Codestates AI bootcamp 15th 정경재

Section2 Project Script

한국형 고혈압 예측 모델 개발

Hypertension Predictive Model

1.

안녕하세요. 코드스테이츠 AI 15기 정경재입니다.

프로젝트 발표 시작하도록 하겠습니다.

2.

시간관계상 빠르게 진행하겠습니다.

3.

4.

고혈압은 침묵의 살인자라고도 불리우는 질환입니다.

사망 위험요인 중 1위를 차지하는 만성질환이며

자료에 따르면 고혈압 환자의 95%는 뚜렷한 원인을 밝힐 수 없는 본태성 고혈압으로 분류됩니다.

5.

고혈압의 진단기준은 오랜기간 동안 수축기 140, 이완기 90 이라는 수치로 진단되어져 왔습니다.

하지만 2017년 미국에서 진단기준을 130, 80으로 강화하였고, 유럽과 우리나라는 진단기준을 유지하는 보수적인 입장을 취해오고 있습니다.

6.

95%가 뚜렷한 원인을 알 수 없는 본태성 고혈압이기 때문에 위험요인을 예측하고 분석하는 것은 매우 중요한 문제이고,

또한 각각의 진단기준에 따른 예측 모델을 개발하고 비교분석을 해볼 필요가 있습니다.

7.

따라서 다음과 같은 세가지의 목표 및 가설을 정의하였습니다.

고혈압 예측모델을 개발하고,

진단기준별로 비교분석 하고,

모델 해석으로 특성 중요도를 파악하는 것입니다.

8.

이번 프로젝트에서 사용한 데이터는 국민건강영양조사 8기 1,2차년도 원시자료입니다.

9.

이번 주제는 고혈압 진단에 대한 문제이기 때문에 분류문제로 정의하였고,

타겟은 고혈압 여부입니다.

10.

대한고혈압학회에서 정의한 위험요인들을 위주로 특성을 정의하여 진행했습니다.

11.

전처리 과정입니다.

12.

13.

우선 필요한 컬럼만 추출하고, 2019년과 2020년 데이터를 병합했습니다.

결측치는 제거하는 방법으로 처리하였습니다.

14.

다음으로 다음과 같은 기준으로 미국기준과 유럽기준 총 두 개의 타겟 데이터를 만들었습니다.

15.

다음으로 위험요인은 크게 세종류로 나누었고, 여러 특성들을 조합하여 11개의 특성으로 가공했습니다.

16.

데이터누수를 방지하기 위해 타겟 데이터를 만들 때 쓴 특성들은 쓰지 않았고

문제없음을 확인했습니다.

17.

그 다음으로 특성 간의 상관관계과 높게 나타난 특성 두개를 제거하였고,

18.

이상치를 살펴보았고,

19.

아래에 표기한 근거자료를 통해 정의하여 이상치를 제거하였습니다.

20.

최종 데이터에 대한 특징은 시간관계상 생략하겠습니다.

21.

22.

23

24.

모델링 과정입니다.

25.

26.

가장 먼저 훈련데이터와 테스트데이터를 8:2의 비율로 나누었습니다.

그리고 모든 검증과정은 교차검증을 통해 진행되었고, CV값은 5로 통일하여 진행했습니다.

27.

우선 분류 모델링에 앞서 타겟의 최빈 클래스로 기준모델을 만들었습니다.

데이터 불균형 문제가 없다고 판단해 정확도와 AUC스코어를 평가지표로 사용하였고 비교분석에는 AUC를 중점적으로 평가했습니다.

28.

우선은 총 네 종류의 모델을 디폴트 설정으로 모델링했습니다.

시간 관계상 바로 검증 스코어 비교분석으로 넘어가겠습니다.

29.

30.

31.

32.

33.

34.

35.

36.

37.

디폴트 모델간의 비교분석을 하겠습니다.

전반적으로 유럽기준이 미국기준보다 검증스코어가 높게 나타나고,

디시젼트리만 유독 낮은 검증스코어를 보여서 디시젼트리를 제외한 세개의 모델로 튜닝을 진행했습니다.

38.

모델 튜닝은 그리드서치CV라는 모듈을 통해서 진행하였고, 시간관계상 바로 비교분석으로 넘어가도록 하겠습니다.

39.

40.

41.

