LG U+ 프로젝트 데이터 모델링 보고서

Step4-1. 데이터 셋 구성 및 전처리 (Diabetes)

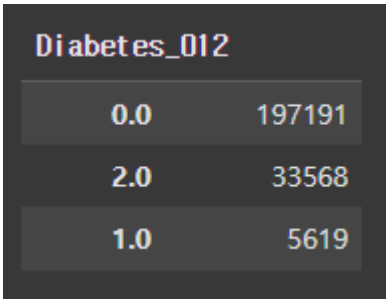
▼ 당뇨 위험군 분류 데이터

1. Data: Diabetes Health Indicators Dataset (kaggle)

2.Data Info확인

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 236378 entries, 0 to 236377
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          236378 non-null  float64
 1   HighBP                236378 non-null  int64
 2   HighChol              236378 non-null  float64
 3   CholCheck             236378 non-null  int64
 4   BMI                   236378 non-null  float64
 5   Smoker                236378 non-null  float64
 6   Stroke                236378 non-null  float64
 7   HeartDiseaseorAttack  236378 non-null  float64
 8   PhysActivity          236378 non-null  int64
 9   Fruits                236378 non-null  int64
 10  Veggies               236378 non-null  int64
 11  HvyAlcoholConsump     236378 non-null  int64
 12  AnyHealthcare         236378 non-null  int64
 13  NoDocbcCost           236378 non-null  float64
 14  GenHlth               236378 non-null  float64
 15  MentHlth              236378 non-null  float64
 16  PhysHlth              236378 non-null  float64
 17  DiffWalk              236378 non-null  float64
 18  Sex                   236378 non-null  int64
 19  Age                   236378 non-null  int64
 20  Education             236378 non-null  float64
 21  Income                236378 non-null  float64
dtypes: float64(13), int64(9)
```

3. data Imbalance 확인

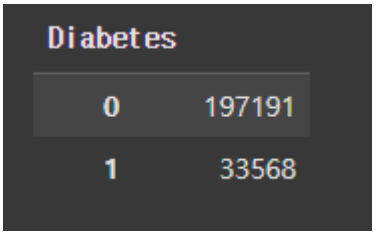


Diabetes_012	
0.0	197191
2.0	33568
1.0	5619

4. 데이터 전처리 과정

- 데이터 컬럼 이름 정리: 원활한 작업을 위해 알아보기 쉬운 컬럼명 으로 변경 (예시: HvyAlcoholConsump을 Alco)

