코로나19 환자 사망 여부 예측:
데이터 불균형 해소와 모델 개선을 통한 재검토

서울대학교 빅데이터 핀테크 AI 고급 전문가 과정 11기
황정현

초록

본 연구는 Kaggle의 COVID-19 환자 데이터를 활용하여 사망 여부를 예측하는 이진 분류 모델을 구축하고, 기존 분석의 한계를 보완함으로써 성능과 해석력을 향상하는 것을 목표로 한다. 기존 보고서는 CLASSIFFICATION_FINAL(최종 코로나 분류) 변수가 코로나 양성 여부를 직접적으로 제공함으로써 모델 성능을 과대평가하는 한계를 보였다. 이에 따라 본 연구는 (1) 전체 환자 시나리오(코로나 분류 변수 포함·제외 비교)와 (2) 확진자만 시나리오(코로나 분류 변수 제거)를 구분하여 설계하였다. 데이터 불균형 문제는 학습 데이터에 언더샘플링과 오버샘플링을 적용하여 완화하였으며, 로지스틱 회귀와 LightGBM 모델을 비교하였다. 또한 임계값 0.5와 검증셋 기반 최적 임계값을 기준으로 성능을 평가하였다. 분석 결과, ROC-AUC와 PR-AUC에서는 LightGBM이 우수하게 나타났으며, 사망(양성) 클래스에 대한 Recall과 F1-score는 리샘플링과 임계값 조정에서 개선 효과가 확인되었다. 본 연구는 단순 정확도 지표의 한계를 지적하고, 시나리오 분리·불균형 대응·임계값 조정의 필요성을 실증적으로 제시한다.

키워드: 코로나19(COVID-19), 사망 예측, 데이터 불균형, 리샘플링, 로지스틱 회귀, LightGBM

목 차

제 1 장 서 론

제 1 절 연구 배경

코로나바이러스감염증-19(COVID-19)는 전 세계적으로 막대한 피해를 초래하며 보건의료 시스템에 큰 부담을 주었다. 특히 감염 이후 중증으로 진행되거나 사망에 이르는 환자를 조기에 예측하는 것은 의료 자원의 효율적 분배와 환자 치료 전략 수립에 있어 매우 중요한 과제이다. 그러나 실제 임상 현장에서는 환자의 연령, 기저질환, 생활 습관 등 다양한 요인들이 복합적으로 작용하여 사망 여부를 단순히 예측하기 어렵다. 따라서 환자 특성에 기반한 데이터 분석을 통해 사망 위험을 보다 정확하게 추정할 수 있는 모델의 필요성이 제기되고 있다.

제 2 절 기존 연구 소개

COVID-19 사망 예측과 관련된 연구는 다양한 국가와 기관에서 진행되었다. 이들 연구는 임상 정보, 역학적 변수, 환자의 생활 습관 요인을 활용하여 환자의 중증도 및 사망 여부를 분류하는 데 목적을 두었다. 예를 들어, Kaggle 플랫폼에 공개된 한 분석 보고서는 멕시코 보건부에서 제공한 대규모 환자 데이터를 기반으로 로지스틱 회귀 등 기계학습 알고리즘을 적용하여 약 90% 수준의 높은 정확도를 보고하였다. 또한 이 보고서에서는 데이터 불균형 문제를 완화하기 위해 언더샘플링(Random Under Sampling)을 적용하여 사망자 클래스의 분류 성능을 개선하고자 하였다.

제 3 절 기존 연구 검토와 한계

기존 Kaggle 보고서는 COVID-19 환자 사망 예측에 대한 중요한 시도였으나 몇 가지 한계가 존재한다.

첫째, CLASSIFFICATION_FINAL 변수를 그대로 포함하여 모델을 학습시켰다는 점이다. 이 변수는 환자의 코로나 양성 여부(확진/음성/의심)를 직접적으로 알려주는데, 이는 사망 여부와 밀접한 상관관계를 가지므로 모델 성능을 과대평가할 위험이 있다. 실제로 “코로나 확진 여부”라는 정보를 알고 있는 상황에서 사망 여부를 예측하는 것은 임상적 의미가 제한적이다. 둘째, 데이터 불균형 문제에 대응하기 위해 언더샘플링은 적용했으나, 오버샘플링이나 다른

방법과의 비교 검증은 이루어지지 않았다. 이로 인해 데이터 손실과 정보 편향의 가능성이 존재하며, 다른 기법을 적용했을 때의 성능 차이를 확인할 수 없었다. 셋째, 평가 지표를 정확도(Accuracy)에 의존한 것도 한계이다. 불균형 데이터셋에서는 정확도가 높게 나오더라도 실제 사망자 분류 성능(Recall, F1-score)은 저조할 수 있다. 따라서 모델 성능을 종합적으로 판단하기 위해서는 ROC-AUC, PR-AUC, Recall, Precision 등 다양한 지표를 활용할 필요가 있다.

제 4 절 본 연구의 필요성

이러한 한계를 극복하기 위해 본 연구는 다음과 같은 방향으로 연구를 설계하였다.

첫째, 분석 시나리오를 이원화하였다. 전체 환자 데이터를 활용한 경우와 코로나 확진자만을 대상으로 한 경우를 분리하여, CLASSIFFICATION_FINAL 변수의 영향을 제거하고 실제로 의미 있는 사망 위험 인자를 규명하고자 하였다. 둘째, 데이터 불균형 문제를 보다 체계적으로 다루기 위해 언더샘플링뿐 아니라 오버샘플링(Random Over Sampling) 기법까지 추가적으로 적용하여 두 방법을 비교하였다. 이를 통해 학습 데이터에서의 클래스 균형 조정이 성능에 미치는 영향을 정량적으로 확인할 수 있도록 하였다. 셋째, 다양한 알고리즘 비교를 통해 모델별 특성을 검토하였다. 기존 연구가 주로 로지스틱 회귀에 집중했던 것과 달리, 본 연구에서는 로지스틱 회귀(Logistic Regression)와 함께 그래디언트 부스팅 기반의 LightGBM 모델을 적용하여 성능을 비교하였다. 넷째, 다양한 성능 지표와 임계값 조정을 통해 실제 활용 가능성을 높이고자 하였다. 단순히 정확도만이 아닌 ROC-AUC, PR-AUC, Precision, Recall, F1-score를 종합적으로 고려하였으며, 기본 임계값(0.5)뿐만 아니라 검증 데이터셋 기반 최적 임계값을 활용하여 결과를 평가하였다.

