

Physical Feature Analysis

for age & gender estimation

Kakao Internship Applicant

임 환 규

문제 정의

- 간단한 EDA 를 실시하고 리포트를 작성
- “성별코드”와 “연령대코드”는 타겟 변수
- 연령대코드는 10 의 배수 단위로 사용
- 모델링 결과에 대한 분석

Abstract

- Simple EDA report
- System Architecture
- Result Analysis
- Conclusion

- Appendix

Statistical Analysis

Modeling AgeNet & GenderNet

Attention & Feature

Are The Results Reliable?

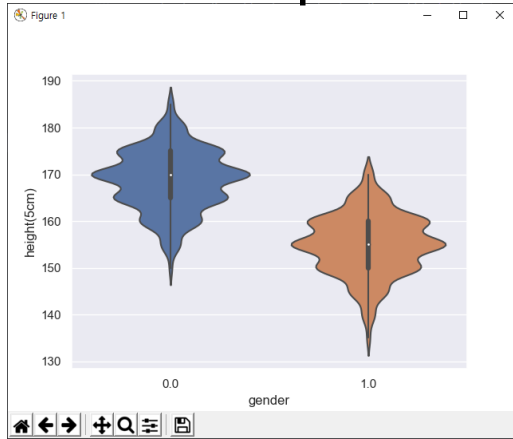
Simple EDA report

데이터 특징 분포 시각화

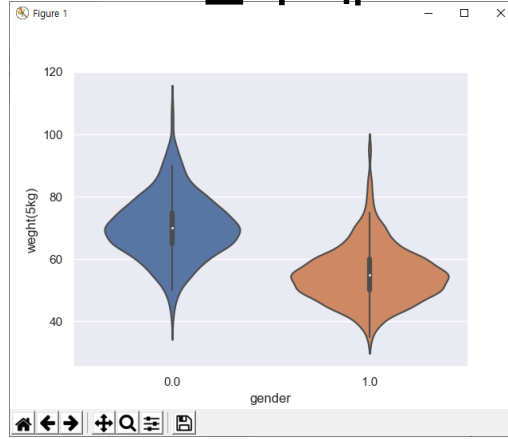
- Violin Plot과 Correlation Matrix를 사용하여 시각화
- 시각적으로 구분되는 특징을 주목함
- 간단한 Threshold 실험 진행

데이터 특징 분포 시각화 (gender)

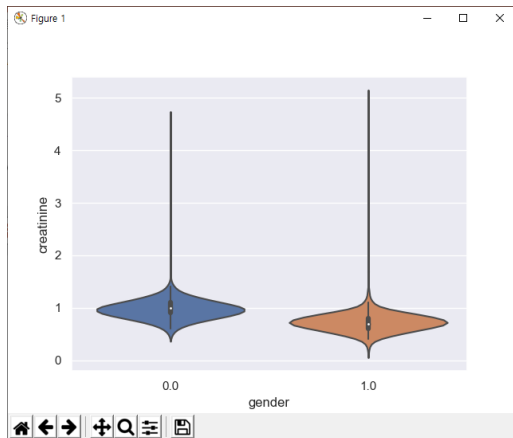
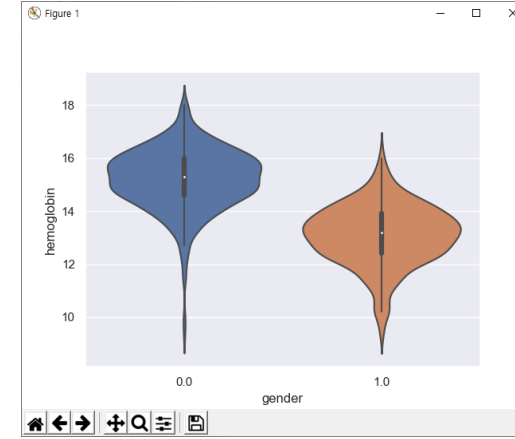
키



