Cover page (1 page)

REPORT

수면개선을 통한 비만관리

· 과 목: 데이터 사이언스

· 조 원: 2022103952 김나리, 2020105716 이건,

2020105742 한지훈, 2020103913 김동규, 2017103709 김거륜

2019110473 김태완, 2022103972 서유정,

2020105707 선신욱, 2020105709 성정윤, 2018102121 유창현

Summary (1 page)

-앞서 Business Understanding, Data Understanding, Data Preparation 단계를 거쳤다. 본 보고서에는 전처리한 데이터를 바탕으로 CRISP-DM의 Modeling, Evaluation을 진행한다. Modeling 단계에서는 Modeling 기법을 선택하고 Overfitting에 유의하며 Model의 성능을 높이기 위한 작업을 실시해야 한다. Evaluation 단계에서는 Training set과 Test Dataset으로 나눈 것을 바탕으로 Test Dataset을 예측하고, 실제 비즈니스 상황에서 작동할 수 있는지 확인한다. 또한 Confusion matrix를 이용하여 ACC, Recall, Precision을 계산해 볼 수 있다.

-Modeling 단계에서 다음 6가지 작업을 수행했다. Feature selection(Filter approach, ElasticNet, forward/backward), Model Validation, RandomForestClassifier, Identify most profitable segments

- 모델의 학습 속도를 높이고 Overfitting을 방지하기 위해 먼저 Feature selection을 진행했다. attribute와 target 간의 상관계수 계산을 통해 연관이 적은 attribute를 찾을 수 있었고 Modeling 시에 제외하는 것을 고려하였다.

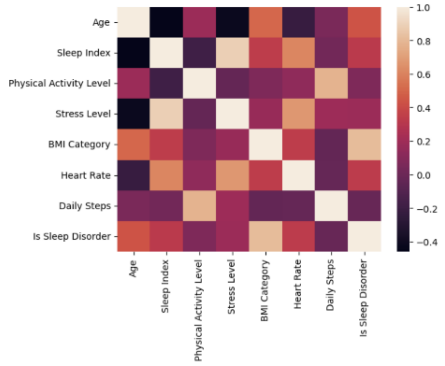
-Model의 성능을 높이기 위해 Feature selection 방법 중 embedded method에 해당하는 ElasticNet을 사용했다. ElasticNet의 hyperparameter을 통해 max_iter와 alpha의 최적의 값을 찾고 도출된 값을 토대로 feature selection을 진행했다. 다음으로 feature selection 기법 중 wrapper method에 속하는 forward, backward selection을 ElasticNet과의 결과와 비교해 보았다. 이때 sweet spot의 성능이 좋아지고 평균적인 accuracy가 향상되어 modeling을 하기에 적절한 것을 확인할 수 있었다.

-Modeling 이전에는 최적의 hyperparameter를 찾을 수 있기에 Model Validation을 활용했다. cross validation의 몇 가지 기법들을 이용해 feature selection 과정 이후 정제된 dataset을 모델링하는데 적합한 최적의 hyperparameter를 찾아보았다. Model Validation으로 hyperparameter tuning해도 성능이 향상되지 않았기 때문에, 수동적 turning의 한계점을 깨닫고 DecisionTree를 앙상블 기법으로 성능을 향상시킨 RandomForestClassifier를 도입했다. hyperparameter tuning을 진행했는데, overfitting 방지를 위해 holdout validation과 early stopping criteria도 도입하여 min_samples_split과 min_samples_leaf도 함께 tuning하였다. 이에 따라 일반화 성능이 향상되었다. 마지막으로 Identify most profitable segments를 통해 프로젝트가 집중해야 할 segment를 구할 수 있었다.

-Evaluation 단계에서는 먼저 training set과 test set으로 나눈 것을 바탕으로 Modeling 단계에서 제작한 Decision Tree 모델을 이용하여 테스트 데이터셋을 직접 예측했다. 또한 최적의 모델을 찾기 위해 Decision Tree에서 Fitting Graph를 활용해 모델 complexity의 node 수에 따라 증가하는 depth를 늘려가며 train acc와 test acc를 비교했다.

-ROC-AUC Curve를 그려 평가를 진행했다. 그래프를 통해 모델이 hold-out data set에서도 좋은 성능을 내며 작동함을 알 수 있었다. 그리고 예측 결과의 Confusion Matrix를 그려 높은 정확도를 확인할 수 있었고 Recall 과 Precision도 계산해 모델이 Recall 과 Precision 중 어디에 집중해야 하는지 찾아냈다.

Modeling: Feature selection-Filter approach



Feature selection 을 진행함으로써 target 과 연관성이 높은 attribute 만을 사용하여 모델의 학습 속도를 높이고, over-fitting 도 방지할 수 있다. Feature selection 의 여러 방법 중 모델링 전에 진행하는 filter approach 를 사용했다. 이를 위해서는 attribute 와 target 간의 상관계수를 알아야 해, 상관계수를 계산하였고, 그 결과를 히트맵을 활용하여 연관성이 높은 순으로 시각화하였다.

그림1 Target과 attribute들 간의 heatmap

상관계수를 계산한 결과, 'Is sleep disorder'(수면장애 여부)가 약 0.8 로 target attribute 인 BMI category 와 가장 강한 linear relationship 을, 'physical activity level', 'daily steps'는 상대적으로 약한 linear relationship 을 보임을 확인했다. 하지만 modeling 단계에서 사용될 decision tree 는 non-linear 한 관계도 예측할 수 있는 기법이기에 최종적으로 모델링에 사용될 feature attribute 들을 고르기 위해서는 추가적인 분석 및 여러 모델들을 시험해 보는 과정이 필요하다.

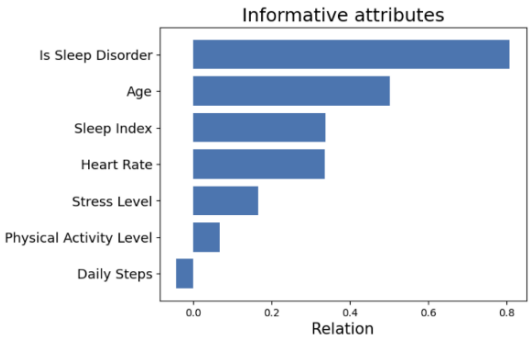
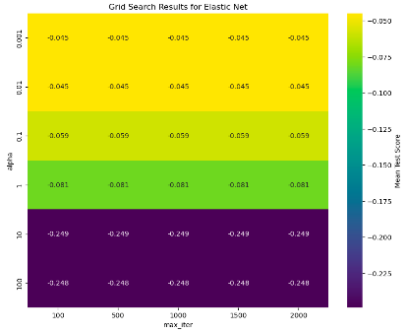
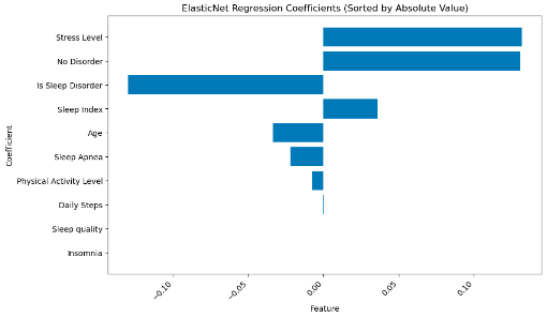


그림2 각 Attribute별 Information gain

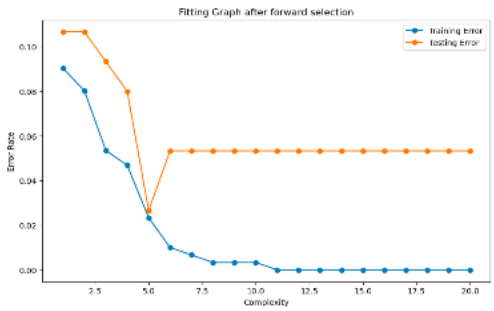
Modeling: Feature Selection-ElasticNet

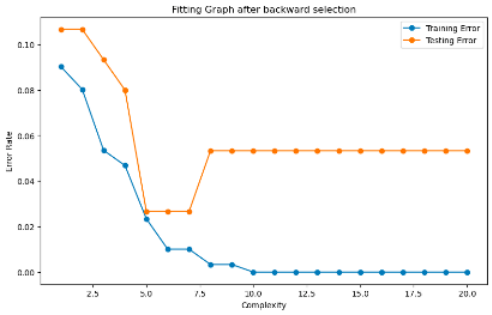




모델의 성능을 높이기 위한 feature selection 방법 중 embedded method 에 해당하는 elasticNet 을 사용하였다. 사용하는 dataset 의 경우 sample 수가 많지 않기 때문에 적용하기에 적합하다고 판단하였다. 실제로 수행하기 전에 GridSearchCV 를 통해 ElasticNet 의 hyperparameter 인 max_iter 와 alpha 의 최적 값을 찾았고, 도출된 값을 토대로 feature selection 을 진행하였다. 해당 feature 들의 coefficient 값들을 비교하여 유의미하다고 판단되는 상위 5 개의 feature 들만을 선택하였고 최종적으로 fitting graph 를 그려본 결과 기초적인 모델링만을 수행하였을 때 보다 평균적으로 성능이 향상된 것을 확인할 수 있었다.

