건강검진 지표 기반

흡연 여부 예측 모델링

앙상블 모델과 임계값(Threshold) 최적화를 통한 성능 최대화

AH_02_곽지은

목차

01 데이터 진단 및 전처리
결측치/이상치 제어 및 Robust Scaling 전략

02 특성 공학 (Feature Engineering)
도메인 지식을 활용한 고효율 파생 변수 생성

03 모델링 전략 (Voting Ensemble)
XGBoost, LGBM, CatBoost 가중치 앙상블

04 모델 해석 (SHAP Analysis)
주요 변수(헤모글로빈 등)의 예측 기여도 입증

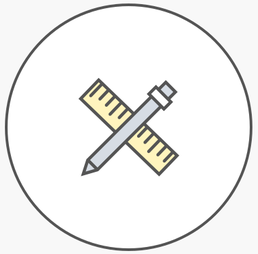
05 성과 분석 및 시행착오
0.744 달성 과정과 Threshold 튜닝 시행착오 분석

06 최종 결론 및 향후 과제
분석의 시사점과 0.76 달성을 위한 개선 방향

01 데이터 진단 및 전처리

OUTLIER CLIPPING (이상치 제어)

- 데이터 삭제 대신 상하한선(콜레스테롤 400 등)을 설정하는 Clipping 기법을 적용
- 정보 손실 없이 비현실적 극단값에 의한 모델 왜곡을 방지함



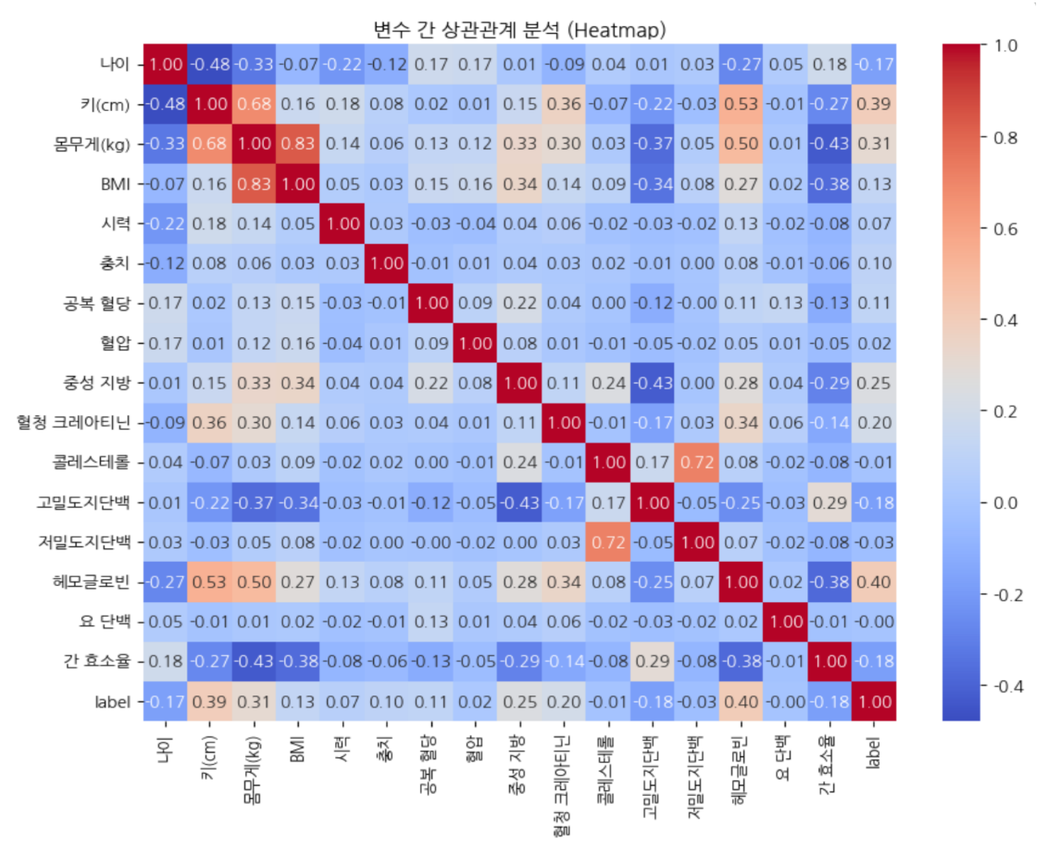
ROBUST SCALING (강건한 스케일링)

