|  |  |
| --- | --- |
| 시작 | 안녕하세요, 2조 발표 시작하겠습니다.  저희 조의 발표 주제는 **건강검진 데이터를 이용한 흡연 여부 예측**입니다. |
| 2 | 발표 순서는 다음과 같이 진행하겠습니다. |
| 3 | 연구 배경 및 목적입니다.  흡연이 인체에 미치는 위험성은 꾸준히 제기되고 있으며 흡연 여부는 국민의 건강 수준을 조사하는 데에 필요한 중요한 자료입니다.  흡연 여부는 건강 검진 시행 시에 문진표 작성, 니코틴 측정 등의 방법으로 판단하기는 하지만 여러 어려움이 존재합니다, 문진 항목의 경우 항목에 거짓 응답을 하거나 응답을 하지 않는 경우가 존재하고 니코틴을 측정하여 검사를 하려면 비용이 들기 때문에 이러한 어려움들을 보완하는 데에 저희의 흡연여부 예측 모델이 도움이 될 것이라 생각했습니다. |
| 4 | 데이터 설명입니다.  먼저 기존 자료 조사 때, 캐글에서 얻은 자료는 조작된 흔적이 발견되어 뒤늦게 공공데이터포털에서 가져온 건강검진보험공단의 원자료를 찾아 진행하였습니다.  결과적으로 신체 정보, 콜레스테롤, 간수치, 혈압 등의 건강검진 변수와 타겟변수 흡연을 포함한 여러 문진 항목 변수를 갖고 분석을 진행했습니다. |
| 5 | 데이터 확인 및 전처리 과정입니다.  건강검진 데이터 특성상, 건강검진 방식의 변화와 문진 항목의 변화로 인해 매우 많은 결측치가 존재했으며, 꼼꼼한 정제를 통해 이용 가치가 있는 데이터만을 활용할 필요가 있다고 생각했습니다.  이에 100만 행 중 66만 행 이상이 결측치인 범주형 변수들을 제거했으며, 분석에 사용하지 않을 변수들도 함께 제거하였습니다. |
| 6 | 이후, 콜레스테롤 정보를 기준으로 결측치인 행들을 최소한으로 제거하고,  타겟변수인 흡연정보가 결측인 행, 대부분의 변수가 결측인 33개의 행을 제거해주어 1차적인 결측치 처리를 마치고, 40만개의 행으로 데이터 분석을 진행했습니다. |
| 7 | 다음은 범주형 변수 전처리 과정입니다.  시각화를 통해 비흡연 외의 나머지 두 범주가 같은 경향을 띈다고 판단했으며, 서울대학교 건강검진센터의 연구결과를 참고하여 흡연 변수의 범주를 비흡연, 과거에는 흡연했으나 현재는 비흡연, 현재 흡연 중이라는 세가지 범주에서 두가지로 재범주화했습니다. |
| 8 | 연령대는 식별가능한 숫자로 바꿔주었습니다. |
| 9 | 요단백 변수는 정상 수치에 해당하는 1번 범주의 수가 지나치게 많은 것을 고려하여, 두 가지 범주로 재범주화하였습니다.  이 과정에서 재범주화된 요단백 변수가 타겟변수인 흡연에 독립적인지를 확인하는 카이스퀘어 검정을 시행했고, 흡연 여부를 예측하는 데에 도움이 되도록 재범주화가 됐다고 판단했습니다. |
| 10 | 이 밖에도 논리적 오류인 이상치 값들을 처리하였고,  의학데이터 특성을 고려하여 우측으로 긴 꼬리 분포들을 갖는 수치형 변수들의 값을 퍼센타일과 도메인 정보를 동시에 고려하여 처리해주었습니다. |
| 11 | 다음은 수치형 변수 중 콜레스테롤 수치 변수들 4가지입니다.  앞서 ldl을 제외한 나머지 세가지 변수를 기준으로 결측행을 모두 지웠기 때문에 ldl만이 결측행이 존재합니다.  화면과 같이, 트리글리세라이드가 400이하이면 프리드워드 방정식을 통해 ldl 결측을 대체하고,  두 번째 조건이면 샘슨 방정식을 적용하는 식으로 값을 대체했습니다. 대체 후에, 추정불가능한 값은 결측치로 두고, 추정된 값이 0보다 작은 행은 이론에 부합하지 않으므로 문제가 있다고 판단하여 행을 삭제하였습니다. |