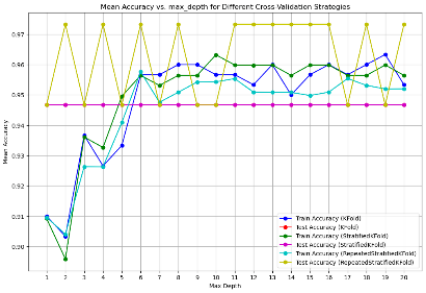
Modeling: Feature Selection-forward, backward



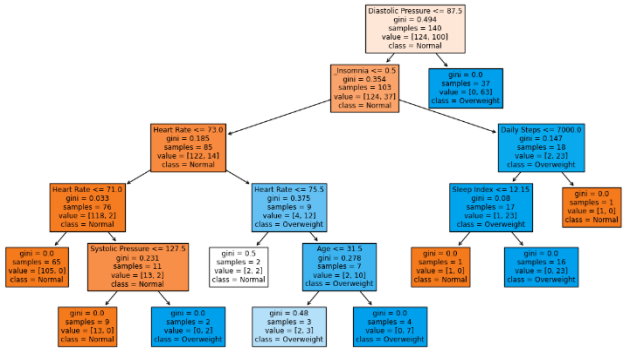


feature selection 기법 중 wrapper method 에 속하는 forward, backward selection 도 모델의 성능을 높일 수 있는 기법이기 때문에 진행해 보았다. 결론적으로 elasticNet 을 적용했을 때와 비교하면, sweet spot 일 때의 성능은 좋아졌다. backward selection 을 수행하였을 때는 평균적인 accuracy 가 향상되어 modeling 을 하기에 가장 적합한 것으로 판단했다.

Modeling: Model Validation



Model validation 은 최적의 hyperparameter 를 찾을 수 있기에 modeling 이전에 활용해 보았다. 그 중 cross validation 의 몇 가지 기법들을 활용하여 feature selection 과정 이후 정제된 dataset 을 모델링하는 데 적합한 최적의 hyperparameter 를 찾아보았다. k-fold, StratifiedKFold, RepeatedStratifiedKFold 방식을 적용했고, StratifiedKFold 을 적용하였을 때 가장 좋은 성능을 보였지만, 성능은 개선되지 않아서 해당 값을 활용하기는 어렵다고 판단했다.



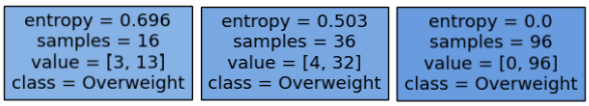
```
검증 세트 정확도: 0.97
하이퍼파라미터 튜닝 후 최적 모델: RandomForestClassifier(max_depth=10, min_samples_split=6, n_estim
ators=130,
                       random_state=42)
최적 파라미터: {'n_estimators': 130, 'min_samples_split': 6, 'min_samples_leaf': 1, 'max_dept
h': 10}
```

Modeling: RandomForestClassifier

Model Validation 으로 hyperparameter tuning 을 하여도 성능이 향상되지는 않았다. 수동적인 tuning 으로는 한계가 있다고 판단했고, 모델링 기법 자체를 변경할 필요가 있어 보였다. 그래서 DecisionTree 를 앙상블 기법으로 성능을 향상시킨 RandomForest 를 도입해 보았다. 이를 더 효율적으로 활용하기 위해 RandomizedSearchCV 로 n_estimators 를 포함한 hyperparameter tuning 을 먼저

진행하였고, overfitting 방지를 위해 holdout validation 과 early stopping criteria 를 도입하였고, 그 과정에서 min_samples_split 과 min_samples_leaf 를 활용하여 tuning 을 진행했다. 결과적으로 training set 에 대한 accuracy 는 눈에 띄게 좋아짐과 동시에 overfitting 도 발생하지 않았고 일반화 성능도 좋아졌다. 다만 모델링의 feature attribute 으로 활용된 불면증 여부는 추후에 손목닥터 9988 에서 측정되는 수면 중 깨어 있는 시간, 얕은 수면 시간, 깊은 수면 시간, 이 3 가지의 지표를 활용하여 적절하게 반영해야 할 것이다.

Modeling: Identify most profitable segments



해당 프로젝트에서 목표로 해야 할 segment 는 위의 3 개와 같이 class 가 'overweight'로 분류되는 segment 들이다. 이를 통해 수면 습관 개선, 적정 수준의 운동 등을 통해 과체중인 사람을 정상체중이 될 수 있게 유도할 수 있다. overweight segment 중에서도 overweight 비율이 더 높은 segment 를 선택하는 것이 목적 달성에 유리하다고 판단했고, 위의 세가지 segment 의 overweight 비율을 계산하면 아래와 같다.

	Normal instance 수	Overweight instance 수	Overweight 비율
Segment 1	3	13	81.25%
Segment 2	4	32	88.88%
Segment 3	0	96	100%

표 1 Segment 별 overweight 비율

Segment 3, 2, 1 순으로 overweight 비율이 높은 것을 확인했고, 프로젝트 투자 비용을 효율적으로 활용하기 위해서는 segment 3 에 집중하는 것이 효율적일 수 있다. 비용은 체중 감량 목표 달성 시 상품권 등 보상 제공, 비만 관리 프로그램 비용 지원 등에 활용할 수 있다. 예시에서는 비율 차이가 크지 않아 투자 효율이 크게 차이 나지 않을 수 있지만, 비율 차이가 더 나는 경우에는 투자 비용에 대한 가치를 극대화할 수 있을 것이다.

Evaluation: Training set/Test set split

제작한 모델의 평가를 위하여, 전체 데이터셋의 25%를 테스트 데이터셋으로 분리하였다.

N BMI Category	count
Normal	49
Overweight	45

분리한 테스트 데이터셋의 target data 의 distribution 을 살펴보니, Normal label 이 49 개, Overweight 인 label 이 45 개로 비슷하게 분포되어 있는 데이터셋인 것을 확인할 수 있다. 그렇기 때문에 도출된 모델로 BMI category 를 normal 로

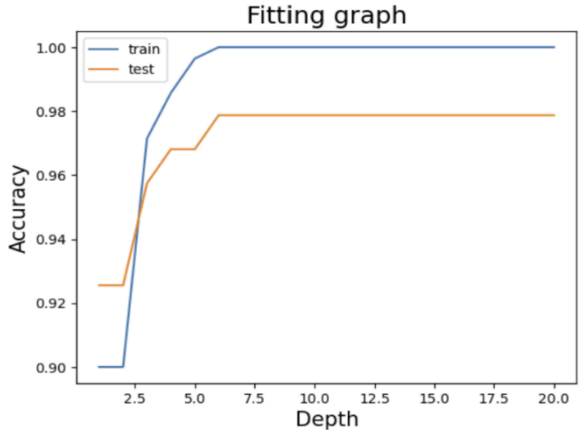
Main text (4/5)

구별해 낼 때와 overweight 으로 구별해 낼 때 양쪽 모두의 성능을 충분히 살펴볼 수 있을 것으로 예상된다.

위의 Modeling 단계에서 제작한 Decision Tree 모델을 테스트 데이터셋에 직접 적용해 보았다.

Evaluation: Fitting Graph

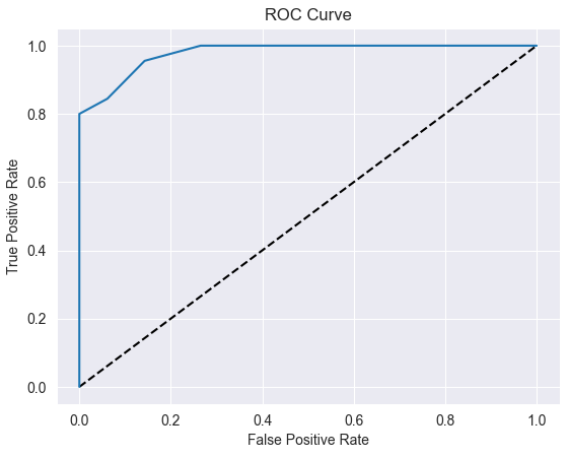
최적의 모델을 찾기 위해서 모델의 complexity 를 증가시키면서 training data 와 test data 의 accuracy 를 비교해 보았다. 사용한 모델에서 complexity 는 node 의 수라고 할 수 있다. 따라서, node 의 수에 따라 증가하는 depth 를 늘려가면서 training data 와 test data 의 accuracy 를 비교해 보았다.