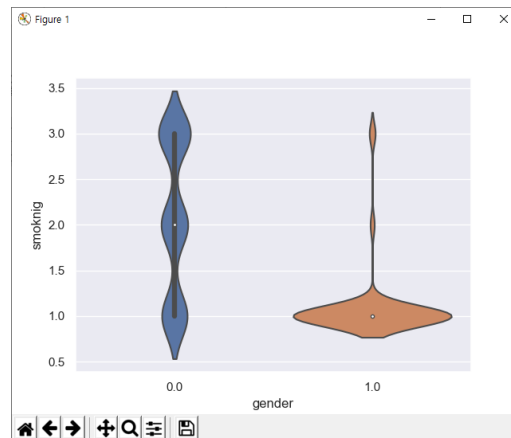
몸무게



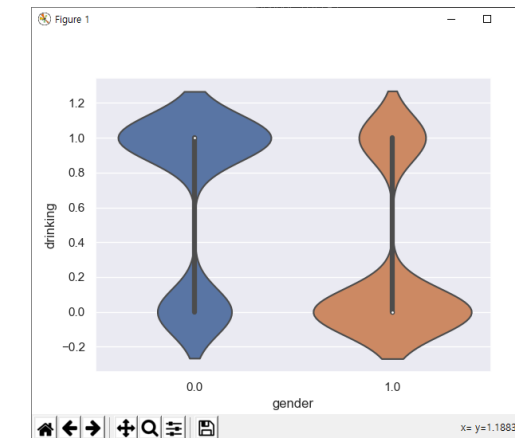
혈색소



크레아티닌

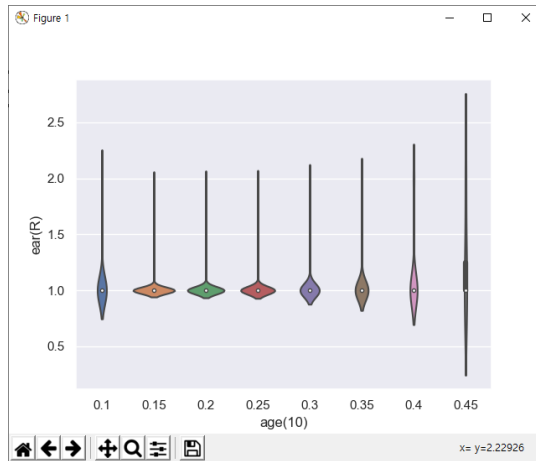


흡연

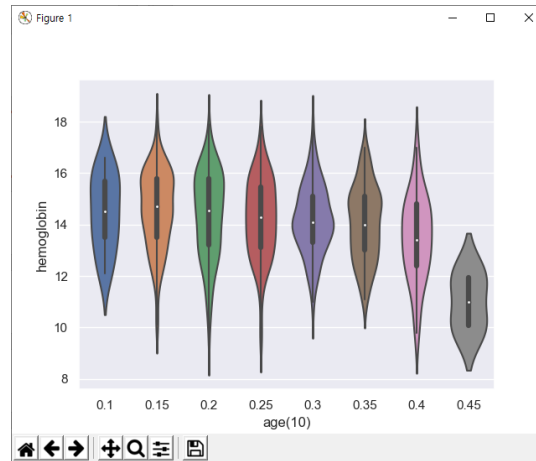


음주

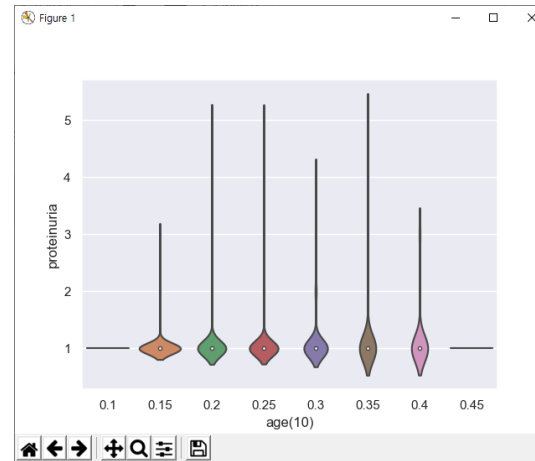
데이터 특징 분포 시각화 (age)