| 12 | 흡연 여부 예측 모델에 사용될 변수를 선택하기 위해서 타겟변수인 흡연에 따른 각 변수들의 차이를 검정했습니다.  먼저 정규성 여부를 확인한 후, 등분산 검정을 진행했고, 이분산이라는 결과를 얻었습니다. 이후 이분산 Ttest를 양측검정방식으로 진행하여 차이를 확인하고 단측검정을 다시 진행하여 대소관계까지 파악했습니다.  이 과정을 통해 흡연집단이 비흡연 집단보다 큰 수치를 갖고 있음을 확인했고 흡연과의 관련성을 파악했습니다. , |
| 13 | 간수치 변수 역시, 도메인 정보를 바탕으로 1%의 극단값을 제거했고 정규성을 따르지 않으므로 모수적 검정이 아닌 비모수검정 맨휘트니 U test를 진행했습니다. 이를 통해 흡연집단 간의 간수치 평균 값 차이가 있음을 확인했습니다. |
| 14 | 1차적인 결측치 처리와 이상치 제거까지 마친 다음, 나머지 결측 범주형 변수는 결측치의 수가 매우 적으므로 단순결측치대체법 중 최빈값 대체를 이용했고, 수치형 변수는 다중대체법 MICE 알고리즘을 통해 모든 결측치를 처리했습니다. |
| 15 | 이후에도 주어진 변수들의 정보가 내재된 파생변수를 생성하여 흡연과의 관계가 두드러지는 변수들을 모델링에 반영하고자 했습니다. |
| 16 | 모델링 방식은 대용량 처리에 장점이 있는 머신러닝 분류모델 랜덤포레스트와 Xgboost를 사용했습니다.  이 때 데이터가 워낙 크기 때문에 랜덤서치나 그리드서치보다는 베이지안최적화기법을 활용하여 효율적인 하이퍼파라미터 튜닝을 했습니다.  여기서 베이지안 최적화란 베이즈 정리를 기반으로 하여 특정 구간 내에서 모델의 성능을 최대로하는 하이퍼파라미터를 찾아주는 알고리즘입니다. |
| 17 | 모델 튜닝과정에서 신장, 체중과 같은 변수 대신 bmi를 넣어주는 등 변수중요도가 낮거나 의미가 겹치는 변수들을 줄여주면서, 모델의 성능 변화를 확인했습니다. 그 결과 성능은 유지하면서 변수가 줄어들어 해석이 더욱 용이해졌습니다. |
| 18 | 랜덤포레스트의 경우, 성별과 간수치gtp 다음으로 콜레스테롤 3가지 변수가 변수중요도가 높았고,  정확도와 F1스코어는 각각 0.82, 0.77이었습니다. |
| 19 | xgboost에서는 변수중요도가 거의 비슷하게 형성되었으나 연령대가 높은 순위에 추가된 것이 특징이었고 성별의 영향력이 너무 크게 모델에 반영되어있기 때문에 성능이 거의 동일하게 나타났음에도 해석이 용이하고 변수중요도가 고르게 적용된 랜덤포레스트를 최좋모델로 선정했습니다. |
| 20 | 마지막으로 ㄴㅁㅇㅁㄴㅇ테스트 데이터 셋을 예측한 결과는 다음과 같으며 준수한 성능을 보였습니다. |
| 21 | 최종 결론입니다.  성별은 흡연을 예측하는 데에 특히 큰 영향을 주었으며, 이는 여성의 흡연자 비율이 남성에 비해 낮은 수치를 보이고 있기 때문입니다. 실제로 눈에 띄게 여성의 흡연율이 낮은 것인지, 그렇다면 여러 이유가 있는지 사후 점검을 하여 본 프로젝트의 결과를 보완할 필요가 있어 보입니다.  음주를 하는 사람은 동시에 흡연을 하는 경향이 있었고, 종합적인 수치를 함께 고려해볼 때, 건강관리에도 더욱 소홀하다고 볼 수 있었습니다. 따라서 특정집단에게는 건강이상 확률이 더 높으므로 건강관리의 필요성을 전달하는 데에 도움이 될 수 있습니다.  모델의 변수중요도에서 나타나듯, 흡연자의 경우 특히 콜레스테롤 수치와 간수치등의 문제 역시 두드러지기 때문에 추가적인 간기능검사나 신장기능검사를 권유해볼 수 있습니다.  마지막으로 저희의 모델을 확장시킨다면, 건강검진 시 흡연자일 확률이 높은 집단에게는 추가적인 검진 또는 문진을 유도하여 해당 집단이 흡연자이기에 발병가능성이 높은 합병증이나 질병을 사전에 발견하도록 도움을 줄 수 있습니다. (끝) |