Term Project: 수면 질과 일상생활 습관과의 상관관계 분석

22101149 강영훈

목차

1. 분석할 데이터 설명 및 분석 목적
2. 분석을 위한 코드 동작 설명
3. 데이터 분석 결과
4. 실습한 Library code 어떻게 반영되었는지에 대한 설명
5. 분석할 데이터 설명 및 분석 목적

이 데이터는 수면 및 생활 습관과 관련된 다양한 13개의 변수를 다룹니다.

* Person ID: 각 참가자를 고유하게 식별하는 ID
* Gender: 참가자의 성별
* Age: 참가자의 나이
* Occupation: 참가자의 직업
* Sleep Duration: 하루 평균 수면 시간 (시간 단위)
* Quality of Sleep: 수면의 질을 1에서 10까지 평가 (1이 최악)
* Physical Activity Level: 일일 신체 활동 시간 (분 단위)
* Stress Level: 주관적인 스트레스 수준을 1에서 10까지 평가 (10이 최악)
* BMI Category: BMI에 따른 체중 분류 (Underweight, Normal, Overweight, Obese)
* Blood Pressure (systolic/diastolic): 혈압 측정값 (최고/최저)
* Heart Rate (bpm): 심박수 (분당 박동수)
* Daily Steps: 일일 걸음수
* Sleep Disorder: 수면 장애 유무 (None, Insomnia, Sleep Apnea)

본 데이터 분석의 목적은 다음과 같습니다.

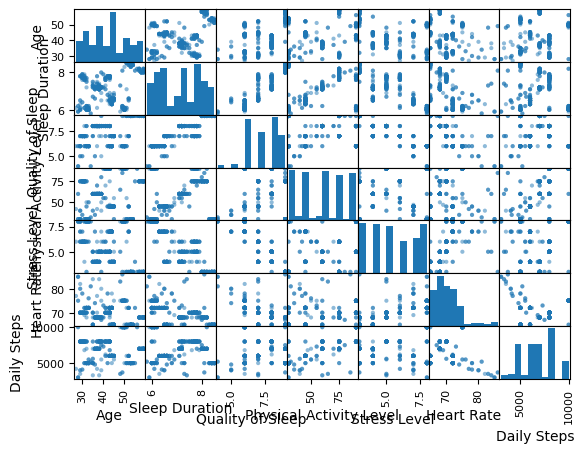
* 수면의 질에 영향을 미치는 주요 요인을 식별합니다.
* 분석 결과를 바탕으로 더 건강한 수면 습관을 위한 지침을 제시합니다

1. 분석을 위한 코드 동작 설명

* 해당 데이터를 multiple linear regression과 Decision Tree classifier 두 가지 방식으로 분석하고자 합니다. 두 가지 방식을 사용한 이유는, 두 방식 중에서 더 높은 prediction을 제공하는 모델을 사용하기 위함입니다. 모든 parameter 간의 correlation을 알아보기 위해 scatter matrix를 관찰했을 때, 데이터가 linear regression, Decision Tree 두 가지 방식 중 어떤 방식이 더 좋을지 정확히 나누기 애매한 부분이 존재하여 두 가지 방식 모두 사용한 뒤에 R2 등의 지표를 활용하여 두 방식을 비교했습니다.
* 준비한 데이터에서 linear regression에 필요한 데이터인 numeric data만 사용할 수 있도록 데이터를 적절히 조작하여 Age, Sleep Duration, Physical Activity Level, Stress Level, Heart Rate, Daily Steps만 column에 남기고 fitting 합니다. Interaction이 큰 값이 있을 수도 있으므로 correlation을 확인하고 상관관계가 큰 Physical Activity Level과 Daily Steps 간의 Interaction을 고려하여 model을 다시 fitting 합니다.
* Decision Tree classifier 방식을 사용하기 위해 linear regression에서 사용했던 데이터에서 intercept만 제외한 데이터를 transform하여 fitting에 사용합니다. 이후에 Boosting을 진행하고 최종적으로 완성된 트리를 제작합니다.
* 마지막으로 두 모델에 대해서 R-sqaured와 MSE를 파악하여 더 잘 fitting된 model을 통해 다른 test case에 대해 수면의 질을 높이기 위해 적절한 생활 습관을 제시할 수 있습니다.

