 있습니다. thal: 지중해 빈혈 유형, 즉 헤모글로빈 결핍으로 인한 빈혈의 유무이며, 1 = 정상, 2 = 고정 결함, 3 = 가역적 결함을 의미합니다.

마지막으로 target은 반응 변수이며, 1 = 심장마비, 0 = 안전을 의미합니다.

<슬라이드 5>

그럼 다음으로 영가설과 대립가설을 설정하겠습니다.

영가설은 저희가 방금 말한 변수들이 심장 마비의 발생과는 관련이 없다. 이며, 대립가설은 해당 변수들이 심장 마비의 발생에 유의미한 영향을 끼친다는 것 입니다.

또한, 해당 내용은 심장마비, 즉 사람의 생명과 관련된 내용이므로 유의수준을 5%가 아닌 3%로 설정하였습니다.

( => 오버피팅 문제 어케 해결할거냐 라고 나오면 k-fold든 random forest든… 말하기)

<슬라이드 6>

다음은 데이터 가공 계획입니다. 우선 결측값과 이상값을 제거 후,

각 파라메터의 상관관계 부석 후, 설명변수를 설정합니다.

이후로 학습중에는 Logistic, RandomForest, SVM 사용할 예정이며, 이후 전체 데이터를 바탕으로 학습한 모델과 선택한 변수를 바탕으로 한 모델을 confusion matrix로 비교할 예정입니다.

<슬라이드 7>

데이터를 보시면 아시다시피, ?표기 혹ㅇ은 2.5 ? 와 같이 불확실한 정보들이 존재하는 것을 확인하였으며, gsub을 사용하여 ?이 포함된 문자열을 모두 NA처리 후, 다시 모든 데이터를 numeric하게 바꿔주었습니다.

이후 na.omit을 이용해 결측값을 제거해 주었습니다. 평균을 낼 수도 있으나, 이렇게 진행한 이유는 13개의 설명변수에 대해 2000개가 넘는 data는 충분히 크다 생각했기 때문입니다.

<슬라이드 8>

다음으로는 이상값의 유무를 확인하기 위해 boxplot을 사용해 모든 데이터와 연속형 데이터에 대해 그래프를 그려보았고, 이상값 처리를 하는 것이 맞다는 판단을 내리게 되었습니다.

<슬라이드 9>

IQR로 구한 범위를 연속형 데이터에 적용시켜 해당 값을 NA로 바꾼 후 다시 전체 데이터에 대해 na.omit을 수행하여 이상값을 제거하였으며, 이 과정을 통해 2181개였던 데이터가 1769개까지 줄어들게 되었습니다.

<슬라이드 10>

다음은 상관관계 분석입니다.

우선 corrr과 performance Analytics를 통해 각 파라메터간의 상관관계를 구하였습니다.

적색이 양의 상관관계, 청색이 음의 상관관계를 나타냅니다.

<슬라이드 11>

해당 내용을 바탕으로 보기 쉽게 정리한 target과 다른 파라메터간의 상관관계와 그 정도를 표시하였습니다.

<슬라이드 12>

이후, 연속형 데이터에 대해서만 다시 상관관계 분석을 수행하였습니다.

범주형 데이터의 경우 우선 학습에 사용한 후, 추후 모델 학습 과정 중, Backward방식을 통해 제거할 예정입니다.

<슬라이드 13>

해당 내용은 연속형 파라메터의 corrr과 performance Analytics분석결과이며, 내용은 이전 슬라이드와 같습니다.

<슬라이드 14>

해당 내용을 바탕으로 age, thalach, oldpeak를 설명변수로 선택하였으며, 기준은 0.15입니다.

<슬라이드 15>

저희가 사용할 설명변수는 age, thalach, oldpeak ,Sex, cp, exang, slope, ca

그리고 반응변수는 target입니다.

<슬라이드 16>

해당 모델로 예측하고자 하는 것은 병원, 혹은 설명변수의 습득이 가능한 환경에서 이용자의 데이터를 입력 받으면, 해당 모델을 통해 심장마비의 위험도를 출력하고, 이때의 근거를 제공하여 심장마비를 사전에 방지하고자 합니다.