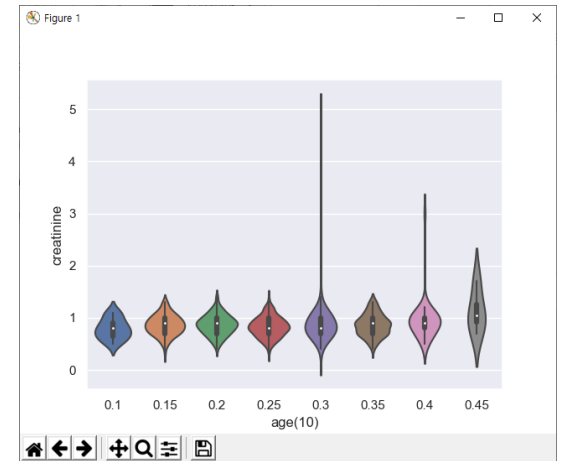
청력



혈색소



요단백



크리아티닌

Simple Age Threshold Test

- Weak Classifier

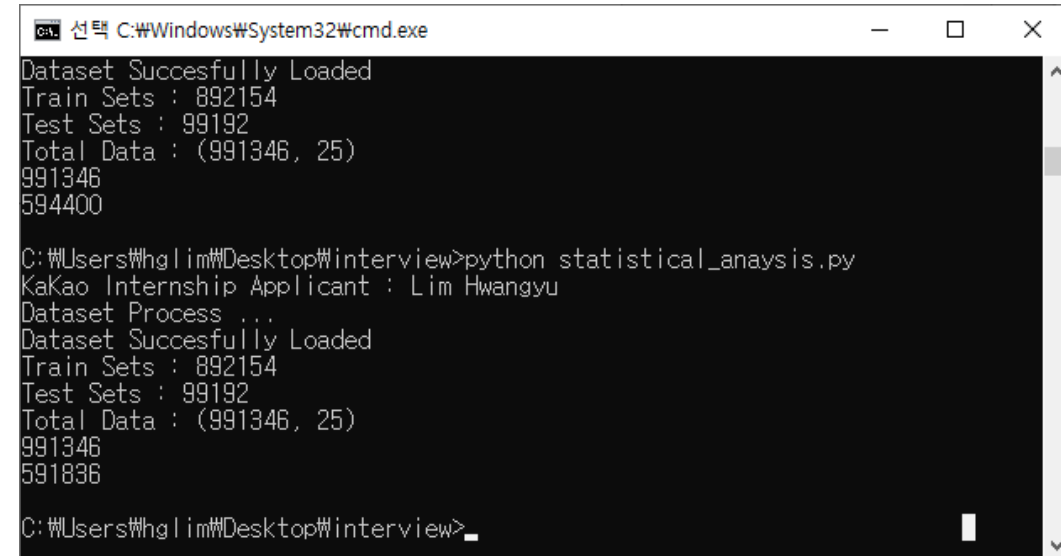
data \geq threshold ? 남자 : 여자

혈청크레아티닌 threshold = 1.0

- 약 78%의 분류 정확도

혈청지오티 threshold = 22.0

- 약 60%의 분류 정확도



```
C:\Windows\System32\cmd.exe
Dataset Successfully Loaded
Train Sets : 892154
Test Sets : 99192
Total Data : (991346, 25)
991346
594400

C:\Users\hnglim\Desktop\interview>python statistical_anaysis.py
KaKao Internship Applicant : Lim Hwangyu
Dataset Process ...
Dataset Successfully Loaded
Train Sets : 892154
Test Sets : 99192
Total Data : (991346, 25)
991346
591836

C:\Users\hnglim\Desktop\interview>
```


Correlation Matrix를 이용한 분석

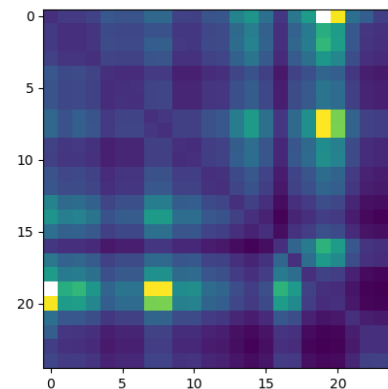
$$\text{Corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$



$$\text{Corr}_{normal}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \frac{1}{\text{corr}(X, X)}, (X < Y)$$

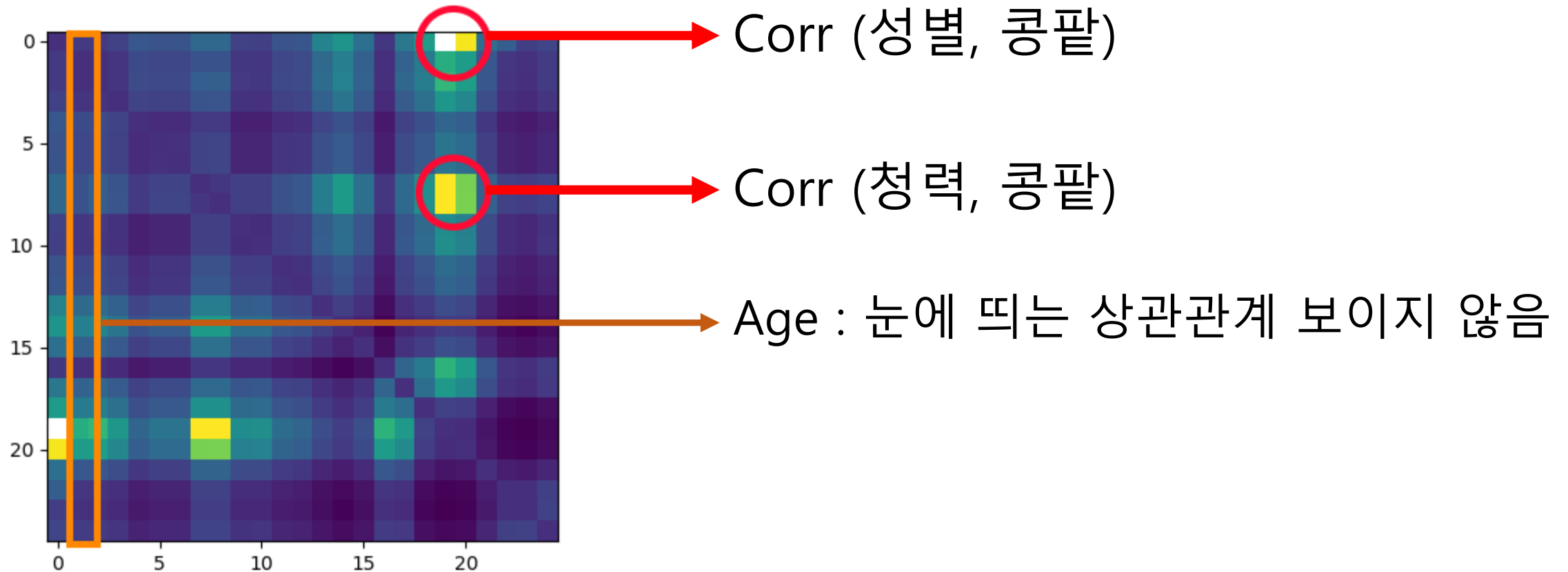
$$\text{Vis}(X, Y) = \log \text{Corr}(X, Y) - \min(\log \text{Corr}(X, Y))$$