1. 데이터 분석 결과

* 숫자형 데이터에 대하여 scatter matrix를 조사한 결과, Quality of Sleep에 대해서 다른 데이터들이 Linear한 경향성을 띄는 것처럼 보입니다.



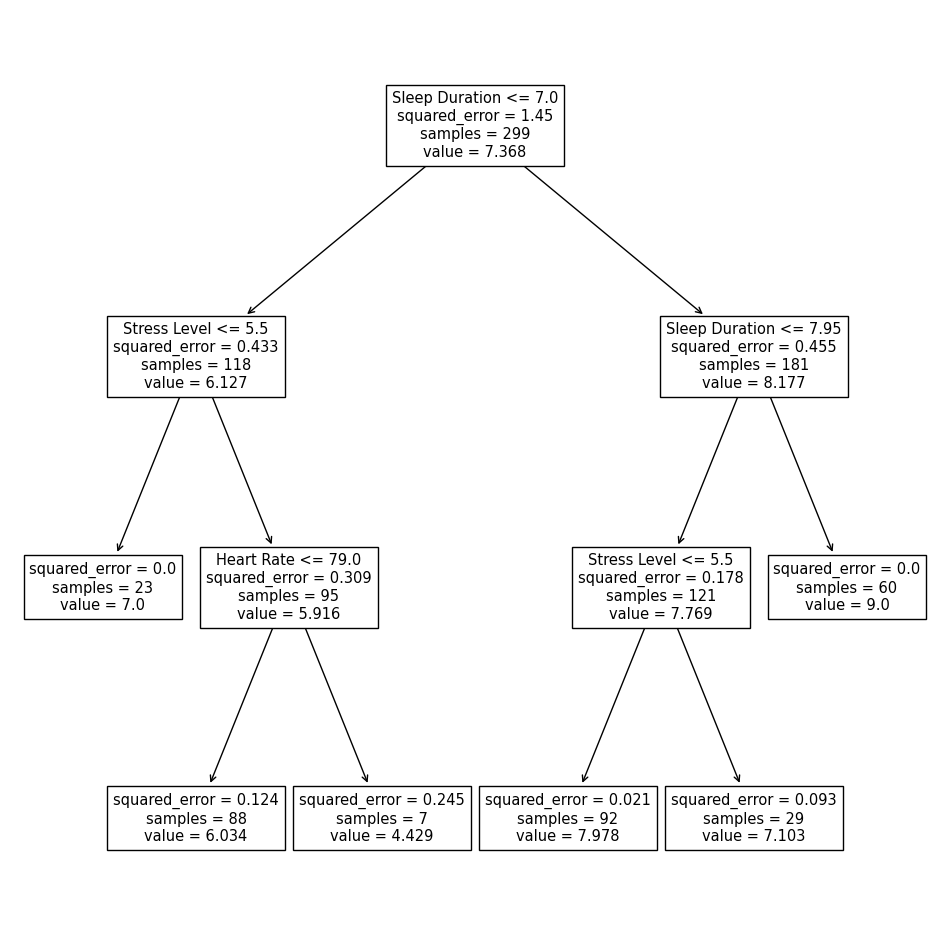
* 우선 numeric data에 대해 linear regression을 진행하였습니다. 이때 Physical Activity Level에서만 t-value가 0.738로 다른 값들에 비해 특히 높아 제곱을 하거나 interaction을 고려한 값을 사용했습니다. 제곱을 했을 때는 R-squared가 낮아 model이 제대로 fitting되지 않았고, Physical Activity Level과 correlation이 다른 parameter에 비해 특히 높은 Daily Steps와 interaction을 고려하는 방법으로 model을 fitting하여 결과를 도출했습니다. 실제로 Physical Activity Level과 Daily Steps의 correlation은 0.77로 나타났고, 다른 value와는 0.2 이하의 수치로 나타났습니다.