Fitting graph 를 확인한 결과, 초기에는 Depth 가 증가함에 따라 training data 와 test data 의 accuracy 가 모두 증가함을 알 수 있다. 그러나 Depth 가 6 이상부터는 training data 의 accuracy 가 1.00 이 되어 overfitting 이 의심되고, test data 의 accuracy 또한 유의미하게 증가하지 않았다. 따라서 모델의 최대 Depth 를 5 로 설정함으로써 overfitting 을 방지하였다.

```
best_dt = tree.DecisionTreeClassifier(criterion = "entropy", max_depth = 5)
```

Evaluation: ROC-AUC Curve

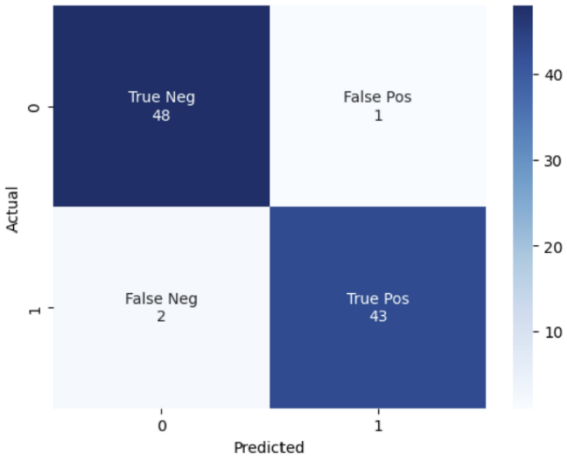


또한, ROC-AUC Curve 를 그려 평가했고, AUC 점수가 0.978 로, 높은 점수를 보이고 있다. 해당 그래프로 미루어 보아, 이 모델은 hold-out data set 을 기준으로 봤을 때는 'normal' 과 'overweight' 사이의 상대적인 차이를 잘 포착할 수 있음을 확인할 수 있다. 그러나 실제 현장에서 이 모델을 그대로 적용시킨다면 원하는 만큼의 정확도가 나오지 않을 수 있어, 실제 현장에서 추가적으로 얻게 되는 데이터를 바탕으로 분석에 사용되는 feature attribute 을 변경하거나 모델의 세부적인 부분을 추가적으로 조정하는 과정이 필수적이다.

Main text (5/5)

Evaluation: Confusion Matrix

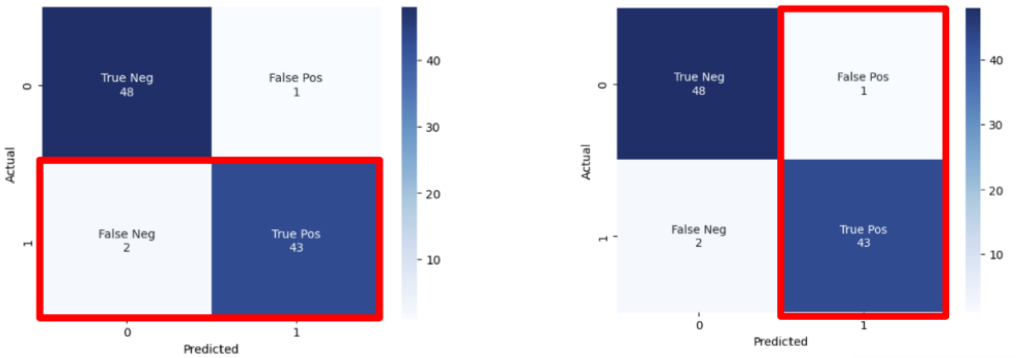
예측 결과의 Confusion Matrix 는 우측의 그림과 같다. Overweight label 을 Overweight 로 예측하는 것을 True Positive 라고 하면, True Positive 는 43, True Negative 는 48 이었다. 우선 Acc 을 계산해보면 (48+43)/(48+43+2+1) = 97%로, 상당히 높은 정확도를 가짐을 확인할 수 있다.



Evaluation: Recall and Precision

Confusion matrix 를 통해 Recall 과 Precision 도 계산해 볼 수 있다.

Recall = TP/(TP+FN) = 43/(43+2) = 96% Precision = TP/(TP+FP) = 43/(43+1) = 98%



$$F_\beta = (1+\beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

이 모델은 비만인 사람들을 구별해 내고 그들에게 적절한 가이드라인을 제공함으로써 사회 전반의 건강을 증진시키는 것이 최종 목표이므로 비만인 사람들을 최대한 많이 비만으로 분류하는 것이 중요하다. 따라서 실제로 과체중인 사람들 중 얼마나 많은 사람을 비만으로 구분할 수 있는지를 나타내는 recall 을 precision 보다 더 중요한 지표로 판단하여 활용할 필요가 있다. 그러므로 추후에 모델의 성능 평가를 위해 F-score 를 활용할 때 recall 에 더 가중치를 두는 F2-score 를 사용하는 것이 좋을 것이다.

Individual Contribution #1 (Name: 김나리)

1) Best Final Report Summary 작성

2) 서브 1팀 Report의 Summary, Team self-evaluation 작성

3) 메인 팀장으로 회의 진행, 일정 관리, 전체 파일 관리

How to check your contribution (1 page)

Best Final Report 중 작성한 Summary 부분입니다.

Summary (1 page)

-앞서 Business Understanding, Data Understanding, Data Preparation 단계를 거쳤다. 본 보고서에는 전처리한 데이터를 바탕으로 CRISP-DM의 Modeling, Evaluation을 진행한다. Modeling 단계에서는 Modeling 기법을 선택하고 Overfitting에 유의하며 Model의 성능을 높이기 위한 작업을 실시해야 한다. Evaluation 단계에서는 Training set과 Test Dataset으로 나눈 것을 바탕으로 Test Dataset을 예측하고, 실제 비즈니스 상황에서 작동할 수 있는지 확인한다. 또한 Confusion matrix를 이용하여 ACC, Recall, Precision을 계산해 볼 수 있다.

-Modeling 단계에서 다음 6가지 작업을 수행했다. Feature selection(Filter approach, ElasticNet, forward/backward), Model Validation, RandomForestClassifier, Identify most profitable segments

- 모델의 학습 속도를 높이고 Overfitting을 방지하기 위해 먼저 Feature selection을 진행했다. attribute와 target 간의 상관계수 계산을 통해 연관이 적은 attribute를 찾을 수 있었고 Modeling 시에 제외하는 것을 고려하였다.

-Model의 성능을 높이기 위해 Feature selection 방법 중 embedded method에 해당하는 ElasticNet을 사용했다. ElasticNet의 hyperparameter을 통해 max_iter와 alpha의 최적의 값을 찾고 도출된 값을 토대로 feature selection을 진행했다. 다음으로 feature selection 기법 중 wrapper method에 속하는 forward, backward selection을 ElasticNet과의 결과와 비교해 보았다. 이때 sweet spot의 성능이 좋아지고 평균적인 accuracy가 향상되어 modeling을 하기에 적절한 것을 확인할 수 있었다.

-Modeling 이전에는 최적의 hyperparameter를 찾을 수 있기에 Model Validation을 활용했다. cross validation의 몇 가지 기법들을 이용해 feature selection 과정 이후 정제된 dataset을 모델링하는데 적합한 최적의 hyperparameter를 찾아보았다. Model Validation으로 hyperparameter tuning해도 성능이 향상되지 않았기 때문에, 수동적 turning의 한계점을 깨닫고 DecisionTree를 앙상블 기법으로 성능을 향상시킨 RandomForestClassifier를 도입했다. hyperparameter tuning을 진행했는데, overfitting 방지를 위해 holdout validation과 early stopping criteria도 도입하여 min_samples_split과 min_samples_leaf도 함께 tuning하였다. 이에 따라 일반화 성능이 향상되었다. 마지막으로 Identify most profitable segments를 통해 프로젝트가 집중해야 할 segment를 구할 수 있었다.

-Evaluation 단계에서는 먼저 training set과 test set으로 나눈 것을 바탕으로 Modeling 단계에서 제작한 Decision Tree 모델을 이용하여 테스트 데이터셋을 직접 예측했다. 또한 최적의 모델을 찾기 위해 Decision Tree에서 Fitting Graph를 활용해 모델 complexity의 node 수에 따라 증가하는 depth를 늘려가며 train acc와 test acc를 비교했다.

-ROC-AUC Curve를 그려 평가를 진행했다. 그래프를 통해 모델이 hold-out data set에서도 좋은 성능을 내며 작동함을 알 수 있었다. 그리고 예측 결과의 Confusion Matrix를 그려 높은 정확도를 확인할 수 있었고 Recall 과 Precision도 계산해 모델이 Recall 과 Precision 중 어디에 집중해야 하는지 찾아냈다.