- 특징 별 직접적인 상관관계 분석



Simple EDA report

- Correlation Matrix를 이용한 분석



Simple EDA report

- Is It Reliable?

저작권© '건강을 위한 정직한 지식'
코메디닷컴(<http://kormedi.com>)

남녀 콩팥 기능 차이가 나는 이유는?

[이태원 박사의 콩팥이야기]

코메디닷컴 입력 2019년 10월 24일 15:01 조회수: 1,318

혈청 크레아티닌의 정상수치는 남녀 간에 다르다. 내과학 교과서에 의하면 여성은 0.5~0.9 mg/dL, 남성은 0.6~1.2 mg/dL로 여성보다 남성에서 높다. 크레아티닌이 근육의 대사산물인데 남성의 근육이 여성보다 많기 때문이다. 이러한 이유로 혈청 크레아티닌 수치에 대한 해석은 남녀를 구분하여 보아야 한다. 예를 들어 어떤 사람에서 측정한 혈청 크레아티닌

신장(콩팥) 정보로 남녀 구분 가능

이현정, "만성 신장질환자, 이명 발생 위험 높아", 헬스조선, 2018

만성 신장질환자, 이명 발생 위험 높아

이현정 헬스조선 기자 2018/02/08 09:03

신장서 못 걸러낸 독소에 혈관 손상... 청각 신경 문제 유발

만성 신장질환자 등 신장 기능이 떨어진 사람은 귀에서 소리가 나는 '이명(耳鳴)'이 생길 수 있다. 국제학술지 '플로스원'에 게재된 '만성 신장질환자의 이명 위험 증가' 연구에 따르면 대만의 만성 신장질환자 18만5430명과 건강한 사람 55만5290명을 대상으로 이명 발생 여부를 조사한 결과, 만성 신장질환자는 건강한 사람에 비해 이명 발생 위험이 3.02배로 높았다.