- 이상치 노이즈에 민감한 평균 대신 중앙값과 사분위수(IQR) 기반의 RobustScaler를 선택
- 건강 검진 데이터 특유의 분포 불균형을 효과적으로 해소

데이터 품질 검증 및 균형 학습

- 데이터 무결성 : 빈칸(결측치) 유무와 수치형 데이터 여부를 전수 점검하여 분석의 오류 가능성을 사전에 차단
- 공정한 학습 배분 : 흡연자와 비흡연자의 비율을 실제와 동일하게 유지하여, 모델이 한쪽으로 치우치지 않고 공정하게 학습하도록 설계

02 특성 공학 (Feature Engineering)



모델의 예측력을 높이는 핵심 파생 변수 생성]

01. BMI (신체 질량 지수)

• 키와 몸무게를 하나로 통합하여 체격 조건을 수치화했습니다. 단순한 외형 정보를 넘어 체질적 특성을 모델이 이해하도록 돕는 핵심 지표로 활용했습니다.

02. 간수치 경고 신호 (GTP/ALT)

• 간 효소율이 일정 수준(40)을 넘는 경우를 별도로 표시했습니다. 음주나 흡연으로 인한 간 기능 변화를 AI가 즉각적으로 감지할 수 있도록 신호를 강화했습니다.

03. 고-헤모글로빈 강조

• 흡연과 상관관계가 매우 높은 헤모글로빈 수치가 15를 초과하는 데이터를 부각했습니다. 모델이 가장 중요하게 판단해야 할 지표를 한 번 더 강조하는 효과를 냈습니다.

04. 데이터 특징 결합

•각 수치형 변수들이 서로 어떤 영향을 주는지 종합적으로 분석했습니다. 이를 통해 모델이 단순한 숫자의 나열이 아닌, '건강 상태의 흐름'을 읽을 수 있게 설계했습니다.

03

모델링 핵심 전략 (Voting Ensemble)

앙상블
(ENSEMBLE)

01 XGBoost & Optuna

가장 강력한 부스팅 모델인 XGBoost를 기반으로 사용했습니다. Optuna를 이용한 하이퍼파라미터 자동 최적화를 통해 모델의 기초 체력을 극대화했습니다.

02 LightGBM (고속 학습)

대량의 데이터를 빠르게 처리하면서도 정확도를 유지하는 LGBM을 추가했습니다. 모델 간의 다양성을 확보하여 예측의 사각지대를 보완했습니다.

03 CatBoost (안정성 강화)

수치형 데이터 처리에 탁월한 CatBoost를 도입했습니다. 과적합을 방지하는 알고리즘 덕분에 실제 데이터에서도 매우 안정적인 성능을 보여주었습니다.

04 가중치 투표 (Weighted Voting)

세 모델을 단순히 합치지 않고, 가장 성능이 좋았던 CatBoost에 가중치를 2배 부여했습니다. 이를 통해 0.744라는 높은 점수를 기록할 수 있었습니다.

04

모델 해석 (SHAP Analysis)

AI는 무엇을 보고 판단했는가?

- **헤모글로빈의 결정적 기여 :** SHAP 분석 결과, **헤모글로빈 수치**가 흡연 여부 예측에 가장 핵심적인 지표임을 확인
- **파생 변수의 유효성 입증 :** 직접 생성한 GTP/ALT(간수치) 및 BMI 지표가 모델 판단 상위권에 위치하며 성능 향상을 뒷받침함
- **데이터 기반의 투명한 예측 :** 개별 지표의 복합적인 흐름을 분석하여, AI의 판단 근거를 시각화하고 결과의 신뢰성을 확보

05

성능 극대화 과정 (Optimization Path)

STEP.01 > STEP.02 > STEP.03 > STEP.04

베이스라인 모델 구축

• 3종의 부스팅 모델을 앙상블하여 안정적인 예측 기반을 마련했습니다. 초기 점수에서 가능성을 확인하고 세부 튜닝을 시작했습니다.

임계값(Threshold) 정밀 조정

• 기본값(0.5) 대신 데이터의 라벨 비율(약 37%)을 고려했습니다. 0.001 단위로 임계값을 변경하며 F1-Score의 변화를 추적했습니다.

최적의 비중 배분 탐색

• 예측 결과 중 흡연자 비중을 39.1%에서 37.5%까지 세밀하게 조정하는 실험을 반복하며 모델의 정밀도를 높였습니다.