따라서 본 연구는 “전체 환자 vs 확진자” 시나리오 분리, 리샘플링 기법 비교, 다양한 알고리즘 적용, 임계값 최적화라는 네 가지 측면에서 기존 연구를 보완한다. 이를 통해 COVID-19 환자의 사망 예측 성능을 개선하고, 임상적으로 해석 가능한 분석 결과를 제시하는 데 그 의의가 있다.

제 2 장 데이터 및 전처리

제 1 절 데이터 출처와 개요

본 연구에서 사용한 데이터는 Kaggle에 공개된 “COVID-19 Dataset”으로, 멕시코 보건부(Secretaría de Salud de México)에서 수집한 환자 정보를 포함하고 있다. 이 데이터셋은 수십만 건 이상의 개별 환자 사례로 구성되어 있으며, 인구통계학적 특성, 기저질환, 생활 습관, 진료 관련 변수, 코로나 확진 여부 및 사망 여부 등을 포함한다. 전체 데이터셋은 약 백만 건 이상의 환자 기록으로 이루어져 있으며, 본 연구에서는 결측 처리와 변수 정제를 거쳐 분석 가능한 최종 데이터셋을 구축하였다. 분석의 목표는 환자의 사망 여부를 예측하는 것이며, 원자료에서 DEATH 변수는 1=사망, 2=생존으로 제공되므로 본 연구에서는 이를 1=사망, 0=생존으로 변환하였다.

제 2 절 변수 체계와 타깃 정의

데이터셋은 크게 인구학적 변수, 기저질환 변수, 생활 습관 변수, 코로나 확진 분류 변수, 그리고 진료 관련 변수로 구분된다. 인구학적 변수에는 나이(AGE), 성별(SEX), 임신 여부(PREGNANT)가 포함되며, 기저질환 변수에는 당뇨(DIABETES), 만성폐쇄성폐질환(COPD), 천식(ASTHMA), 면역억제(INMSUPR), 고혈압(HIPERTENSION), 심혈관질환(CARDIOVASCULAR), 비만(OBESITY), 만성신부전(RENAL_CHRONIC), 기타 질환(OTHER_DISEASE) 등이 있다. 생활 습관 변수에는 흡연 여부(TOBACCO)가 있으며, 코로나 분류 변수인 CLASSIFICATION_FINAL은 1~3은 코로나 양성, 4~7은 음성 혹은 의심으로 구분된다. 진료 관련 변수로는 환자의 입원 여부(PATIENT_TYPE), 진료 기관 단위(MEDICAL_UNIT) 등이 존재한다. 이 중에서 분석 시나리오에 따라 CLASSIFFICATION_FINAL 변수를 포함하거나 제외하였으며, 사망 여부는 최종적으로 예측해야 하는 타깃 변수로 설정하였다.

제 3 절 데이터 품질 문제와 처리

원자료에는 결측치 및 특수 코드 값이 다수 존재한다. 예를 들어 97, 98, 99와 같은 값들은 각각 '정보 없음', '적용 불가', '무응답'을 의미한다. 본 연구에서는 이러한 값을 변수별 맥락에 따라 적절히 처리하였다. 예를 들어 PREGNANT 변수에서 남성은 98로 표기되는데, 이를 '임신하지 않음(2=No)'으로 변환하여 PREGNANT_CLEAN 변수를 새로 생성하였다. 또한 INTUBED, ICU, PATIENT_TYPE과 같이 내생성이 강한 변수들은 예측 성능을 저해할 가능성이 크다고 판단하여 모델링에서 제외하였다. 나머지 결측치와 특수 코드는 상황에 따라 새로운 범주로 통합하거나 드롭하여 데이터 품질을 정제하였다.

제 4 절 전처리 절차

전처리는 라벨 변환, 변수 변환, 인코딩, 스케일링, 데이터 분할의 과정을 거쳤다. 먼저 DEATH 변수는 0과 1로 변환하였으며, 앞서 언급한 PREGNANT_CLEAN 변수를 생성하여 원래의 PREGNANT를 대체하였다. 모든 범주형 변수는 원-핫 인코딩을 적용하여 수치형 변수로 변환하였으며, 이때 모든 범주를 보존하기 위해 drop_first 옵션은 False로 설정하였다. 연속형 변수인 AGE는 StandardScaler를 적용하여 평균 0, 분산 1의 값으로 변환하였는데, 데이터 누수를 방지하기 위해 훈련 데이터에서만 fit한 후 검증 및 테스트 데이터에는 동일한 변환을 적용하였다. 데이터셋은 훈련, 검증, 테스트 세트로 6:2:2 비율로 분할하였으며, stratify 옵션을 사용하여 각 세트에서 사망 여부의 비율이 전체와 동일하게 유지되도록 하였다.