튜닝 전과 후를 시각화한 결과입니다.

튜닝 결과에서도 미국기준보다 유럽기준의 검증 스코어가 높게 나타났고,

비교분석시 AUC 스코어를 판단 근거로 삼았기 때문에 최종모델은 AUC스코어가 가장 높은 XG부스트 모델로 선정하였습니다.

42.

선택한 최종 모델에 대한 검증을 진행하였고,

43.

44.

이젠 분리해 놓았던 테스트데이터로 일반화가능성을 살펴보도록 하겠습니다.

미국 진단기준의 경우엔 차이가 크게 나타나서 과적합이 발생한 것으로 보이고,

유럽 진단기준의 경우엔 차이가 작게 나타나서 일반화가 매우 잘 된 모델이라고 볼 수 있습니다.

45.

이 모두를 종합해서 보았을 때 성능과 일반화 가능성 모두 유럽의 진단기준이 훨씬 좋다고 볼 수 있습니다.

46.

이젠 모델에 대한 해석을 진행하도록 하겠습니다.

47.

48.

우선 특성간 상관관계가 크지 않아서, 상관관계는 고려하지 않고 해석하겠습니다.

트리기반 모델의 특성중요도를 살펴본 것인데,

특성중요도는 값이 다양할 수록,

즉 카디널리티가 클수록 높은 값을 부여하는 문제가 있어서,

49.

그 단점을 보완하는 것 중 하나가 순열중요도 입니다.

공통적으로 나이, 비만도, 가족력을 가장 중요한 특성이라고 판단하고 있습니다.

여기서 0에 거의 수렴하는 흡연여부를 제외한 나머지 특성들에 대한 경향성을 살펴보도록 하겠습니다.

50.

특성의 분포에 따른 영향력을 볼 수 있는 PDP를 통해 경향성을 파악하였습니다.

우선 조절 불가 요인에서는 연령이 높을수록 고혈압 예측 확률이 꾸준히 증가하고, 여성보다 남성이 더 예측확률이 높게 나오고, 가족력이 강할수록 예측확률이 더 커집니다.

51.

그 다음으로 만성질환 요인들에서는 중성지방의 경우에도 꾸준히 예측확률이 커지고, 고콜레스테롤혈증이 있거나 당뇨가 있는경우 예측확률이 더 크게 나타납니다.

52.

그 다음엔 생활습관과 관련 요인에서는 BMI는 비만 1단계 진단기준인 25를 기준으로 급증하는 것을 알 수 있고, 폭음을 하는 사람의 고혈압 예측확률이 더 높게 나타납니다.

53.

앞에서 언급한 내용을 정리한 것이며, 시간관계상 넘어가겠습니다.

54.

결론입니다.

55.

목표 및 가설에 따라 요약해보겠습니다.

예측모델은 분류 문제로 정의했고, 최종 모델은 튜닝과정을 거친 XG부스트 모델이었습니다.

다음으로 유럽 진단기준이 성능과 일반화가능성이 더 높게 나타났습니다.

다음으로 공통적으로 가장 중요한 특성은 나이 특성이었고, 순열 중요도에선 진단기준에 관계없이 나이 비만도 가족력 순으로 중요도가 높게 나타났습니다.

56.

이를 통해 결론을 지어보면, 유럽의 진단기준이 모델의 성능과 일반화 가능성이 더 높게 나타나므로 예측의 관점에서는 보수적인 입장을 유지하는 것이 더 나을 것이라 보고,

특성들 중에서 나이가 가장 중요도가 높고, 고령일수록 예측확률이 증가하므로 고혈압 정책은 고령층을 최우선으로 고려하는 것이 좋을 것이라 보고,

나이에 이어서 비만, 가족력도 중요도가 높은 특성이었기 때문에 비만이거나 가족력이 있는 경우 고혈압에 좀 더 주의를 기울일 필요가 있습니다.

제가 생각하는 본 프로젝트의 한계점으로는 복부비만, 나트륨 과잉섭취 처럼 중요하다 알려진 변수를 사용할 수 없었던 점과

가족력 변수의 기준이 유럽기준과 동일한 보수적인 진단기준이라서 미국기준과 맞지 않았던 점 등을 꼽을 수 있을 듯 합니다.

57.

이것으로 발표를 마치도록 하겠습니다.

감사합니다.