- Diabetes_012 컬럼 값 처리: Diabetes_012값이 0,1,2로 구성 되어있는데, 기획한 서비스 내용에 맞게 1을 제거 후, 2를 1로 변경



Diabetes	
0	197191
1	33568

- Smoker, HighChol, Diabetes_012의 데이터 값과 형식은 0/1 (float64)이다. 이를 int값으로 들어가도록 처리

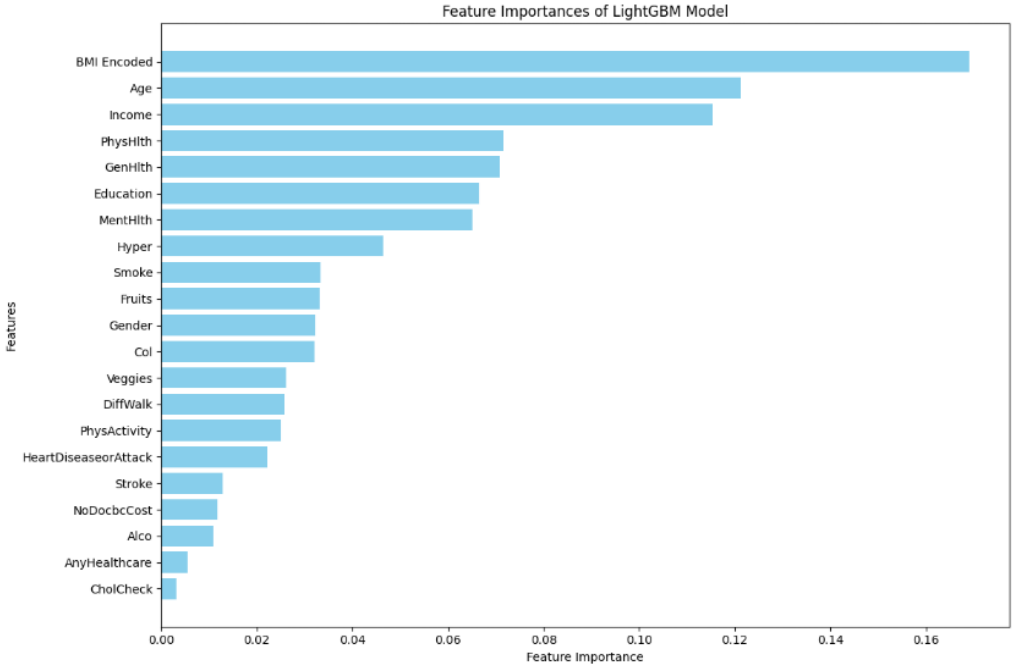
- 데이터 불균형 해결: SMOTE-Tomek을 사용해 데이터 불균형 해결

Step4-2. 데이터 모델링 (Diabetes)

▼ 당뇨 위험군 분류 데이터

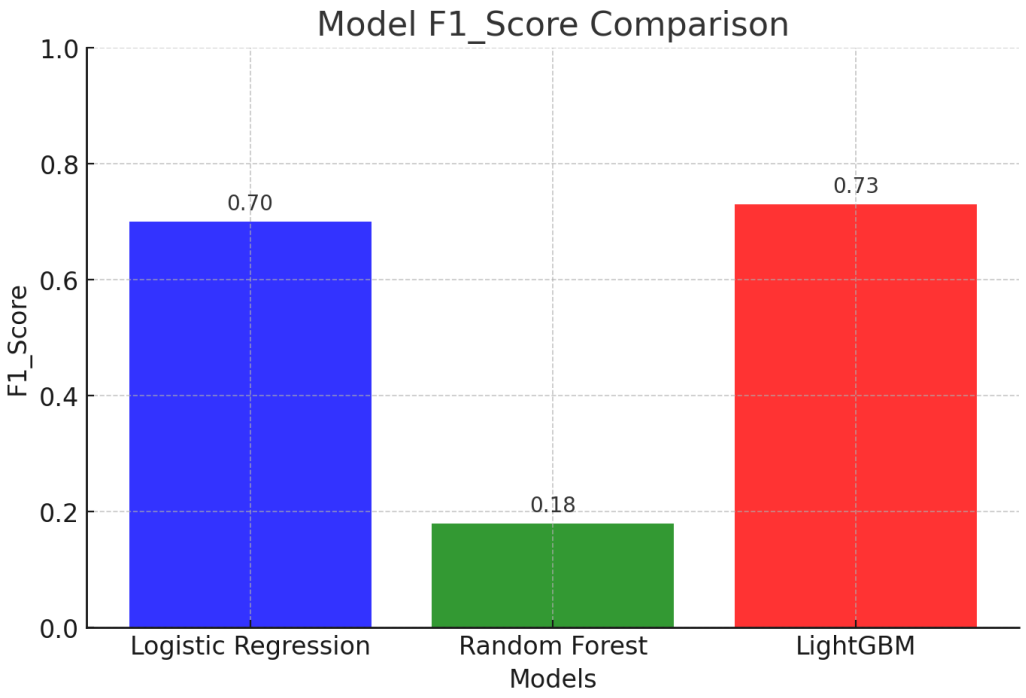
1. 사용 모델: LightGBM
2. 모델링 과정

- 하이퍼 파라미터 튜닝: RandomizedSearchCV을 사용하여 최적의 파라미터 선택
- Feature Selection: Feature Importance를 계산 후 점수가 높은 것과 사용자에게 받기 쉬운 정보 인지를 고려하여 Feature를 선택



- 선택된 Feature: "Gender", "Age", "Hyper", "BMI Encoded", "Smoke", "Col", "Alco"
- 가중치: 선택 변수인 Diabetes의 예측 성능을 개선하기 위해 Diabetes 가중치 조정
- 모델 평가 지표: F1_Score = 0.73

3. 다른 분류 모델과 성능 비교 (F1_Score)



Step5-1 데이터 셋 구성 및 전처리 (lung cancer)

▼ 폐암 위험군 분류 데이터

1. Data: Survey Lung Cancer

2. Data Info확인

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Gender                 309 non-null    object
 1   Age                    309 non-null    int64
 2   Smoke                  309 non-null    int64
 3   YELLOW_FINGERS         309 non-null    int64
 4   ANXIETY                309 non-null    int64
 5   PEER_PRESSURE          309 non-null    int64
 6   CHRONIC DISEASE        309 non-null    int64
 7   Tired                  309 non-null    int64
 8   ALLERGY                309 non-null    int64
 9   WHEEZING               309 non-null    int64
 10  Alco                   309 non-null    int64
 11  COUGHING               309 non-null    int64
 12  SHORTNESS OF BREATH    309 non-null    int64
 13  SWALLOWING DIFFICULTY  309 non-null    int64
 14  CHEST PAIN             309 non-null    int64
 15  LUNG_CANCER            309 non-null    object
dtypes: int64(14), object(2)
```