제 5 절 불균형 처리

본 데이터의 중요한 특징 중 하나는 사망 여부의 불균형이다. 원자료에서 사망자보다 생존자가 훨씬 많기 때문에 단순히 모델을 학습하면 사망자를 제대로 분류하지 못할 위험이 존재한다. 이를 보완하기 위해 본 연구에서는 훈련 데이터에 한정하여 리샘플링 기법을 적용하였다. 첫째, 언더샘플링(Random Under Sampling)은 다수 클래스인 생존자를 줄여 사망자와 생존자의 비율을 1대1로 맞추는 방식이다. 둘째, 오버샘플링(Random Over Sampling)은 소수 클래스인 사망자를 중복 생성하여 동일하게 1대1의 비율을 만드는 방법이다. 이러한 리샘플링은 훈련 데이터에서만 수행하였으며, 검증 및 테스트 데이터는 원래의 불균형 분포를 그대로 유지하였다.

이는 실제 배치 환경에서의 데이터 분포를 반영하면서도 학습 단계에서는 불균형 문제를 완화하기 위함이다.

제 3 장 모델링

제 1 절 분석 시나리오 설계

본 연구는 기존 Kaggle 보고서의 한계를 보완하기 위해 세 가지 시나리오로 나누어 모델링을 진행하였다. 첫째, 전체 환자 데이터를 대상으로 한 분석에서는 코로나 확진 여부를 나타내는 CLASSIFFICATION_FINAL 변수를 포함한 경우와 제외한 경우를 모두 비교하였다. 둘째, 코로나 확진자만을 필터링하여 사망 여부를 예측하는 시나리오를 설계하였으며, 이 경우에는 CLASSIFFICATION_FINAL 변수를 제거하여 보다 임상적으로 의미 있는 위험 요인 분석을 수행하였다. 셋째, 데이터 불균형 해소를 위한 개선 실험으로 훈련 데이터에 한정해 언더샘플링과 오버샘플링을 적용하고, 그 성능을 원래 데이터와 비교하였다. 이를 통해 단일 데이터셋에서 여러 시나리오를 설계하고 결과를 종합적으로 비교할 수 있도록 하였다.

제 2 절 알고리즘 선정

모델링에는 두 가지 알고리즘을 사용하였다. 첫째, 로지스틱 회귀(Logistic Regression)는 해석력이 뛰어나고 기본적인 이진 분류 문제에서 널리 활용되는 알고리즘이다. 특히 각 변수의 계수를 통해 사망 여부에 미치는 영향을 직관적으로 확인할 수 있기 때문에, 의학적 해석 가능성이 중요시되는 본 연구의 목적에 적합하다. 둘째, LightGBM은 그라디언트 부스팅 프레임워크 기반의 앙상블 학습 알고리즘으로, 대규모 데이터셋에서 빠른 학습 속도와 높은 예측 성능을 제공한다. LightGBM은 복잡한 변수 간 비선형 관계를 잘 포착할 수 있어, 로지스틱 회귀와 비교함으로써 단순 선형 모델 대비 비선형 모델의 성능 개선 효과를 확인할 수 있도록 하였다.

제 3 절 하이퍼파라미터 튜닝

로지스틱 회귀에서는 규제 형태(L1, L2, ElasticNet)와 규제 강도(C 값)를 주요 하이퍼파라미터로 설정하였다. GridSearchCV를 활용하여 교차 검증(AUC 기준)을 수행하였으며, 최적의 조합을 선택한 후 이를 고정하여 반복 학습을 진행하였다. LightGBM의 경우에는 num_leaves, min_child_samples, subsample, colsample_bytree, reg_lambda 등을 탐색 대상으로 설정하였다. 이 역시 교차 검증 기반의 탐색을 통해 최적의 하이퍼파라미터를 도출하였으며, Early Stopping을 적용하여 과적합을 방지하였다. 최종적으로는 고정된 최적 하이퍼파라미터를 사용하여 모델을 학습하고 평가하였다.

제 4 절 임계값 설정

불균형 데이터의 특성을 고려하여, 단순히 기본값인 0.5 임계값만 사용하는 것은 한계가 있다. 사망자(양성 클래스)의 Recall과 F1-score를 높이는 것이 목표인 만큼, 검증 데이터셋에서 F1-score를 기준으로 최적 임계값을 탐색하여 이를 테스트 평가에도 적용하였다. 따라서 모든 모델은 두 가지 기준에서 평가되었다. 첫째, 기본 임계값(0.5)을 적용하여 일반적인 분류 성능을 확인하였고, 둘째, 검증셋 기반 최적 임계값을 적용하여 Recall과 F1-score의 개선 여부를 검증하였다.

제 5 절 성능 평가 지표

모델의 성능 평가는 단순히 정확도(Accuracy) 하나만으로는 불균형 데이터셋의 특성을 제대로 반영하기 어렵다. 따라서 본 연구에서는 여러 지표를 종합적으로 활용하여 모델의 성능을 다각도로 평가하였다. 우선 ROC-AUC는 모델이 사망자와 생존자를 얼마나 잘 구분하는지를 전반적으로 보여주는 지표로 사용되었다. 그러나 불균형 데이터 상황에서는 ROC-AUC만으로는 충분하지 않기 때문에, 양성 클래스인 사망자 예측 성능을 보다 민감하게 확인할 수 있는 PR-AUC도 함께 평가하였다. 또한 실제 사망자를 놓치지 않고 예측하는 능력을 나타내는 Recall, 예측된 사망자 중 실제로 사망자인 비율을 나타내는 Precision을 지표로 포함하였다. 이 두 지표는 각각 민감도와 정밀도를 의미하며, F1-score는 이 둘의 조화 평균으로 균형 잡힌 성능을 측정하는 데 활용되었다. 마지막으로 혼동 행렬(Confusion Matrix)을 통해 참음성(TN), 거짓양성(FP), 거짓음성(FN), 참양성(TP)의 분포를 확인함으로써 모델이 어떤 유형의 오분류를 주로 범하는지도 함께 검토하였다. 이와 같은 다각도의 평가 체계를 통해 본 연구는 모델별, 시나리오별 성능

차이를 정량적으로 비교하고, 실제 임상적 의사결정 과정에서 유용하게 활용될 수 있는 시사점을 도출하고자 하였다.