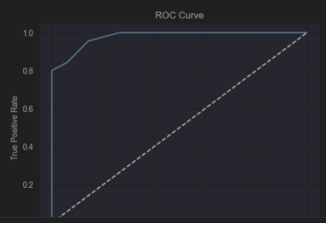
Individual Contribution #2 (Name: 김동규)

- Evaluation 부분을 맡아 작성
- Confusion Matrix, accuracy , recall, precision, F1 score 도출
- ROC-AUC graph 제작
- 전반적인 Evaluation 보고서 작성

How to check your contribution (1 page)

```
In 43
from sklearn import metrics
# Compute predicted probabilities: y_pred_prob
y_pred_prob = clf.predict_proba(x_test)[:,1]

from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
y_test_le = le.fit_transform(y_test)

# Generate ROC curve values: fpr, tpr, thresholds
fpr, tpr, thresholds = metrics.roc_curve(y_test_le, y_pred_prob)

# Compute and print AUC score
print("AUC: {}".format(metrics.auc(fpr, tpr)))

# Plot ROC curve
plt.plot([0, 1], [0, 1], 'k--')
plt.plot(fpr, tpr)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.show()
Executed at 2023.11.28 21:05:00 in 78ms
```

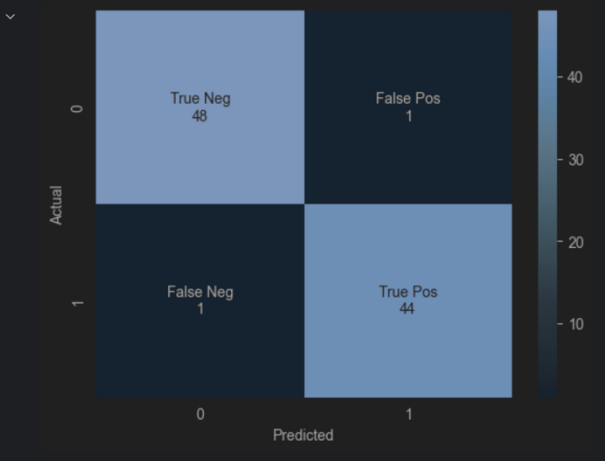
```
AUC: 0.978231292517006
/Users/jeffrey/PycharmProjects/data_science_project/venv/lib/python3.9/site-packages/sklearn/pre
passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example
  y = column_or_1d(y, warn=True)
```




```
In 30
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score

print('Accuracy:  ', round(accuracy_score(y_test, y_pred), 2))
print('Precision: ', round(precision_score(y_test, y_pred, pos_label='Overweight'),2))
print('Recall:    ', round(recall_score(y_test, y_pred, pos_label='Overweight'),2))
print('F1_score:  ', round(f1_score(y_test, y_pred, pos_label='Overweight'),2))
Executed at 2023.11.28 20:56:01 in 6ms
```

```
Accuracy:  0.98
Precision:  0.98
Recall:  0.98
F1_score:  0.98
```

```
In 25
import numpy as np

group_names = ['True Neg','False Pos','False Neg','True Pos']
group_counts = ["{0:0.0f}".format(value) for value in
                cm.flatten()]

labels = [f"{v1}\n{v2}" for v1, v2 in
          zip(group_names,group_counts)]

labels = np.asarray(labels).reshape(2,2)

sns.heatmap(cm, annot=labels, fmt='', cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
Executed at 2023.11.28 20:30:53 in 82ms
```

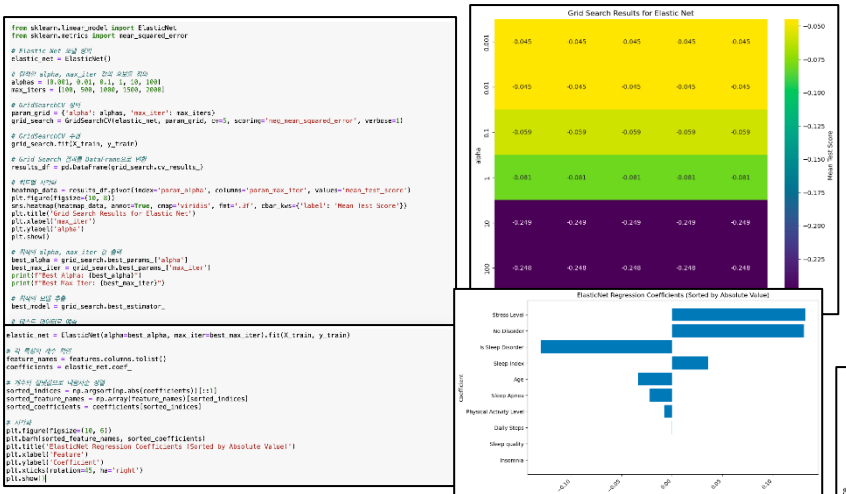


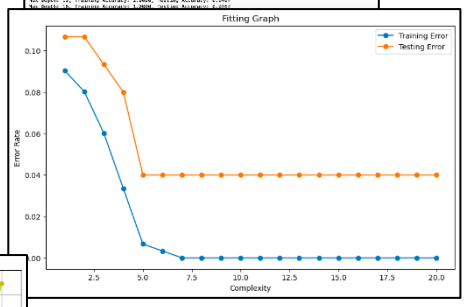
Individual Contribution #3 (Name: 이건)

1. Modeling 단계 일부 수행
    a. feature selection 진행
        i. embedded method-elasticNet을 적용하기 위해 GridSearchCV로 hyperparameter tuning 이후 feature selection을 진행하여 feature간의 상관계수 파악
        ii. wrapper method-forward, backward selection과 elasticNet의 fitting graph를 비교하여 더 적합한 방식을 적용
    b. Hyper Parameter tuning
        i. model validation 중 cross validation의 기법인 k-fold 방식을 통해 최적의 max_depth 값을 도출
        ii. RandomizedSearchCV를 통해 RandomForestClassifier의 hyperparameters(n_estimators, min_samples_split, min_samples_leaf, max_depth)의 최적의 값들을 찾음
    c. RandomForest
        i. RandomForest를 적용시켜 모델의 성능을 향상시킴
        ii. CCP Pruning을 시도해보았지만 tree 구조가 변화되지 않아 일반화 성능이 한계에 도달했다는 것을 확인

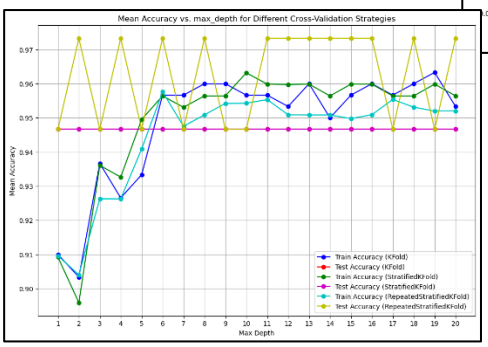
How to check your contribution (1 page)

1. feature selection

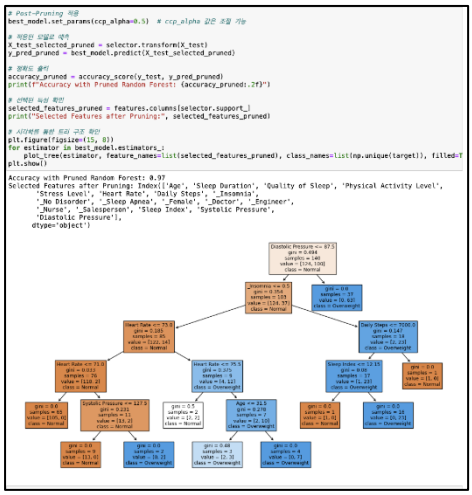




2. hyper parameter tuning



3. RandomForest



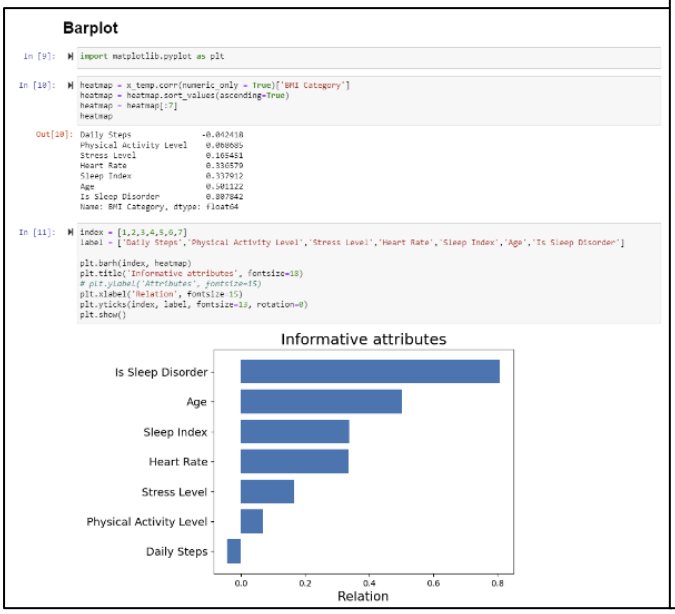
Individual Contribution #4 (Name: 한지훈)