텍스트, 스크린샷, 메뉴, 폰트이(가) 표시된 사진

자동 생성된 설명

Linear regression의 결과로 Quality of Sleep에 큰 영향을 주는 것은 Sleep Duration으로 평균적으로 1시간을 자는 것이 0.62만큼 Quality of Sleep을 증가시킵니다. 또한, Stress Level이 1만큼 낮아질 때 0.29만큼 Quality of Sleep을 증가시킬 수 있습니다. 그 외에도 Physical Activity Level, Daily Steps도 Quality of Sleep에 작은 영향을 줄 수 있다고 볼 수 있습니다.

* Regression Tree를 만들기 위해 data를 training, test sets으로 8:2로 구분하고, training data에 대해 fitting 합니다. 그리고 boosting을 사용하여 fitting 합니다. fitting 결과, Sleep Duration이 7.95보다 큰 경우에 Quality of Sleep이 Sample mean이 9였고, Sleep duration이 7보다 크고, Stress Level이 5.5보다 작을 때 sample mean이 7.978로 약 8의 Quality of Sleep이었습니다. 최악의 경우엔 4.429로 Sleep Duration이 7보다 작고 Stress Level이 5.5보다 크며 Heart Rate가 79보다 컸습니다.



* Linear regression의 경우 R2의 값은 0.906으로 알고 있고, MSE를 알아내기 위해 sklearn library를 사용하였습니다. 그 결과 0.225를 도출했습니다. Tree model은 MSE가 0.036으로 알고 있고, R2를 알아내기 위해 sklearn library를 사용했습니다. 그 결과 0.942를 도출했습니다. Tree model이 R2는 더 크고, MSE는 더 작기 때문에 Tree model이 Linear regression model보다 더 잘 fitting 되었다고 볼 수 있습니다.
* 따라서 Tree model을 통해서 좋은 수면의 질을 얻기 위해 몇 가지 제안을 할 수 있습니다. 최우선으로 수면 시간을 7시간 이상으로 해야 하며 7.95시간 이상이면 더 좋다고 할 수 있습니다. 그 다음으로는 Stress Level을 5.5 이하로 관리해야 합니다. 하지만 Stress Level 이란 것은 개인적인 요소라고 할 수 있기에 평소에도 스트레스 수치를 5보다 작게 할 수 있도록 Stress Level과 관련 있는 Heart Rate를 높지 않게 관리해야 합니다. Heart Rate는 79보다 낮게 관리하는 것이 좋다고 할 수 있습니다.

1. 실습했던 Library code가 어떻게 사용되었는지 확인

* 데이터가 어떻게 분포되어 있는지 확인하기 위해서 scatter\_matrix() 메서드를 사용했습니다.

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

* Multiple linear regression을 fitting 하는데 Chapter 3에서 Linear regression에 사용했던 ISLP library를 활용했고, predictor를 non-linear 형태, Interaction을 고려한 형태 모두 사용하였습니다. Data간 correlation을 확인하는 코드도 실습에서 사용했던 corr() 메서드를 사용했습니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

* Decision Tree model을 만들기 위해 Chapter 8에서 Regression Tree, Boosting를 구성하는 데 사용했던 과정을 참고했습니다.

텍스트, 스크린샷, 폰트, 디스플레이이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

* Model fitting의 결과를 비교하기 위해 Linear regression의 R2, Boosting의 MSE를 확인할 때 각각 Chapter 3, 8에서 활용했던 방법을 사용했습니다.

텍스트, 폰트, 스크린샷, 디자인이(가) 표시된 사진

자동 생성된 설명 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명