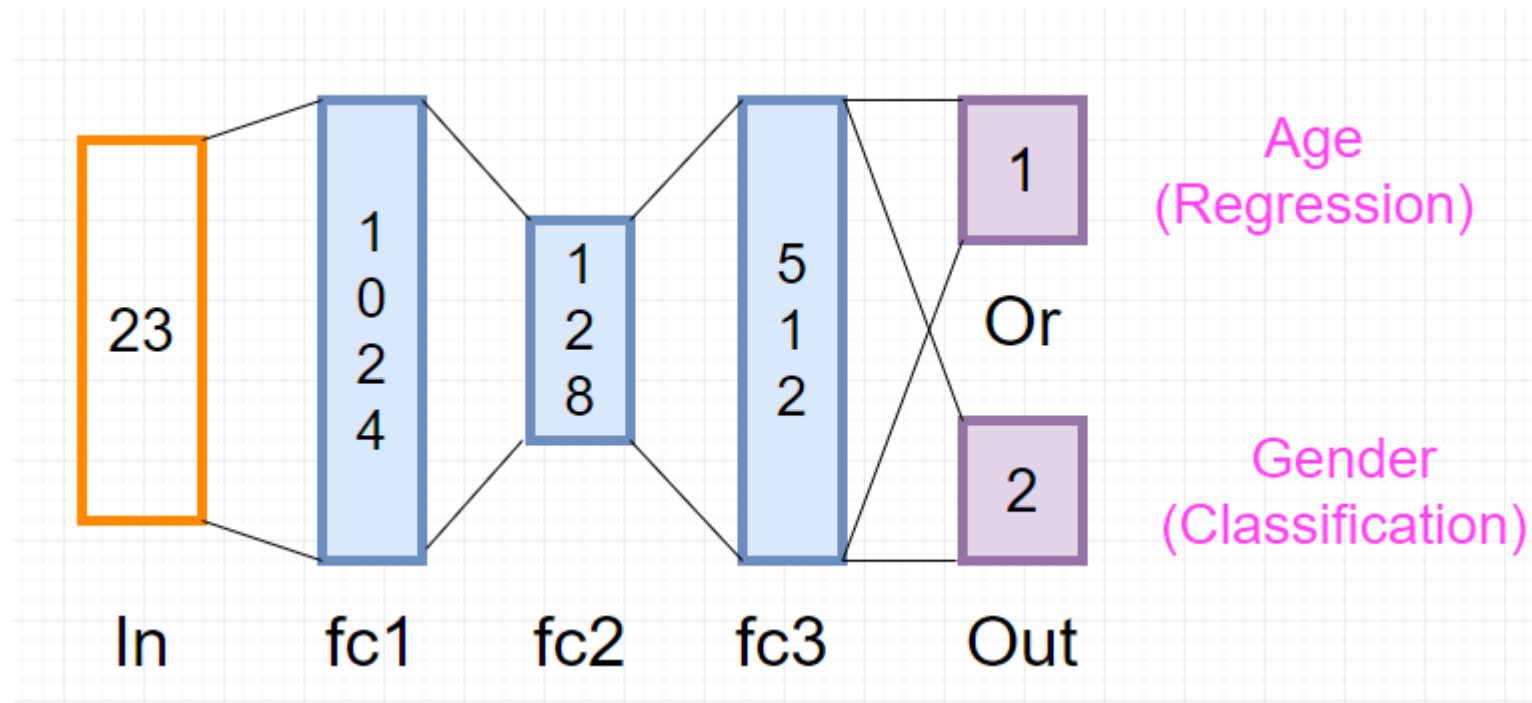
신장과 청력은 상관관계가 있음

EDA Conclusion

- Gender를 결정하는 특징은 상대적으로 명확함
 - Ensemble learning으로도 높은 정확도 획득 가능 예상
- Age의 특징은 명확하지 않음
 - Perceptron을 이용하여 Semantic Feature를 활용해야 함

System Architecture

Three Level Perceptron

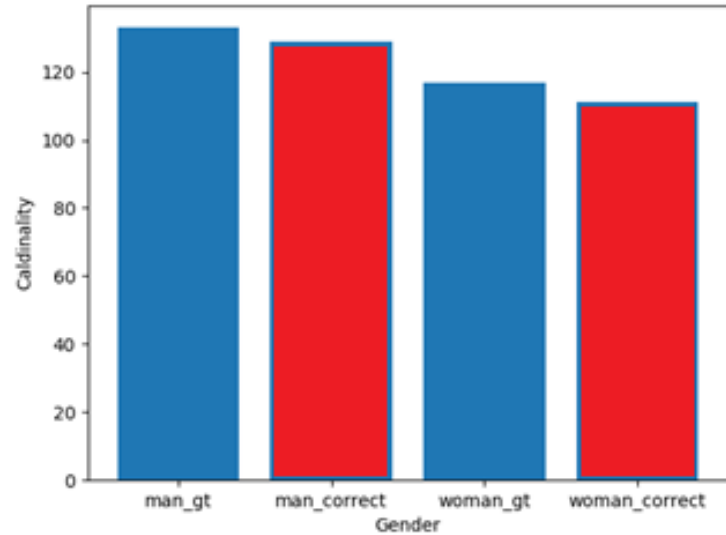


- (모델 상세는 Appendix에 수록)

Training

- Gender
 - Binary Classification
 - 학습 수행 후 각각의 성별에 대한 정확도를 측정하여 시각화
- Age
 - 10세 단위의 학습 데이터로 Regression를 통한 나이 추정
 - 추정된 나이를 다시 10세 단위로 양자화 하여 정확도 분석