3. data Imbalance 확인

LUNG_CANCER	count
YES	270
NO	39

4. 데이터 전처리 과정
    - 데이터 컬럼 이름 정리: 원활한 작업을 위해 알아보기 쉬운 컬럼명 으로 변경 (예시: FATIGUE을 Tired)
    - 데이터 공백 제거

- 범주형 변수인 'LUNG_CANCER'와 'Gender' 컬럼에 각각 문자를 이진 숫자 데이터로 처리, 라벨 인코딩 적용

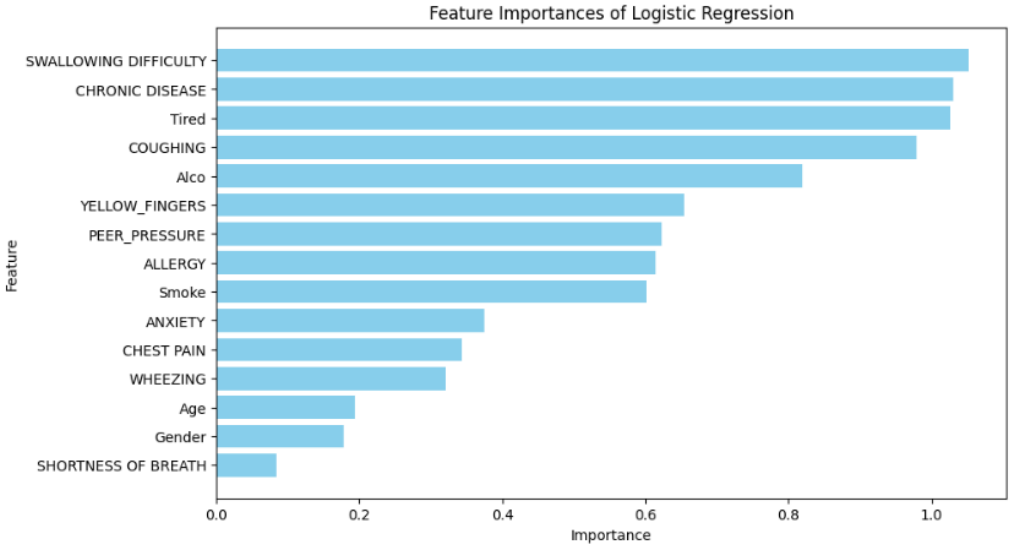
- 데이터 불균형 해결: SMOTE을 사용해 데이터 불균형 해결

Step5-2. 데이터 모델링 (lung cancer)