제 4 장 결과 분석 – 시나리오 1: 전체 환자 사망 여부 예측

제 1 절 베이스라인 성능

본 절에서는 전체 환자 데이터를 대상으로 리샘플링을 적용하지 않은 상태에서의 모델 성능을 비교하였다(로지스틱 회귀는 class_weight='balanced' 적용, LightGBM은 표준 설정). 로지스틱 회귀와 LightGBM 모두 ROC-AUC 기준으로 양호한 분류 성능을 보였으나, 데이터 불균형의 영향으로 PR-AUC와 F1-score에서는 한계가 확인되었다. 테스트 기준으로 로지스틱 회귀는 ROC-AUC 0.9269, PR-AUC 0.5363, F1-score 0.4860(임계값 0.5)였고, LightGBM은 ROC-AUC 0.9332, PR-AUC 0.5610, F1-score 0.4583으로 나타났다. 두 모델 모두 임계값 0.5에서 Recall이 높은 대신 Precision이 낮은 전형적 양상을 보였으며, 특히 LightGBM은 ROC-AUC와 PR-AUC에서 로지스틱 대비 소폭 우수했다.

표 1. 베이스라인 @0.50 성능 비교

모델명	ROC-AUC	PR-AUC	F1-score	Precision	Recall	ACC
Logistic	0.9269	0.5363	0.4860	0.3412	0.8448	0.8689
LightGBM	0.9332	0.5610	0.4583	0.3095	0.8824	0.8469

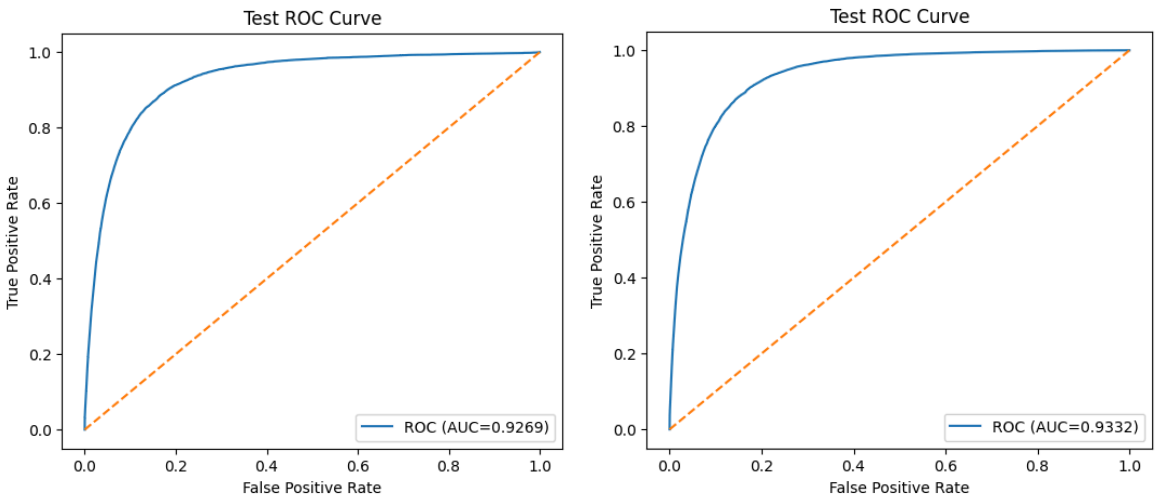

그림 1. ROC 곡선(좌: 로지스틱, 우: LightGBM)

제 2 절 임계값 조정 효과

검증셋에서 F1-score를 최대화하는 최적 임계값을 탐색하여 테스트셋에 적용한 결과, 두 모델 모두 Precision과 F1-score가 유의미하게 향상되었으나, 그 대가로 Recall은 감소하였다. 로지스틱 회귀는 임계값을 0.50에서 0.82로 상향 조정했을 때 F1-score가 0.4860에서 0.5569로 개선되었으며, Precision은 0.3412에서 0.4946으로 크게 증가한 반면 Recall은 0.8448에서 0.6371로 감소하였다. LightGBM 역시 유사한 패턴을 보였으며, 임계값 조정 후 F1-score가 0.4583에서 0.5598로 상승하고 Precision이 0.3095에서 0.4882로 높아졌으나, Recall은 0.8824에서 0.6560으로 감소하였다.

이는 오탐(False Positive)을 줄여 '사망 예측'의 신뢰도를 높이는 대신, 일부 사망자를 더 놓치는(False Negative 증가) 보수적 의사결정으로 해석된다. 따라서 실제 운영 목적이 사망자 누락 최소화(Recall 우선)인지, 불필요 경고 최소화(Precision 우선)인지에 따라 적정 임계값의 선택이 달라질 수 있음을 시사한다.

표 2. 임계값 0.50 vs 최적 임계값(*, F1 기준) 성능 비교

모델명	ROC-AUC	PR-AUC	F1-score	Precision	Recall	ACC
Logistic	0.9269	0.5363	0.4860	0.3412	0.8448	0.8689
Logistic*	0.9269	0.5363	0.5569	0.4946	0.6371	0.9256
LightGBM	0.9332	0.5610	0.4583	0.3095	0.8824	0.8469
LightGBM*	0.9332	0.5610	0.5598	0.4882	0.6560	0.9243