1. Modeling 단계 task 일부 수행
    a. Feature selection - Filter approach
        - Heatmap을 통해 attribute들의 상관계수를 계산하고, target과의 연관이 적은 attribute를 제거하여 overfitting 방지
    b. Identify most profitable segments
        - Segment의 overweight 비율 계산을 통해 투자 효율을 극대화할 수 있는 segment를 확인
2. Evaluation 단계 task 일부 수행
    a. Fitting graph
        - Decision Tree의 Depth를 바꿔가면서 Fitting graph를 plot하여 최적을 결과를 만들어내는 모델 생성
    b. Recall, Precision 비교
        - Confusion Matrix에서 Recall, Precision을 계산하고, 프로젝트 모델에서는 두 값 중 비중을 RecalI에 둬야 한다는 것 확인
3. PPT 내용 일부 제작

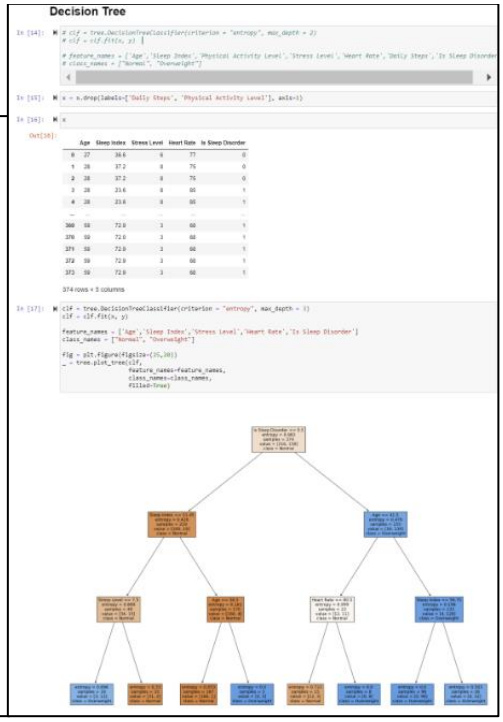
    - Modeling/Evaluation 에 대한 제 구현 부분을 PPT로 제작하였습니다.

How to check your contribution (1 page)

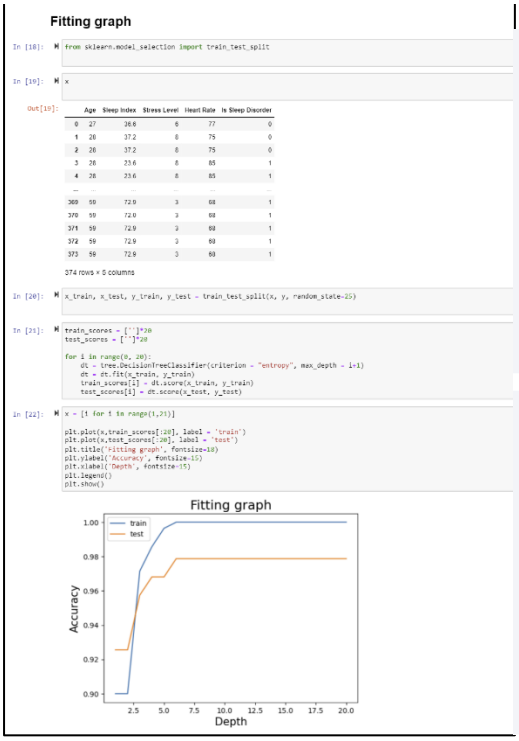
1. Feature selection - Filter approach



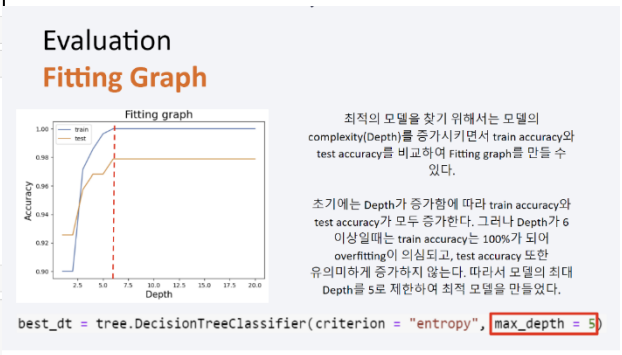
2. Identify most profitable segments

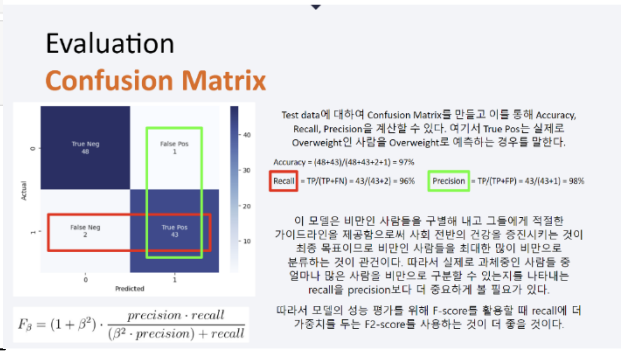


3. Fitting graph



4. PPT 내용 일부 제작(일부 슬라이드)





Individual Contribution #5 (Name: 김거륜)

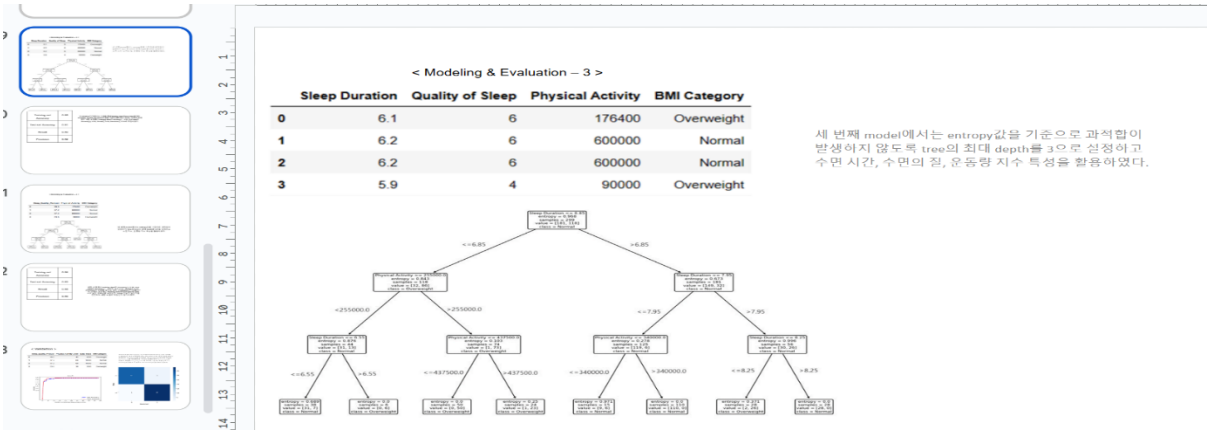
How to check your contribution (1 page)

Individual Contribution #6 (Name: 김태완)

1. PPT 내용 일부 제작

   - Modeling/Evaluation 에 대한 부분을 보고서를 참고하여 PPT로 정리하였다. 최종 ppt의 참고자료로 쓰였다.

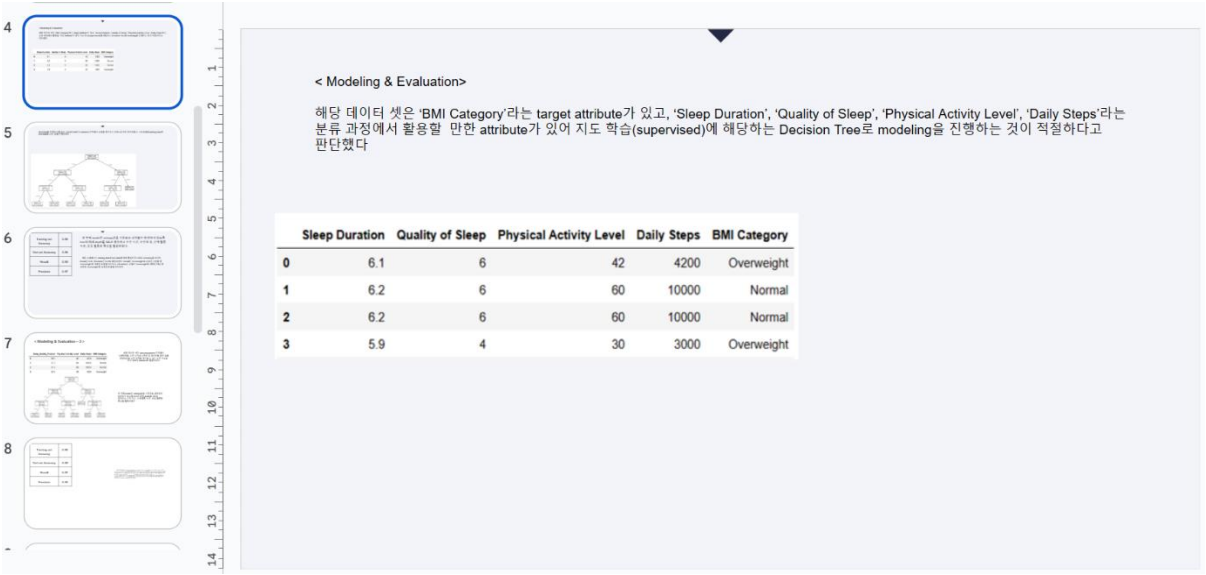
How to check your contribution (1 page)