▼ 폐암 위험군 분류 데이터

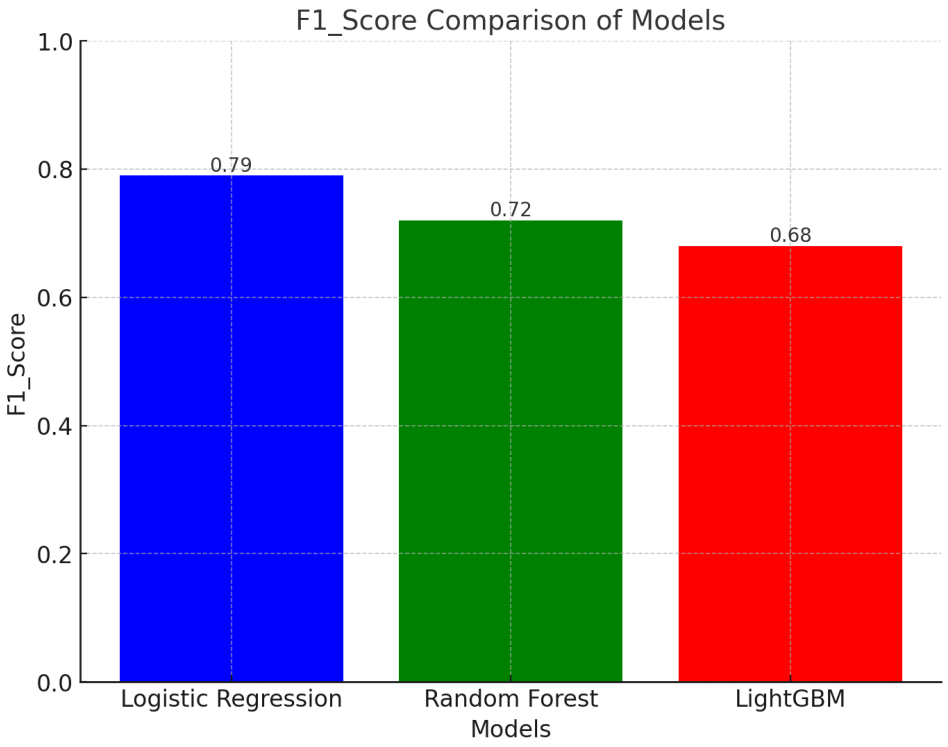
1. 사용 모델: LogisticRegression
2. 모델링 과정

- 하이퍼 파라미터 튜닝: GridSearchCV을 사용하여 최적의 파라미터 선택
- Feature Selection: Feature Importance를 계산 후 점수가 높은 것과 사용자에게 받기 쉬운 정보 인지를 고려하여 Feature를 선택



- 선택된 Feature: 'Gender', 'Age', 'Smoke', 'Tired', 'Alco'
- 모델 평가 지표: F1_Score = 0.79

3. 다른 분류 모델과 성능 비교 (F1_Score)



Step6-1. 데이터 셋 구성 및 전처리 (Liver_disease)