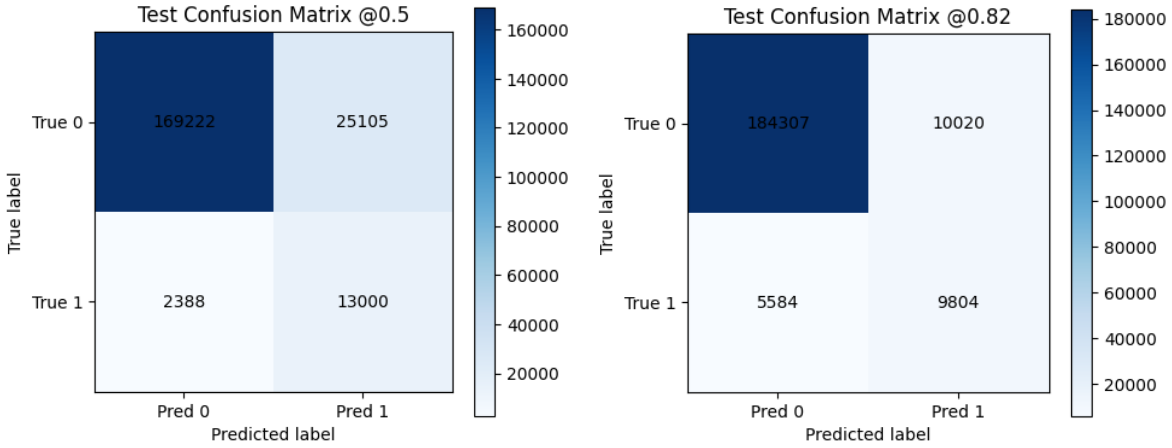

그림 2. 임계값 0.5 vs 최적 임계값의 혼동 행렬 (Logistic)

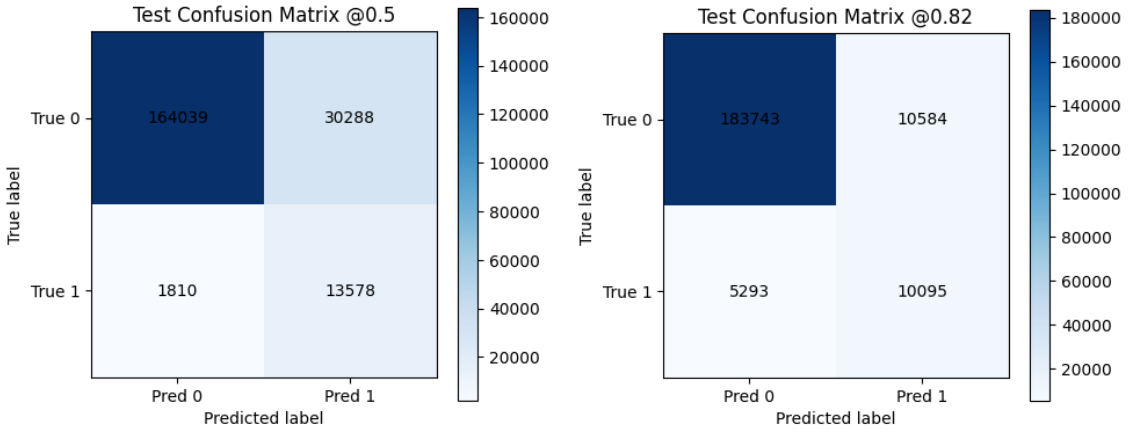

그림 3. 임계값 0.5 vs 최적 임계값의 혼동 행렬 (LightGBM)

제 3 절 리샘플링 적용 결과

훈련 데이터에 한정하여 언더샘플링(Random Under Sampling, RUS)과 오버샘플링(Random Over Sampling, ROS)을 각각 적용하고, 원래 분포의 검증·테스트셋에서 평가하였다. 임계값 0.5 기준 결과, 리샘플링을 적용한 모델들의 F1-score와 PR-AUC는 베이스라인 대비 거의 차이가 없었다. 예를 들어, 로지스틱 회귀는 기본 모델의 F1-score가 0.4860이었으며, RUS 적용 시 0.4850, ROS 적용 시 0.4853으로 사실상 동일한 수준이었다. LightGBM 역시 유사한 경향을 보였으며, ROS 적용 시 Precision이 0.3095에서 0.3108로 소폭 개선되었으나, F1-score는 0.4583에서 0.4595로 미미한 차이에 그쳤다.

이러한 결과는 현재의 파이프라인(AGE 스케일링, 원-핫 인코딩, 적절한 규제 및 부스팅 적용)에서는 리샘플링 자체의 효과가 제한적이며, 오히려 임계값 조정이 성능 개선에 더 직접적임을 시사한다. 다만, 특정 비용 구조(예: 거짓양성 비용이 매우 큰 상황)에서는 오버샘플링과 임계값 조정을 결합한 정책적 접근이 실질적 효과를 가질 수 있다.

표 3. 리샘플링(RUS/ROS) 적용 성능 비교 (Test, thr=0.50)

모델명	ROC-AUC	PR-AUC	F1	Precision	Recall	ACC
Logistic	0.9269	0.5363	0.4860	0.3412	0.8448	0.8689
Logistic(RUS)	0.9267	0.5332	0.4850	0.3400	0.8459	0.8682
Logistic(ROS)	0.9268	0.5361	0.4853	0.3404	0.8451	0.8685
LightGBM	0.9332	0.5610	0.4583	0.3095	0.8824	0.8469
LightGBM(RUS)	0.9327	0.5563	0.4578	0.3090	0.8826	0.8466
LightGBM(ROS)	0.9332	0.5614	0.4595	0.3108	0.8809	0.8479

제 4 절 오류 양상 및 해석

혼동행렬 분석 결과, 베이스라인(@0.5)에서는 거짓양성(FP)이 상대적으로 많이 발생하여 Precision이 낮게 나타났다. 그러나 검증셋에서 도출한 최적 임계값을 적용하면 FP가 감소하면서

Precision이 개선되었다. 반대로 거짓음성(FN)이 증가하여 일부 사망자를 더 놓치는 양상이 나타났는데, 이는 Recall의 감소로 이어졌다. 따라서 임계값 조정은 "사망자 탐지(Recall)보다는 사망 예측의 신뢰도(Precision)를 높이는 보수적 접근"으로 해석할 수 있다.