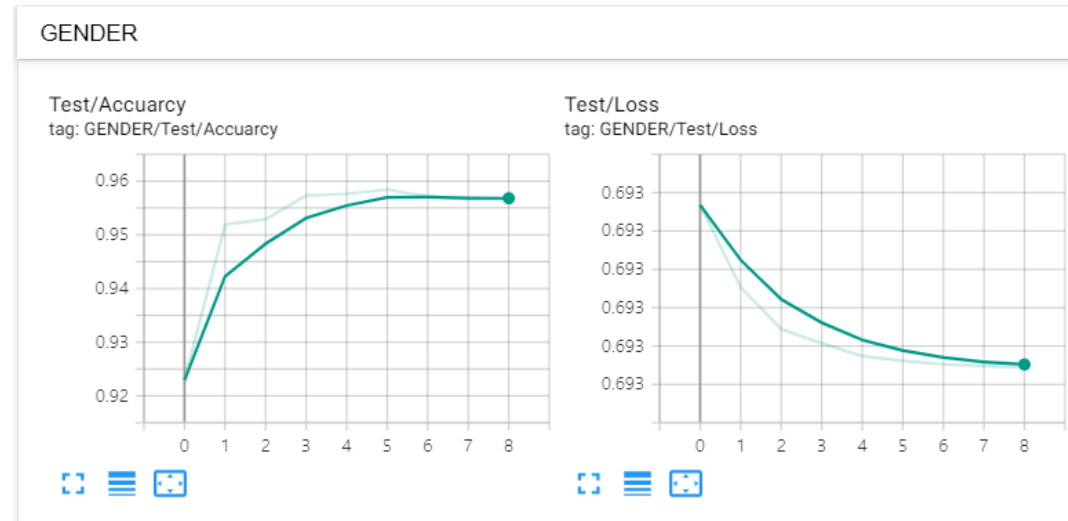
Gender Estimation



랜덤 샘플 Test 결과

실험 항목 개수 (파랑)

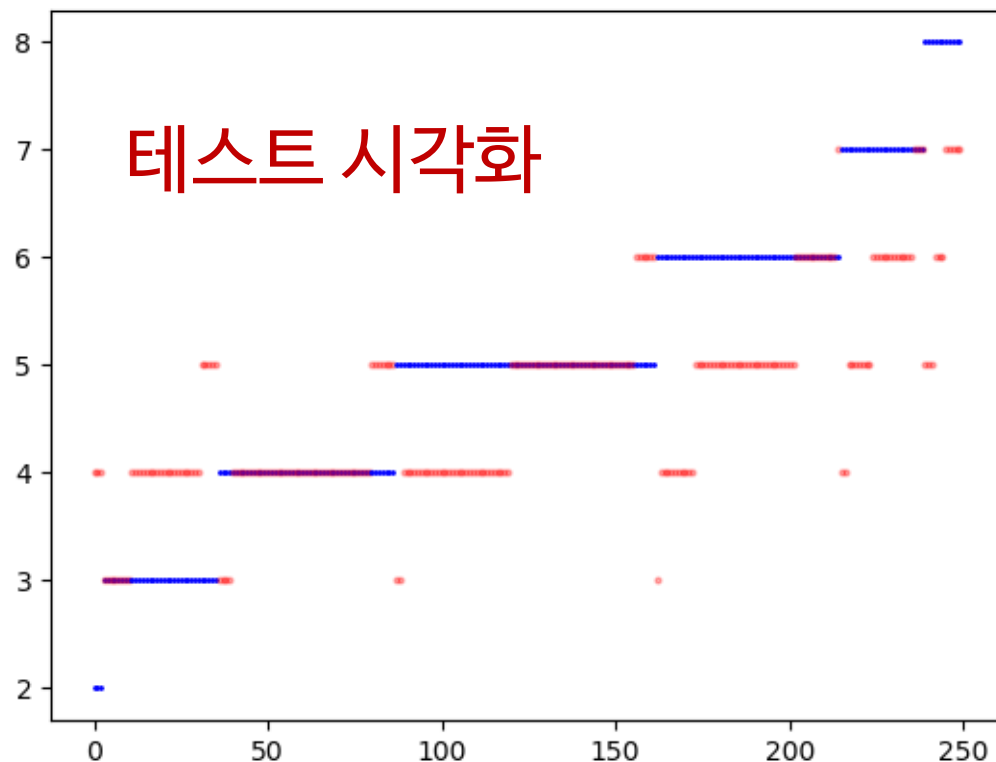
맞춘 항목 개수 (빨강)