< Modeling & Evaluation – 3 >

	Sleep Duration	Quality of Sleep	Physical Activity	BMI Category
0	6.1	6	176400	Overweight
1	6.2	6	600000	Normal
2	6.2	6	600000	Normal
3	5.9	4	90000	Overweight

세 번째 model에서는 entropy값을 기준으로 과적합이 발생하지 않도록 tree의 최대 depth를 3으로 설정하고 수면 시간, 수면의 질, 운동량 지수 특성을 활용하였다.



< Modeling & Evaluation>

해당 데이터 셋은 'BMI Category'라는 target attribute가 있고, 'Sleep Duration', 'Quality of Sleep', 'Physical Activity Level', 'Daily Steps'라는 분류 과정에서 활용할 만한 attribute가 있어 지도 학습(supervised)에 해당하는 Decision Tree로 modeling을 진행하는 것이 적절하다고 판단했다

	Sleep Duration	Quality of Sleep	Physical Activity Level	Daily Steps	BMI Category
0	6.1	6	42	4200	Overweight
1	6.2	6	60	10000	Normal
2	6.2	6	60	10000	Normal
3	5.9	4	30	3000	Overweight

Individual Contribution #7 (Name: 서유정)

1. 서브팀 Final report summary 작성

   팀원들이 작성한 모델링과 평가 부분을 바탕으로 CRISP-DM 각 단계에 해당하는 부분으로 나누어 summary를 작성하였다.

2. 서브팀 self-evaluation을 작성

   evaluation의 각 항목에 맞게 보고서의 충실성, 구체성, 타당성, 완성도, 산업 기여도를 작성하였다.

3. 최종 ppt 제작

   완성된 보고서와 내용 기반PPT를 중심으로 하여 최종본 PPT를 제작하였다.

How to check your contribution (1 page)

Summary (1 page)

'Sleep Health and Lifestyle Dataset'을 활용한 비만 예측 모델링 프로세스

1. 데이터 이해와 준비

-데이터 셋에는 다양한 특성이 포함되어 있다. CRISP-DM 프로세스에 따라 필요한 특성을 분류해서 남기고, 새로운 특성을 도입하여 모델의 예측력을 향상시킨다.

2. 모델링과 평가

-해당 데이터 셋은 'BMI Category'라는 target attribute가 있고 'Sleep Duration', 'Quality of Sleep', 'Physical Activity Level', 'Daily Steps'라는 분류 과정에서 활용할 만한 attribute가 있어 지도 학습(supervised)에 해당하는 Decision Tree로 modeling을 진행하는 것이 적절하다고 판단했다.

-모델의 평가는 training data와 test data에 대한 Accuracy, Recall, Precision의 값을 구해 진행하였고, 이 프로젝트는 비만인 사람들을 위해 적절한 가이드라인을 제공함으로써 삶의 변화를 이끌어내는 것이 목적이므로 가능한 'Overweight'인 사람들을 모두 'Overweight'로 분류하는 것이 중요하다. 따라서, Evaluation 단계에서 Precision보다 Recall이 더 중요한 평가 지표라고 판단하였다.

3. 모델 최적화

-특성 조합을 통해 새로운 지표인 수면 지수와 운동량 지수를 도입하여 모델을 최적화한다.

-entropy 값을 기준으로 max leaf nodes 값 조절과 early stopping을 활용하여 과적합을 방지한다.

4. 결과 선택 및 검증

-네 가지 모델 평가하였을 때 training data 와 test data에 대해 모두 높은 Accuracy를 보이며 프로젝트의 목표에 부합하게 Recall이 높게 측정된 두 번째 모델을 최적의 모델로 선택하였다.

-두 번째 모델에 대해 추가적인 검증 단계를 진행하였다. 해당 모델에서 max leaf nodes 값을 2~200까지 차례대로 Decision Tree를 구성하고, 각 Decision Tree에 대해 각각의 Test Accuracy를 측정하였다. 그 결과 모델의 max leaf nodes의 크기가 11일 때, 가장 높은 Test Accuracy가 나왔고 그 후로는 Test Accuracy가 증가하지 않았다.

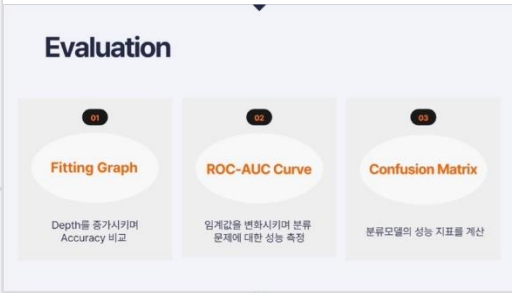
5. 모델 적용과 주의사항

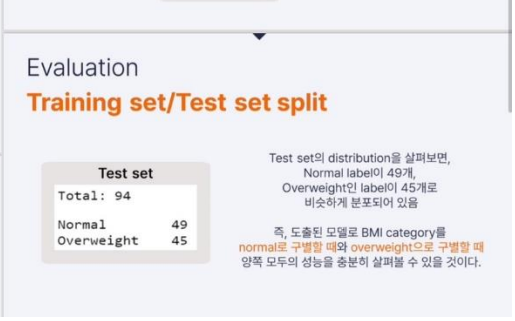
-training-set accuracy와 test-set accuracy가 너무 높은 점을 감안했을 때 이 모델을 실제 현장에 적용시킨다면 다른 결과가 나올 수 있다. 이런 높은 accuracy가 도출된 원인은 비교적 작은 데이터 셋의 규모나 데이터 셋이 모집단의 분포를 충분히 반영하고 있지 않은 가능성 등이 예상된다.

-손목 닥터 9988의 데이터에 이 모델을 적용시킨다면 추가적으로 얻는 데이터를 통해 모델링을 다시 실시하고, 최대 depth나 feature attribute 등의 조정을 통해 모델의 일반화 성능 향상을 고려해야 한다.

Team Self-Evaluation

항목	점수	사유
충실성 (20)	20	CRISP-DM의 프로세스를 기반으로 하여 데이터 마이닝의 각 단계를 체계적으로 수행하고 있다. Data Understanding, Data Preparation, Modeling, Evaulation 등의 단계를 명확하게 구분하고 각각의 작업을 설명하고 있다. 의사결정 나무 기법을 사용하여 최적의 모델을 찾기 위해 여러 모델들을 만들어보고 비교하는 과정을 담고 있다. Modeling을 진행하기 전에 데이터를 8:2 비율로 나누어 training data와 test data로 나누어 진행하여 검증을 위한 hold-out data를 만들었다.
구체성 (20)	19	모델링에서 사용된 특성과 평가 지표들이 어떠한 과정을 거쳐 발생되었는지 그에 대한 과정이 명확히 보고서에 작성돼 있다. Target attribute가 있다는 점에 기인하여 지도학습에 해당하는 Decision Tree로 모델링을 진행하는 것에 대한 이유를 서술하고 있다. 1특성 조합을 통해 새로운 지표인 수면 지수와 운동량 지수를 도입하여 모델을 최적화하는 과정을 포함한다.
타당성 (20)	18	모델링에서는 다양한 특성을 선택하였고, 손목 닥터 9988에서 수집 가능한 데이터들을 기반으로 모델링을 만들었다. 또한, 각 모델에 대한 정확도, 재현율, 정밀도 등의 평가 지표를 활용하여 결과를 검증하고 있다. 최종 선택된 최적의 모델에 대해서는 여러 max leaf nodes 값들 중에 가장 높은 test accuracy를 확인하는 등 추가적인 검증을 거쳐 결과의 타당성을 확인할 수 있다.
완성도 (20)	19	데이터를 선정하는 과정부터 추가적인 검증까지의 내용을 인과적으로 설명하고 있고, 한 가지의 모델만을 만드는 것이 아닌 여러 가지 모델을 만들어보고 그 중 프로젝트 목적에 잘 부합하는 최적의 모델을 선택하여 완성도를 높였다. 또한, 손목 닥터 9988에 적용될 때 생길 수 있는 문제점을 고려하며 일반화를 위해 개선해야 할 점이 무엇인지에 대해 서술하고 있다.
산업/학계/사회 기여도 (20)	19	비만과 관련된 요인들을 분석하고 모델링 하여 비만과 관련된 요인을 모델링하고 평가를 내렸다. 이 프로젝트는 비만인 사람들을 위해 적절한 가이드라인을 제공함으로써 삶의 변화를 이끌어는 목적을 가지고 있다. 그에 따라 Evaluation 단계에서 Recall을 더 중요하게 평가하며 프로젝트 목적을 달성하려 한다. 따라서 보고서에 나온 모델링을 적용시키고 일반화를 위한 개선점들을 고쳐나간다면 비만 관리에 큰 도움이 될 수 있을 것이다.