제 5 절 소결

시나리오 1의 분석을 종합하면, LightGBM이 로지스틱 회귀에 비해 전반적인 분류 성능(ROC-AUC, PR-AUC)에서 우수한 결과를 보였다. 그러나 데이터 불균형 환경에서는 단순히 알고리즘 차이보다 임계값 조정이 성능 개선에 더 직접적인 효과를 제공하였다. 특히 Precision과 F1-score의 개선은 리샘플링보다 임계값 최적화에서 더 뚜렷하게 나타났다.

따라서 실제 응용에서는 먼저 검증 데이터 기반의 임계값 정책을 최적화하는 것이 바람직하다. 이후 필요에 따라 비용 민감도 조정(class_weight)이나 모델 확률 보정(calibration) 기법을 추가적으로 적용하는 접근이 권장된다. 이러한 단계적 전략은 불균형 의료 데이터에서 보다 신뢰성 있는 사망 위험 예측 모델을 구축하는 데 기여할 수 있다.

제 5 장 결과 분석 - 시나리오 2: 코로나 양성 환자 사망 여부 예측

제 1 절 코호트 특성 및 분석 설정

본 시나리오는 CLASSIFFICATION_FINAL $\in \{1, 2, 3\}$으로 필터링된 코로나 양성 환자만을 대상으로 한다. 이때 CLASSIFFICATION_FINAL 변수는 집단이 이미 확진자로 한정되어 있어 상수와 유사하게 작동하므로 모델 입력에서 제거하였다. 따라서 예측은 인구학적 변수, 기저질환, 생활 습관, 진료 관련 변수들을 중심으로 수행되었다. 전체 환자 대비 표본 규모는 감소하였으며, 클래스 불균형은 여전히 존재하지만 그 정도는 일부 완화된 것으로 나타났다.

제 2 절 베이스라인 성능

확진자 코호트에서 리샘플링을 적용하지 않고 학습·평가한 결과, 두 모델 모두 전체 환자 시나리오와 유사한 경향을 보였다. 다만 코로나 여부 변수가 제거되면서, 다른 변수들의 기여도가 상대적으로 균형 있게 반영되는 특성이 확인되었다.

로지스틱 회귀의 테스트 성능은 ROC-AUC 0.9104, PR-AUC 0.5918, F1-score 0.5910으로 나타났으며, LightGBM은 ROC-AUC 0.9120, PR-AUC 0.5944, F1-score 0.5754로 측정되었다. 전반적으로 두 모델 모두 안정적인 ROC-AUC를 보였으며, PR-AUC는 전체 환자 분석 대비 소폭 변동을 보였다. 이는 코호트 규모 축소 및 사망자 비율 변화의 영향으로 해석된다.

표 4. 베이스라인 @0.50 성능 비교

모델명	ROC-AUC	PR-AUC	F1-score	Precision	Recall	ACC
Logistic	0.9104	0.5918	0.5910	0.4552	0.8424	0.8387
LightGBM	0.9120	0.5944	0.5754	0.4287	0.8748	0.8214

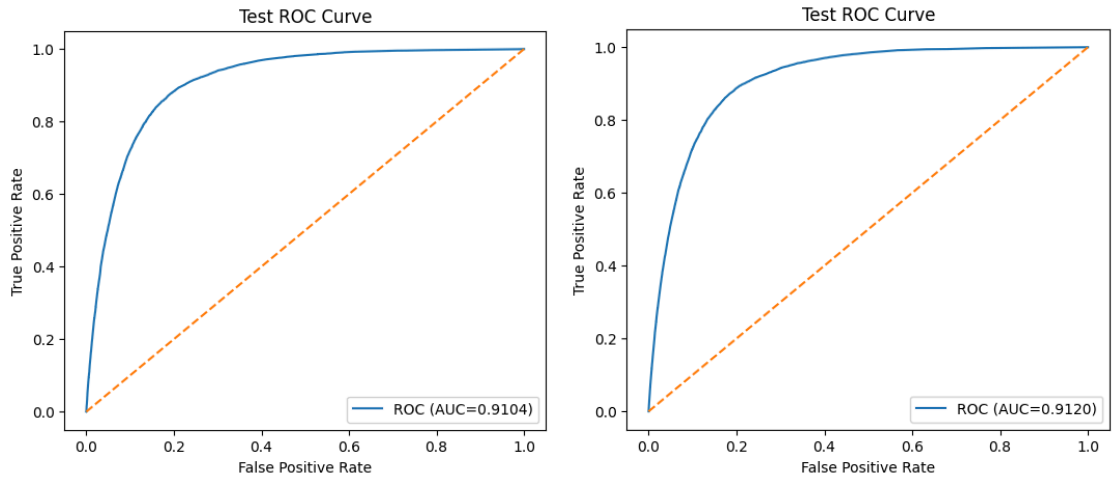

그림 4. ROC 곡선(좌: 로지스틱, 우: LightGBM)

제 3 절 임계값 조정 효과

검증셋에서 F1-score를 기준으로 최적 임계값을 탐색하여 적용한 결과, 시나리오 1과 동일하게 Precision과 F1-score는 향상되고 Recall은 감소하는 양상이 확인되었다. 이는 임계값 상향 조정을 통해 거짓양성(FP)이 줄어든 반면, 거짓음성(FN)이 늘어난 결과로 해석된다.

코로나 확진자 코호트의 경우, 임상적 맥락에서는 사망자를 최대한 놓치지 않는 것(Recall 우선)이 중요할 수 있다. 따라서 운영 목적에 따라서는 기본값인 0.50보다 더 낮은 임계값을 적용하여 Recall을 확보하는 전략이 필요하다. 반대로 불필요한 경고(FP)의 비용이 큰 환경이라면, 본 연구에서 도출된 최적 임계값(예: 로지스틱 0.68, LightGBM 0.70)을 활용하는 것이 더 합리적일 수 있다. 즉, 임계값 설정은 모델의 성능 최적화뿐 아니라 실제 적용 맥락에 따라 달라져야 함을 보여준다.