- Test Set에서 96% 분류 정확도

Age Estimation

Scatter: Blue is Ground Truth

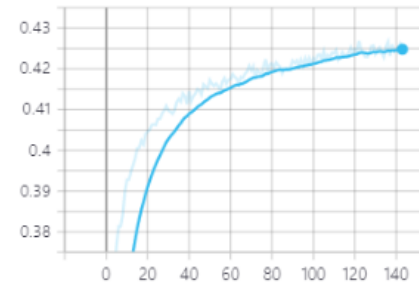


파란색 점이 Ground Truth
빨간색 점이 Prediction Result

Test

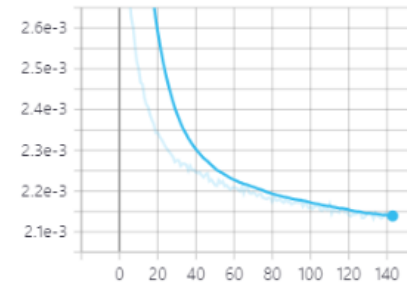
Accuracy/Age

tag: Test/Accuracy/Age



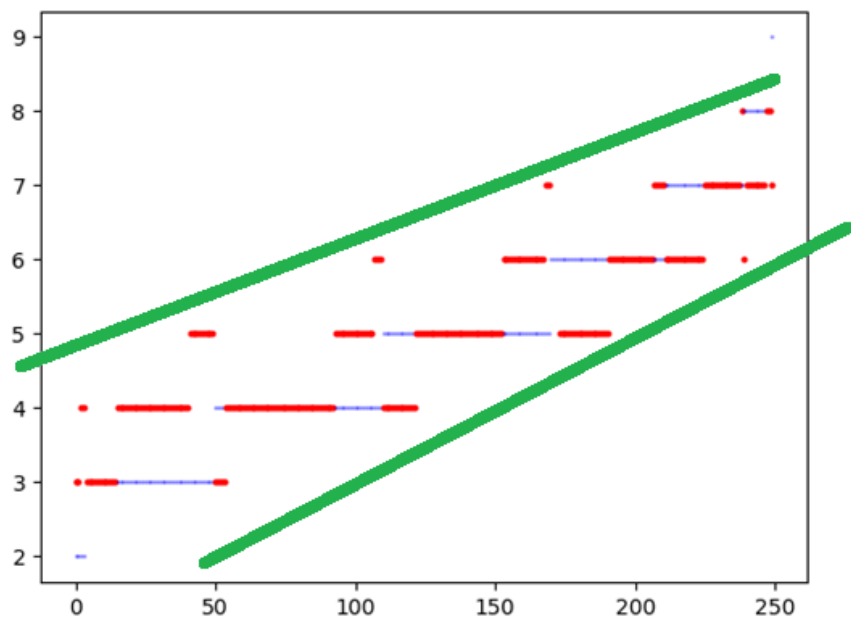
Loss/Age

tag: Test/Loss/Age



Age Estimation

항상 Ground Truth와 비슷한 분포 나타남



- Train data를 10대 단위로 양자화
오차 발생 불가피

Ex) Cls(20세) = Cls(19세) : 한살 차이지만 다른 Class

Cls(20세) = Cls(29세) : 아홉살 차이지만 같은 Class

- 인접 연령대 허용 정확도
 - Test Set에서 91% 분류 정확도
- 엄밀한 정확도
 - Test Set에서 43% 분류 정확도

Result Analysis

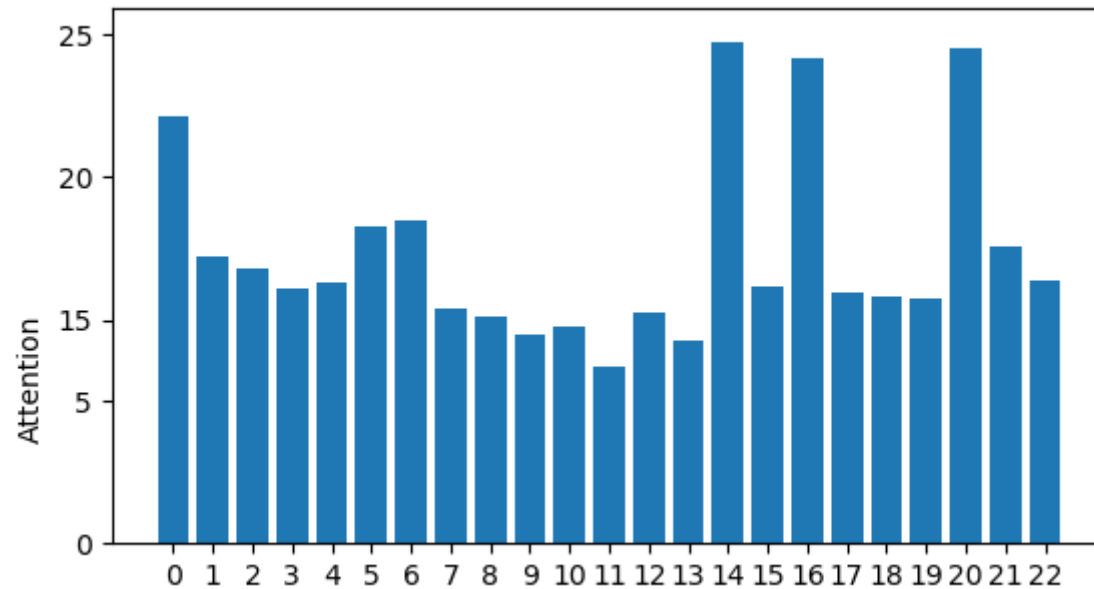
Attention Visualization

- Interpretable level에 대한 Attention 분석
- 특징 별 대략적인 결정력 추정 가능
- Correlation보단 Semantic한 분석 가능