▼ 간암 위험군 분류 데이터

1. Data: Survey Lung Cancer

2. Data Info 확인

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1700 entries, 0 to 1699
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 1700 non-null   int64
 1   Gender              1700 non-null   int64
 2   BMI                 1700 non-null   float64
 3   AlcoholConsumption  1700 non-null   float64
 4   Smoking             1700 non-null   int64
 5   GeneticRisk         1700 non-null   int64
 6   PhysicalActivity    1700 non-null   float64
 7   Diabetes            1700 non-null   int64
 8   Hypertension        1700 non-null   int64
 9   LiverFunctionTest   1700 non-null   float64
 10  Diagnosis           1700 non-null   int64
dtypes: float64(4), int64(7)
```

3. Data Imbalance 확인

Diagnosis	count
1	936
0	764

dtype: int64

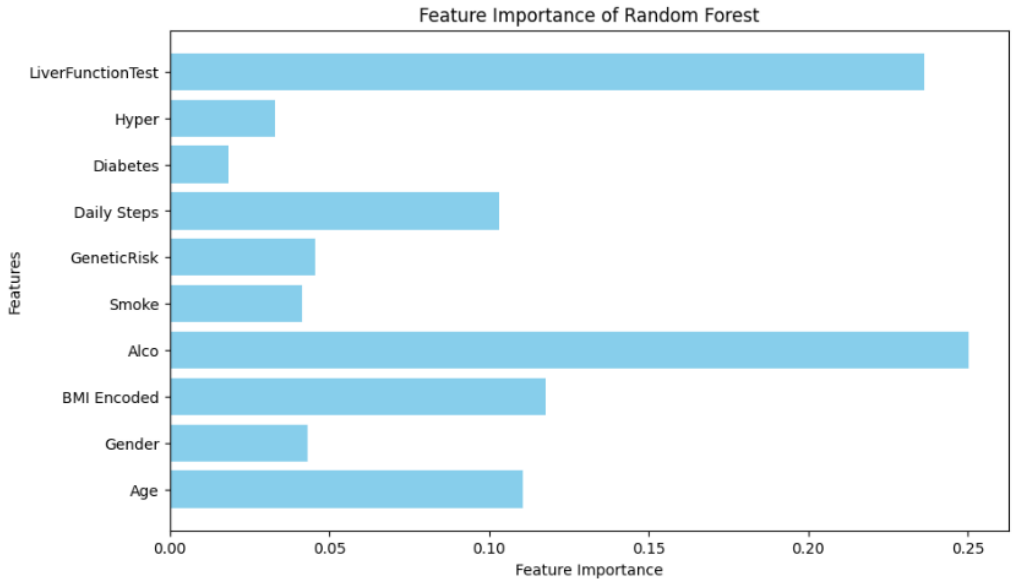
4. 데이터 전처리 과정
    - 데이터 컬럼 이름 정리: 원활한 작업을 위해 알아보기 쉬운 컬럼명 으로 변경 (예시: PhysicalActivity을 Daily Steps)
    - 데이터 공백 제거
    - Alco 컬럼의 값을 이진화

- Daily Steps 컬럼의 값을 신체 활동 기준으로 30분 걷기를 약 3000걸음으로 간주해서 계산, 1~5의 범주형 값으로 변환, int로 변환, Nan 값을 최소값(1) 으로 처리

Step6-2. 데이터 모델링 (Liver_disease)

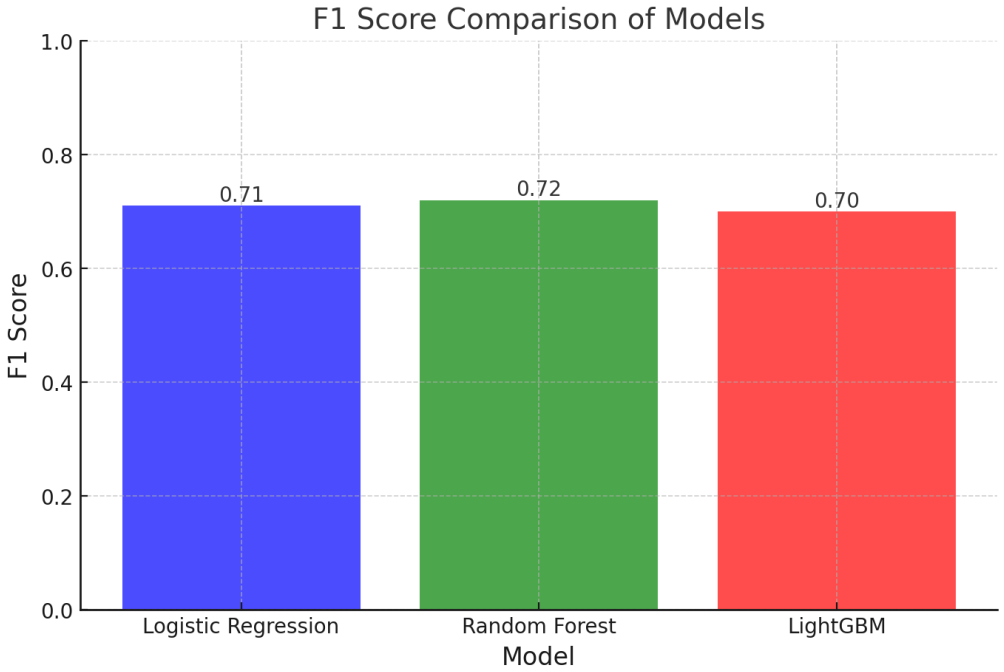
▼ 간암 위험군 분류 데이터

1. 사용 모델: RandomForest
2. 모델링 과정
    - 하이퍼 파라미터 튜닝: GridSearchCV을 사용하여 최적의 파라미터 선택
    - Feature Selection: Feature Importance를 계산 후 점수가 높은 것과 사용자에게 받기 쉬운 정보 인지를 고려하여 Feature를 선택



    - 선택된 Feature: 'Age', 'Gender', 'BMI Encoded', 'Alco', 'Smoke', 'Daily Steps', 'Hyper'
    - 모델 평가 지표: F1_Score = 0.72

3. 다른 분류 모델과 성능 비교 (F1_Score)


F1 Score Comparison of Models
F1 Score
1.0
0.8
0.6
0.4
0.2
0.0
0.71
0.72
0.70
Logistic Regression
Random Forest
LightGBM
Model