표 5. 임계값 0.50 vs 최적 임계값(*, F1 기준) 성능 비교

모델명	ROC-AUC	PR-AUC	F1-score	Precision	Recall	ACC
Logistic	0.9104	0.5918	0.5910	0.4552	0.8424	0.8387
Logistic*	0.9104	0.5918	0.6144	0.5249	0.7407	0.8387
LightGBM	0.9120	0.5944	0.5754	0.4287	0.8748	0.8214
LightGBM*	0.9120	0.5944	0.6141	0.5185	0.7529	0.8691

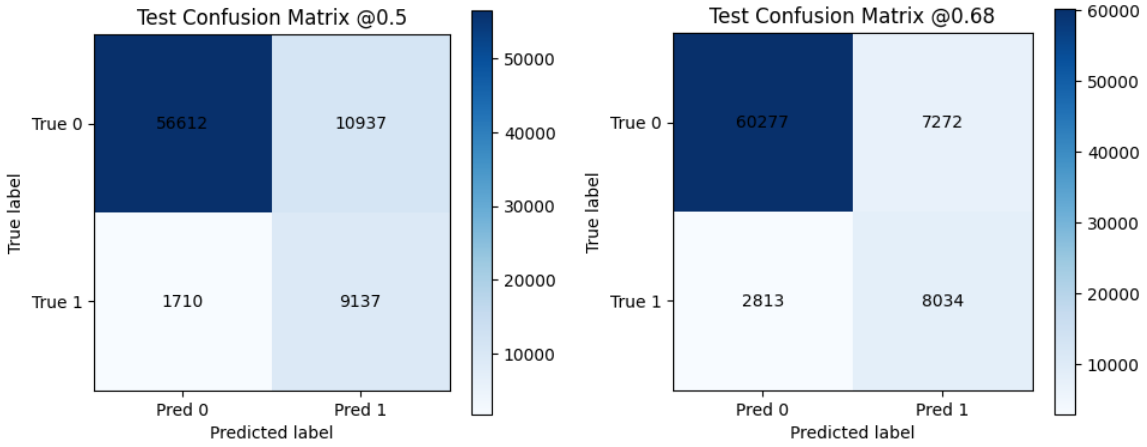

그림 5. 임계값 0.5 vs 최적 임계값의 혼동 행렬 (Logistic)

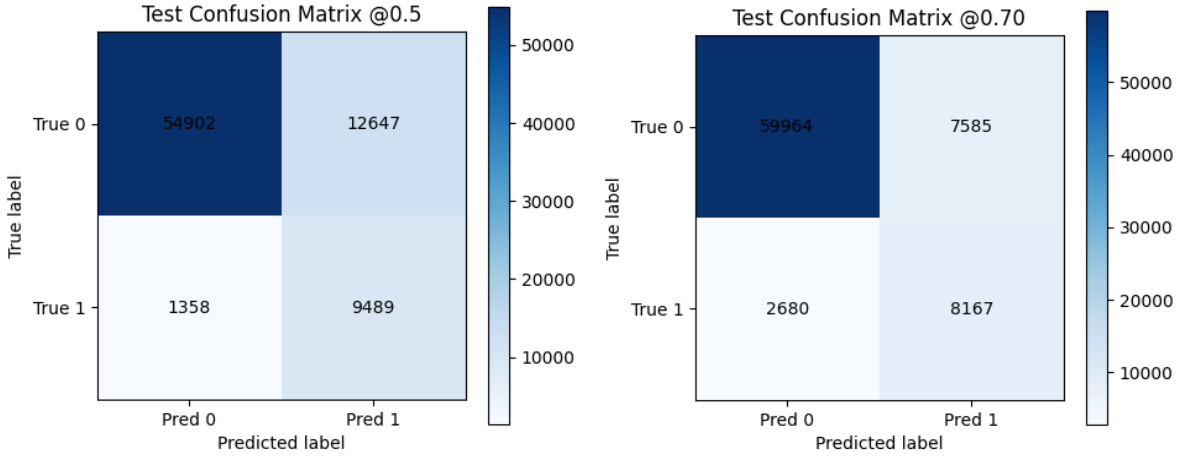

그림 6. 임계값 0.5 vs 최적 임계값의 혼동 행렬 (LightGBM)

제 4 절 리샘플링 적용 결과

훈련 데이터에 한정하여 언더샘플링(Random Under Sampling)과 오버샘플링(Random Over Sampling)을 각각 적용하고, 검증·테스트 단계에서는 원래의 불균형 분포를 유지한 채 평가를 수행하였다. 그 결과, 임계값 0.50 기준에서 리샘플링을 적용한 모델들의 F1-score와 PR-AUC는 베이스라인과 거의 유사하거나 근소한 개선 수준에 머물렀다. 이는 코로나 확진자 코호트에서도 리샘플링 기법 자체가 모델 성능을 획기적으로 향상시키지는 못하며, 오히려 임계값 조정의 효과가 더 크게 나타난다는 점을 재확인시켜준다.

따라서 불균형 데이터 대응 전략으로 단순 리샘플링만을 사용하는 것은 한계가 있으며, 실무적으로는 class_weight 조정이나 확률 보정(calibration)과 같은 방법이 더 효율적인 대안이 될 수 있음을 시사한다.