$$\text{Attention}(f) = \sum_i w_{f,i} (fc1)$$

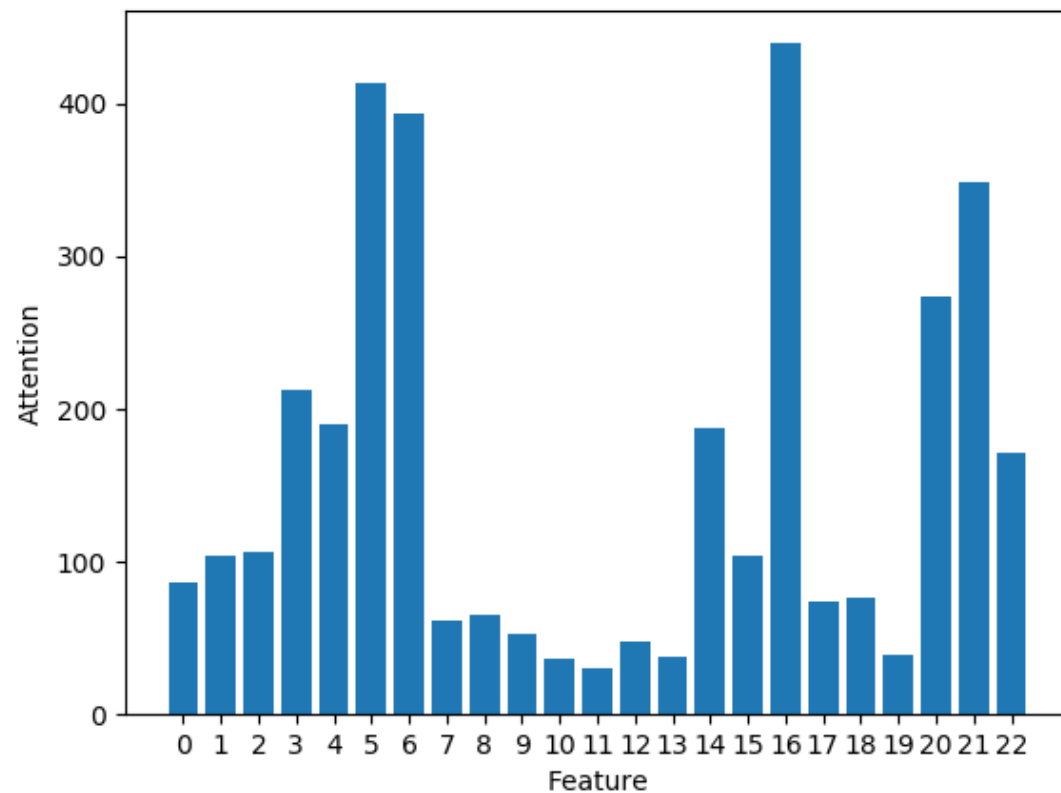
Attention Gender

- Top : 키, 혈색소, 혈청크레아티닌, 흡연여부



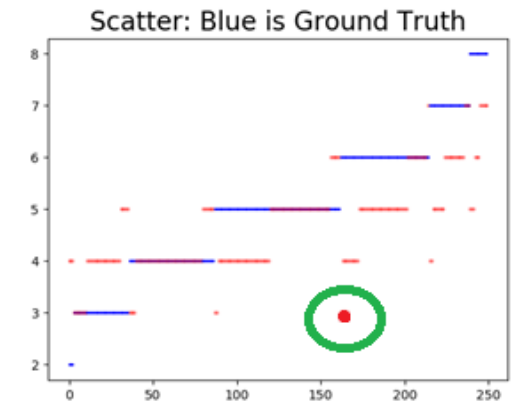
Attention Age

- Top : 청력, 혈청크레아티닌, 시력, 음주, 흡연



Conclusion

- Three-Level-Perceptron의 학습이 잘 이루어 진다
 - 의학적 해석과 Attention이 일치함을 확인할 수 있다.
- 검진정보는 남녀를 구분할 수 있다
- 검진정보와 나이에 대한 상관관계가 존재한다
 - 단, 건강정보에 대한 나이의 Outlier가 많다
 - 건강관리를 잘하면 50대에도 30대의 몸을 가질 수 있음



추가 연구

- 전처리를 이용한 성능향상
 - 학습데이터의 심한 이상치 제거를 이용하여 정확도 향상 (ex 허리둘레 999)
- 데이터셋의 개선
 - 청력 데이터 : 나이에 대한 식별력이 높지만 현재 데이터는 2-categories인 한계
(가청주파수는 연령대마다 1000Hz씩 감소한다)
 - 나이 데이터의 세분화
목적 데이터가 10대 단위로 분할되었기 때문에 Regression이 어려움
(Ex: 19세와 20세의 신체정보와 10세와 29세의 신체정보에 따른 loss 차별을 둘 수 없음)

감사합니다