Individual Contribution #8 (Name: 선신욱)

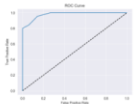
1. Evaluation 단계 기여

   - 현재 프로젝트에서는 Evaluation 단계에서 recall이 precision보다 더 중요한 지표임을 언급한 내용이 구체적으로 보고서에 포함되어 있다.

2. Best proposal에서 보고서 전체적인 내용 수정 및 자체 평가 작성

   - Best proposal의 초안을 보고서의 조건에 맞게 수정하고 글의 흐름을 자연스럽게 하기 위해 문장 수정 및 맞춤법 수정을 했다. 또한, 조원들과의 토의 끝에 종합된 의견을 바탕으로 자체 평가를 작성하였다.

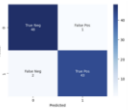
How to check your contribution (1 page)



또한, ROC-AUC Curve 를 그려 평가했고, AUC 점수가 0.978 로, 높은 점수를 보이고 있다. 해당 그래프로 미루어 보아, 이 모델은 hold-out data set 을 기준으로 봤을 때는 'normal' 과 'overweight' 사이의 상대적인 차이를 잘 포착할 수 있음을 확인할 수 있다. 그러나 실제 현장에서 이 모델을 그대로 적용시킨다면 원하는 만큼의 정확도가 나오지 않을 수 있어, 실제 현장에서 추가적으로 얻게 되는 데이터를 바탕으로 분석에 사용되는 feature attribute 을 변경하거나 모델의 세부적인 부분을 추가적으로 조정하는 과정이 필수적이다.

Main text (5/5)

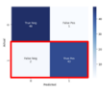
Evaluation: Confusion Matrix

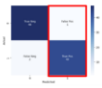


예측 결과의 Confusion Matrix 는 우측의 그림과 같다. Overweight 을 Overweight 로 예측하는 것을 True Positive 라고 하면, True Positive 는 43, True Negative 는 48 이었다. 우선 Acc 을 계산해보면 (48+43)/(48+43+2+1) = 97%로, 상당히 높은 정확도를 가짐을 확인할 수 있다.

Evaluation: Recall and Precision

Confusion matrix 를 통해 Recall 과 Precision 도 계산해 볼 수 있다.





Recall = TP/(TP+FN) = 43/(43+2) = 96%

Precision = TP/(TP+FP) = 43/(43+1) = 98%

$$F_\beta = (1+\beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

이 모델은 비만인 사람들을 구별해 내고 그들에게 적절한 가이드라인을 제공함으로써 사회 전반의 건강을 증진시키는 것이 최종 목표이므로 비만인 사람들을 최대한 많이 비만으로 분류하는 것이 중요하다. 따라서 실제로 과체중인 사람들 중 얼마나 많은 사람을 비만으로 구분할 수 있는지를 나타내는 recall 을 precision 보다 더 중요한 지표로 판단하여 활용할 필요가 있다. 그러므로 추후에 모델의 성능 평가를 위해 F-score 를 활용할 때 recall 에 더 가중치를 두는 F2-score 를 사용하는 것이 좋을 것이다.

Team Self-Evaluation

항목	점수	사유
충실성 (20)	20	CRISP-DM 단계 중 보고서에 나타나야 하는 Modeling, Evaluation & Validation 단계의 핵심 과정들이 포함되어 있다. Feature selection의 여러 방법(Filter approach/ElasticNet/Forward, Backward)를 활용해 모델의 학습 속도를 높이고 overfitting을 방지하며 성능을 향상시킬 수 있는 방안을 고려했다. 또한 최적의 hyperparameter를 찾아 모델 성능을 향상시켰고, DecisionTree를 앙상블 기법으로 성능을 향상시킨 RandomForest를 도입하는 등 다양한 방법을 통해 목적을 달성하였다.
구체성 (20)	19	진행 과정을 히트맵, 그래프 등을 활용하여 시각화하였다. Evaluation 단계에서는 Fitting graph를 만들고, Depth 변화에 따른 training data와 test data의 accuracy 변화를 알아보았다. 또한 ROC-AUC Curve를 통해 test data의 성능을 확인해 보았으며, confusion matrix를 활용하여 recall, precision 값을 확인했다.
타당성 (20)	19	본 보고서는 다양한 방법을 통해 모델의 성능을 높이고자 했다. 수동적인 hyperparameter tuning의 한계점을 파악하고 모델링 기법을 변경하는 것이 좋다고 생각해 DecisionTree를 앙상블 기법으로 성능을 향상시킨 RandomForest를 도입하는 등 구체적인 작업을 통해 입증한 내용을 바탕으로 모델을 개선하는 모습도 보였다. Evaluation 단계에서도 overfitting 관련 성능을 보았고 test data에서도 모델이 잘 작동함을 확인할 수 있었다. 구체적인 과정들과 그에 대한 적절한 근거를 가지고 Modeling 단계를 수행했고, Evaluation에서도 모델의 성능이 적절하다고 평가되는 것으로 보아 이번 모델이 타당한 편이라고 볼 수 있다.
완성도 (20)	19	구체적인 Modeling과 Evaluation 과정을 거쳤다. 그 과정에서 Feature selection의 여러 가지 방법을 활용해 모델의 성능을 높이기 위한 고민도, Decision Tree에서 더 발전시킨 RandomForest도 포함되어 있다. 모델 생성과 성능 향상, Evaluation 내용과 추가적으로 실제 현장에 도입하였을 때의 고려해야 할 부분까지 자세하게 서술되어 있어 높은 완성도를 가졌음을 알 수 있다.
산업/학계/사회 기여도 (20)	19	본 보고서의 궁극적인 목표는 수면개선을 통한 비만관리이고, 시민들이 손목닥터 9988 등 스마트 헬스케어를 통해 자신의 비만 정도를 알고 개선할 수 있게 하는 것이다. 따라서 본 보고서의 충실성, 구체성, 타당성, 완성도에 포함된 모든 내용이 좋을수록 산업/학계/사회 기여도를 높여준다고 할 수 있다. 모델의 성능이 좋아야, 목표에 부합할 가능성이 높아지기 때문이다. 특히 이번 단계에서 overweight segment 중 프로젝트에서 더 집중해야 하는 주요 segment도 찾아내, 구체적인 비용 투자도 고려하였다.

위의 사진과 같이 best proposal의 전체적인 내용과 자체 평가에 대해 추가적인 내용 추가 및 수정을 진행하였다.

Individual Contribution #9 (Name: 성정윤)

1. Filter approach 항목에서 잘못된 내용 수정
- 상관 계수가 낮게 나왔으므로 target attribute과 큰 관련이 없다는 내용이 잘못되었음을 지적
- 상관 계수 하나만으로 두 attribute 사이에 관계가 없다고 결론을 내릴 수는 없음을 지적
- linear한 관계만 보이지 않을 뿐 non-linear한 관계는 충분히 존재할 수 있다는 내용으로 수정
2. RandomForestClassifer 항목에서 도출된 모델에 대한 설명에서 부족한 부분 추가
- 손목 닥터 9988 에서는 직접적으로 측정되지 않는 불면증 여부를 간접적으로 분석할 수 있는 방법에 대해 설명
3. Evaluation: Training set/Test set split 항목에서 부족한 설명 추가
- test set 의 분포가 어떤 점에서 유리한지에 대한 설명 추가
4. Evaluation: ROC-AUC Curve 항목에서 높은 AUC 에 대한 내용 중 잘못된 부분 수정
- 높은 AUC 가 가지는 의미에 대한 잘못된 이해를 수정
- 실제 현장에서도 높은 성능을 보이며 작동할 것이라는 내용을 지우고, AUC 의 실질적 의미와 한계에 대해 설명
5. Evaluation: Confusion Matrix 항목에서 불필요한 부분 삭제
- 강의 교재에 등장하는 Confusion Matrix 와 항목들의 위치가 다른 점에 대한 설명을 생략
6. Evaluation: Recall and Precision 항목에서 부족한 내용 추가
- F2-Score 가 모델을 평가하는 데 더 적합할 것이라는 결론의 근거 추가
7. 보고서 Main 항목의 맞춤법 교정 및 모호한 부분 수정

How to check your contribution (1 page)

수정하기 전 원본과 수정한 내용을 캡처한 사진들 (밑줄 친 빨간 부분이 수정 전 내용, 빨간 글씨가 직접 수정한 내용)

상관계수를 계산해본 결과, 'Is sleep disorder'(수면장애 여부)가 약 0.8로 target인 BMI category와 가장 연관되어있음을 알 수 있다. 반면 'Physical activity level', 'Daily steps'는 상대적으로 target과 큰 관련이 없음을 알 수 있다. 이런 연관이 적은 attribute는 모델링 시에 제외하는 것을 고려해볼 수 있다.