표 6. 리샘플링(RUS/ROS) @0.5 성능 비교 (Test, thr=0.50)

모델명	ROC-AUC	PR-AUC	F1	Precision	Recall	ACC
Logistic	0.9104	0.5918	0.5910	0.4552	0.8424	0.8387
Logistic-RUS	0.9103	0.5918	0.5922	0.4569	0.8415	0.8397
Logistic-ROS	0.9103	0.5919	0.5917	0.4562	0.8418	0.8393
LightGBM	0.9120	0.5944	0.5754	0.4287	0.8748	0.8214
LightGBM-RUS	0.9117	0.5931	0.5753	0.4288	0.8739	0.8215
LightGBM-ROS	0.9119	0.5959	0.5754	0.4289	0.8738	0.8216

제 5 절 해석 및 임상적 함의

코로나 확진자만을 대상으로 한 분석에서 로지스틱 회귀와 LightGBM은 모두 안정적인 분류 성능을 보였다. ROC-AUC와 PR-AUC 수치는 시나리오 1과 유사하거나 근소한 차이를 나타냈으며, 전반적으로 두 알고리즘 모두 사망 예측에 활용 가능한 수준임을 확인하였다.

임계값을 조정한 결과, Precision과 Recall 간의 상충 관계가 뚜렷하게 나타났다. 임계값을 높이면 거짓양성(FP)이 줄어 Precision과 F1-score가 향상되었으나, Recall은 감소하여 일부 사망자를 놓치는 경우가 증가하였다. 반대로 임계값을 낮추면 Recall은 개선되지만 불필요한 경고(거짓양성)가 늘어나는 양상이 확인되었다.

이러한 결과는 실제 의료 의사결정 맥락에서 중요한 의미를 가진다. 환자 안전을 우선할 경우 Recall을 높여 가능한 많은 고위험군을 포착하는 전략이 필요하며, 이는 1차 선별(screening) 상황에 적합하다. 반면, 불필요한 자원 소모를 줄이고 효율성을 강조할 경우 Precision을 중시하는 접근이 바람직하다. 따라서 실제 적용에서는 목적에 따라 임계값을 조정하거나, 단계적 선별 전략을 도입하는 것이 합리적이다.

제 6 절 소결

코로나 확진자 코호트에서의 실험 결과, LightGBM은 전반적 분류력(ROC-AUC, PR-AUC)에서 로지스틱 회귀보다 근소하게 우세하였다. 그러나 두 모델 모두 성능 차이가 크지 않았으며, 실제 활용 가능성은 충분하였다. 또한, 리샘플링보다는 임계값 조정이 성능 개선에 더 효과적인 방법으로 확인되었다. 이러한 결과는 "연령과 기저질환 중심의 예측 변수 → 모델링 기법 선택 → 임계값 정책 최적화"라는 단계적 접근이 사망 위험 예측에서 중요함을 시사한다.

제 6 장 결론 및 한계

제 1 절 연구 요약

본 연구는 Kaggle에서 제공하는 COVID-19 환자 데이터를 활용하여 사망 여부를 예측하는 이진 분류 모델을 구축하고, 기존 분석의 한계를 보완하는 것을 목표로 하였다. 기존 보고서는 전체 환자를 대상으로 분석하면서 CLASSIFFICATION_FINAL 변수를 포함하였는데, 이 변수는 코로나 양성 여부를 사실상 직접적으로 알려주어 모델 성능을 과대평가할 수 있는 문제가 있었다. 이에 본 연구는 전체 환자와 코로나 확진자만을 분리한 두 시나리오를 설계하고, 불균형 문제 해결을 위해 언더샘플링과 오버샘플링을 적용하였으며, 로지스틱 회귀와 LightGBM 알고리즘을 비교하였다.

제 2 절 주요 결과

분석 결과, LightGBM은 로지스틱 회귀보다 일관되게 높은 ROC-AUC와 PR-AUC를 기록하며 전반적 분류 성능에서 우위를 보였다. 임계값을 0.5에서 검증셋 기반 최적값으로 조정했을 때 Precision과 F1-score가 개선되었으나, Recall은 감소하는 trade-off가 확인되었다. 이는 모델이 오탐(False Positive)을 줄여 "사망 예측의 신뢰도"를 높이는 대신 일부 사망자를 놓치는 방향으로 작동했음을 의미한다. 반면 언더샘플링과 오버샘플링은 테스트셋의 불균형 분포를 그대로 둔 조건에서 성능 향상 효과가 제한적이었다. 코로나 확진자만을 대상으로 한 시나리오에서는 코로나

여부 변수의 영향력이 제거되면서, 연령과 기저질환 등 환자 특성이 보다 직접적으로 반영되는 경향이 나타났다.

제 3 절 연구의 한계

첫째, 본 연구에서 활용한 Kaggle 데이터는 행정 데이터 기반이어서 진단 및 보고 편향의 위험이 존재한다. 일부 변수에는 97, 98, 99와 같은 결측·불명 코드가 포함되어 있어 데이터 정확성이 떨어진다. 둘째, 사망자의 비율이 낮아 불균형 문제가 여전히 완전히 해소되지 않았다. 셋째, 혈액검사 결과나 산소포화도와 같은 세부 임상 지표가 포함되지 않아 의학적 해석력이 제한된다. 마지막으로, 비교한 모델이 로지스틱 회귀와 LightGBM에 국한되어 더 다양한 알고리즘과의 비교는 이루어지지 못했다.

제 4 절 향후 연구 방향

향후 연구에서는 SMOTE, ADASYN 등 고급 오버샘플링 기법과 비용 민감 학습(cost-sensitive learning)을 적용해 불균형 문제를 보다 효과적으로 해결할 수 있다. 또한 XGBoost나 심층 신경망을 포함한 다양한 알고리즘과의 비교가 필요하다. 더 나아가 실제 임상 현장에서는 사망자를 놓치지 않는 Recall 중심 전략과 불필요 경고를 줄이는 Precision 중심 전략 간 균형이 중요하므로, 목적에 따른 임계값 정책 시뮬레이션이 요구된다. 마지막으로, 다른 국가 및 기관의 데이터를 활용한 교차 검증을 통해 본 연구 결과의 일반화 가능성을 점검하는 것도 중요한 과제이다.