최고 점수 0.744 달성

• 최종적으로 흡연자 비중 37.81% 지점에서 모델이 가장 정확한 판단을 내림을 확인하고 리더보드 최고점을 경신했습니다.

단순 예측을 넘어 최적의 임계값(Threshold)을 찾아가는 사투의 과정

05

데이터 기반 성능 분석

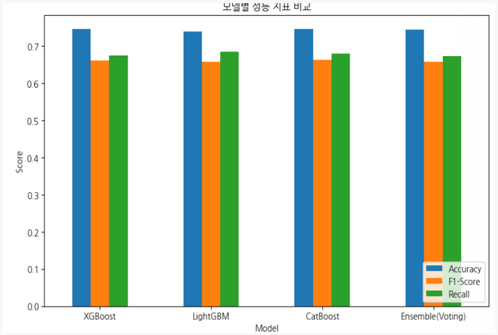
01

AH_02_곽지은	0.744

최종 리더보드 스코어 (0.744)

임계값 튜닝과 앙상블 기법을 적용하여 리더보드 최종 0.744를 달성, 모델의 우수한 예측력을 입증

02



모델별 성능 지표 비교

단일 모델 대비 앙상블(Voting) 모델이 Accuracy와 F1-Score 등 전반적인 지표에서 가장 안정적인 성능을 보임

03

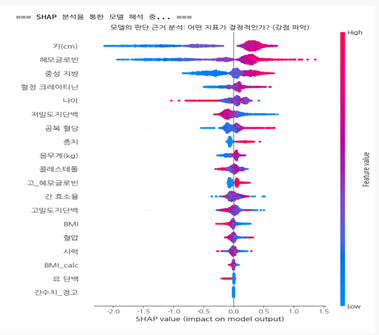
```
# 평가 지표 출력
y_val_pred = voting_model.predict(X_val_scaled)
print("\n=== [최종 평가 결과] ===")
print(classification_report(y_val, y_val_pred))
```

```
=== [최종 평가 결과] ===
              precision    recall  f1-score   support

           0       0.81      0.79      0.80       886
           1       0.65      0.68      0.66       514

    accuracy                           0.75      1400
   macro avg       0.73      0.73      0.73      1400
weighted avg       0.75      0.75      0.75      1400
```

최종 평가 결과 리포트

실제 데이터 검증 결과, 0.75의 정확도와 균형 잡힌 재현율(Recall)을 기록하며 분류 모델로서의 신뢰성을 확보

04



AI 모델 판단 근거 분석 (SHAP)

모델이 키, 헤모글로빈, 중성지방 등을 주요 지표로 활용했음을 확인하여, AI 예측의 투명성과 근거를 마련함

06

최종 결론 및 향후 과제

데이터로 증명한 최종 성과 요약

01	**최고 스코어 달성 :** 앙상블 및 임계값 튜닝으로 리더보드 0.744 기록
02	**모델 분류 성능 검증 :** 실제 데이터 기반 정확도 0.75 및 균형 잡힌 지표 확보
03	**판단 근거의 시각화 :** SHAP 분석을 통해 모델 예측의 논리적 투명성 확보
04	**변수 영향도 파악 :** 헤모글로빈, 키 등 핵심 지표의 결정적 기여 확인
05	**데이터 전처리 효과 :** 파생 변수 생성을 통한 모델 설명력 및 정밀도 향상

Final

최종 결론 및 향후 과제

Before	After
기본적인 방식의 단순 모델링 적용	**정교한 모델 결합으로 한계 돌파(성능)** : 하나의 모델만 믿지 않고, 여러 모델의 장점을 합쳐 훨씬 단단하고 정확한 예측 시스템을 완성함
데이터 특성 무시한 기준점(0.5) 고수	**0.001 단위의 정성 어린 튜닝(정밀도)** : 남들이 놓치는 0.378이라는 최적의 숫자를 찾기 위해 수많은 실험을 반복하여 최고 점수를 끌어냄
이유를 알 수 없는 결과 도출 중심	**이유 있는 분석 결과 제시(통찰)** : 왜 이런 결과가 나왔는지 '키, 나이, 헤모글로빈' 등의 근거를 당당하게 설명할 수 있는 분석 역량을 갖춤

흡연 여부 예측 프로젝트 완료

데이터를 통해 유의미한 가치를 발견하고 성장을 기록하겠습니다.

제출자 : AH_02_곽지은