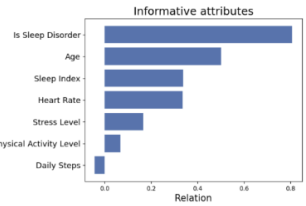


그림2 각 Attribute별 Information gain

강한 linear relationship을 보임을 알 수 있다. 반면 'physical activity level', 'daily steps'는 상대적으로 약한 linear relationship을 보인다. 하지만 이번 연구에서 사용될 decision tree는 non-linear한 관계도 예측할 수 있는 기법이기에 최종적인 모델링에 사용될 feature attribute들을 고르기 위해서는 추가적인 분석이나 직접 여러 모델들을 시험해 보는 것이 필요하다.

방지를 위해 holdout validation과 early stopping criteria도 도입하여 min_samples_split과 min_samples_leaf도 함께 tuning 하였다. 결과적으로 training set에 대한 accuracy는 눈에 띄게 좋아짐과 동시에 overfitting도 되지

Evaluation: Training set/Test set split

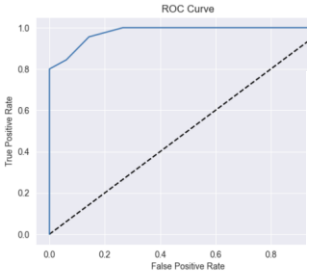
제작한 모델의 평가를 위하여, 전체 데이터셋의 25%를 테스트

분리한 테스트 데이터셋의 target data의 distribution을 살펴보았더니, Normal label이 49개, Overweight인 label이 45개로 비슷하게 분포되어 있는 데이터셋인 것을 알 수 있다. 그렇기 때문에 도출된 모델로 BMI category를 normal로 구별해 낼 때와 overweight으로 구별해 낼 때 양쪽 모두의 성능을 충분히 살펴볼 수 있을 것이다.

않아서 일반화 성능 또한 좋아졌다. 모델링의 feature attribute으로 사용된 불면증 여부는 추후에 손목닥터 9988에서 측정되는 수면 중 깨어 있는 시간, 얕은 수면 시간, 깊은 수면 시간, 이 3가지의 지표를 활용하여 적절하게 반영해야 할 것이다.

Evaluation: ROC-AUC Curve

또한, ROC-AUC Curve를 그려 평가해 보았다. AUC 점수가 0.978으로, 높은 점수를 나타내고 있다.



해당 그래프로 미루어 보아, 최종적으로 모델링이 훌륭하게 진행되어 train 상황 뿐 아니라 실제 비즈니스 상황을 가정한 테스트 데이터셋에서도 높은 성능을 보이며 작동한다는 사실을 확인할 수 있다.

이 모델은 hold-out data set을 기준으로 봤을 때는 'normal' 과 'overweight' 사이의 상대적인 차이를 잘 포착할 수 있음을 확인할 수 있다. 그러나 실제 현장에서 이 모델을 그대로 적용시킨다면 원하는 만큼의 정확도가 나오지 않을 수 있다. 따라서 실제 현장에서 추가적으로 얻게 되는 데이터를

우리의 모델은 비만 관리 솔루션을 제공하기 위해 과체중인 사람을 최대한 많이 찾는 것이 관건이라고 할 수 있다. 즉 Recall 과 Precision 중에서는 p(실제 과체중인 사람) 탐지를 얼마나 잘하는지를 묻는 Recall을 높이는 것이 적절하다.

$$F_\beta = (1+\beta^2)\cdot\frac{precision\cdot recall}{(\beta^2\cdot precision)+recall}$$

따라서 F-measure 계산시에 Recall을 4배 더 중요시하는 F2 값을 사용하는 것이 좋을 것이다. (B = 2)

바탕으로 분석에 사용되는 feature attribute을 변경하거나 모델의 세부적인 부분을 추가적으로 조정하는 것이 필수적이다.

이 모델은 비만인 사람들을 구별해 내고 그들에게 적절한 가이드라인을 제공함으로써 사회 전반의 건강을 증진시키는 것이 최종 목표이므로 비만인 사람들을 최대한 많이 비만으로 분류하는 것이 관건이다. 따라서 실제로 과체중인 사람들 중 얼마나 많은 사람을 비만으로 구분할 수 있는지를 나타내는 recall이 precision보다 더 중요하게 볼 필요가 있다. 그러므로 모델의 성능 평가를 위해 F-score를 활용할 때 recall에 더 가중치를 두는 F2-score를 사용하는 것이 더 좋을 것이다.

Individual Contribution #10 (Name: 유창현)

1. 발표 영상 제작

- 최종 보고서 및 최종 PPT 발표자료 기반 발표 영상 녹화

- 녹화 영상 편집하여 발표 영상 제작

How to check your contribution (1 page)

1 조 발표 영상 참고

Team Self-Evaluation

항목	점수	사유
충실성 (20)	20	CRISP-DM 단계 중 보고서에 나타나야 하는 Modeling, Evaluation & Validation 단계의 핵심 과정들이 포함되어 있다. Feature selection의 여러 방법(Filter approach/ElasticNet/Forward, Backward)을 활용해 모델의 학습 속도를 높이고 overfitting을 방지하며 성능을 향상시킬 수 있는 방안을 고려했다. 또한 최적의 hyperparameter을 찾아 모델 성능을 향상시켰고, DecisionTree를 앙상블 기법으로 성능을 향상시킨 RandomForest을 도입하는 등 다양한 방법을 통해 목적을 달성하였다.
구체성 (20)	19	진행 과정을 히트맵, 그래프 등을 활용하여 시각화하였다. Evaluation 단계에서는 Fitting graph를 만들고, Depth 변화에 따른 training data와 test data의 accuracy 변화를 알아보았다. 또한 ROC-AUC Curve를 통해 test data의 성능을 확인해 보았으며, confusion matrix를 활용하여 recall, precision 값을 확인했다.
타당성 (20)	19	본 보고서는 다양한 방법을 통해 모델의 성능을 높이고자 했다. 수동적인 hyperparameter tuning의 한계점을 파악하고 모델링 기법을 변경하는 것이 좋다고 생각해 DecisionTree를 앙상블 기법으로 성능을 향상시킨 RandomForest를 도입하는 등 구체적인 작업을 통해 입증한 내용을 바탕으로 모델을 개선하는 모습도 보였다. Evaluation 단계에서도 overfitting 관련 성능을 보았고 test data에서도 모델이 잘 작동함을 확인할 수 있었다. 구체적인 과정들과 그에 대한 적절한 근거를 가지고 Modeling 단계를 수행했고, Evaluation에서도 모델의 성능이 적절하다고 평가되는 것으로 보아 이번 모델이 타당한 편이라고 볼 수 있다.
완성도 (20)	19	구체적인 Modeling과 Evaluation 과정을 거쳤다. 그 과정에서 Feature selection의 여러 가지 방법을 활용해 모델의 성능을 높이기 위한 고민도, Decision Tree에서 더 발전시킨 RandomForest도 포함되어 있다. 모델 생성과 성능 향상, Evaluation 내용과 추가적으로 실제 현장에 도입하였을 때의 고려해야 할 부분까지 자세하게 서술되어 있어 높은 완성도를 가졌음을 알 수 있다.
산업/학계/사회 기여도 (20)	19	본 보고서의 궁극적인 목표는 수면개선을 통한 비만관리이고, 시민들이 손목닥터 9988 등 스마트 헬스케어를 통해 자신의 비만 정도를 알고 개선할 수 있게 하는 것이다. 따라서 본 보고서의 충실성, 구체성, 타당성, 완성도에 포함된 모든 내용이 좋을수록 산업/학계/사회 기여도를 높여준다고 할 수 있다. 모델의 성능이 좋아야, 목표에 부합할 가능성이 높아지기 때문이다. 특히 이번 단계에서 overweight segment 중 프로젝트에서 더 집중해야 하는 주요 segment도 찾아내, 구체적인 비용 투자도 고려하였다.