2024.08.24

회원 데이터 분석을 통한 만료/중지/탈퇴 예측 문제의 해결 방안에 대한 보고

빅데이터 9기 신유라

[목차]

[본문]

1. 요약

본 보고서는 회원의 상태를 "만료", "중지", "탈퇴"의 세 가지로 분류하고, 이를 예측하기 위한 머신러닝 모델의 성능을 평가하는 것을 목적으로 한다. 본 프로젝트에서는 변수 변환 방법에 따라 각각의 머신러닝 모델을 학습시켰으며, 그 결과 세 가지 모델의 성능을 비교 평가하였다.

2. 서론

1) 문제 정의

고객 이탈은 많은 기업에게 중요한 문제로, 이를 예측하고 효과적으로 대응하는 것이

사업 성과에 직결된다. 회원 상태를 "만료", "중지", "탈퇴"로 분류하여 고객 이탈을 조기에 감지하고, 이를 통해 적절한 마케팅 및 고객 유지 전략을 실행할 수 있다.

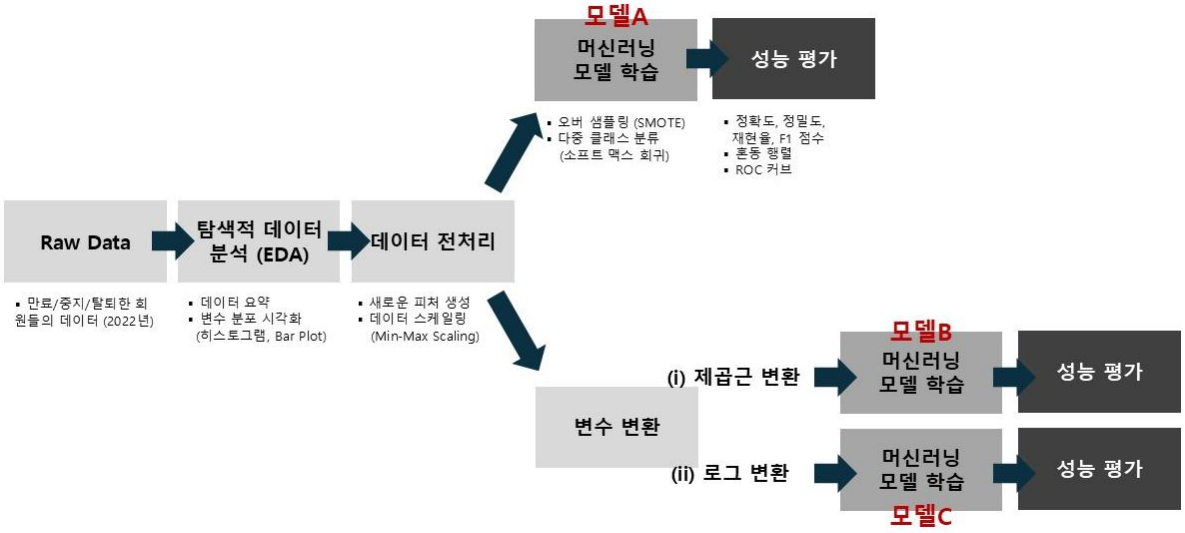
2) 프로젝트 기획의도

기업의 데이터는 점점 더 풍부해지고 있으며, 이러한 데이터를 활용하여 고객 행동을 예측하는 기술은 비즈니스 성공에 중요한 역할을 하고 있다. 본 프로젝트는 데이터 기반 의사 결정을 지원하기 위해, 고객의 이탈 여부를 예측할 수 있는 머신러닝 모델을 개발하는 데 중점을 두고 있다. 특히, 고객의 다양한 활동 데이터를 기반으로 다중 클래스 분류 문제를 해결하고자 한다.

3. 개발 환경

라이브러리	버전	라이선스
Python	3.12.4	Python Software Foundation License
Pandas	2.1.2	BSD 3-Clause License
Seaborn	0.12.2	BSD 3-Clause License
NumPy	1.26.0	BSD 3-Clause License
Matplotlib	3.8.0	Matplotlib License
scikit-learn	1.3.1	BSD 3-Clause License
imblearn	0.11.0	MIT License
SciPy	1.11.3	BSD 3-Clause License

4. 개발 프로세스



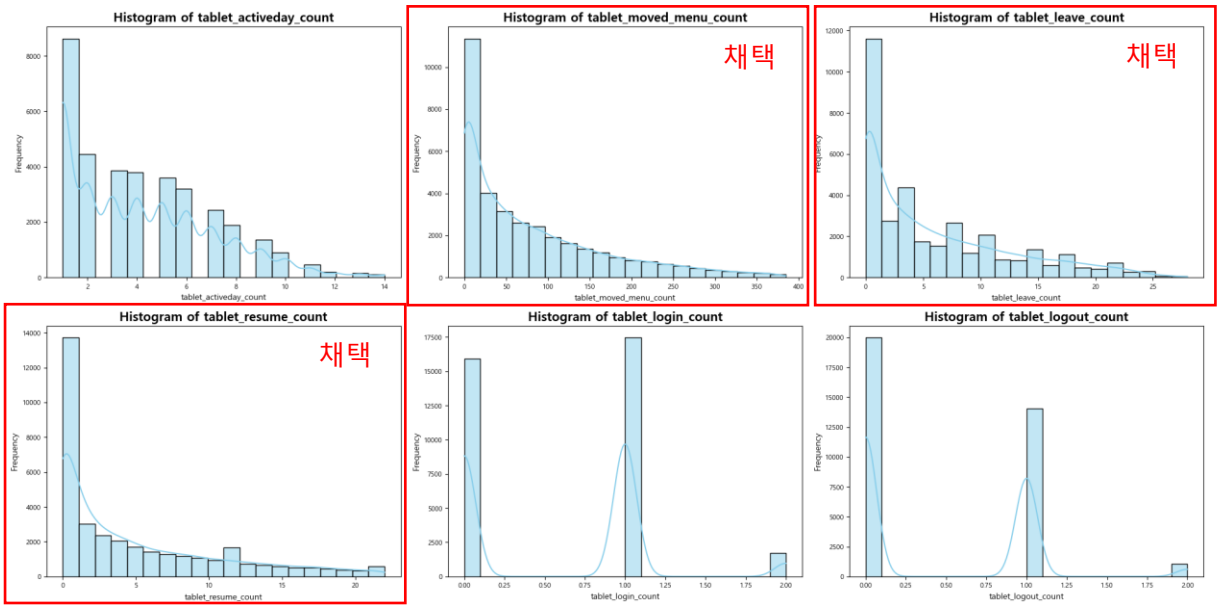
5. 사용된 데이터

- 프로젝트의 데이터는 별첨 자료('만료및탈퇴회원.csv')로 첨부하였다.
  - 데이터 셋 구성 : 천재교육 서비스 만료/중지/탈퇴입회원 데이터

6. Result

1) 탐색적 데이터 분석(EDA)

- 변수 분포 확인
  - 활동(tablet, study, media&video, test) 관련 변수들의 분포를 히스토그램으로 시각화하였다.
  - 변수 분포 히스토그램(대표 이미지)

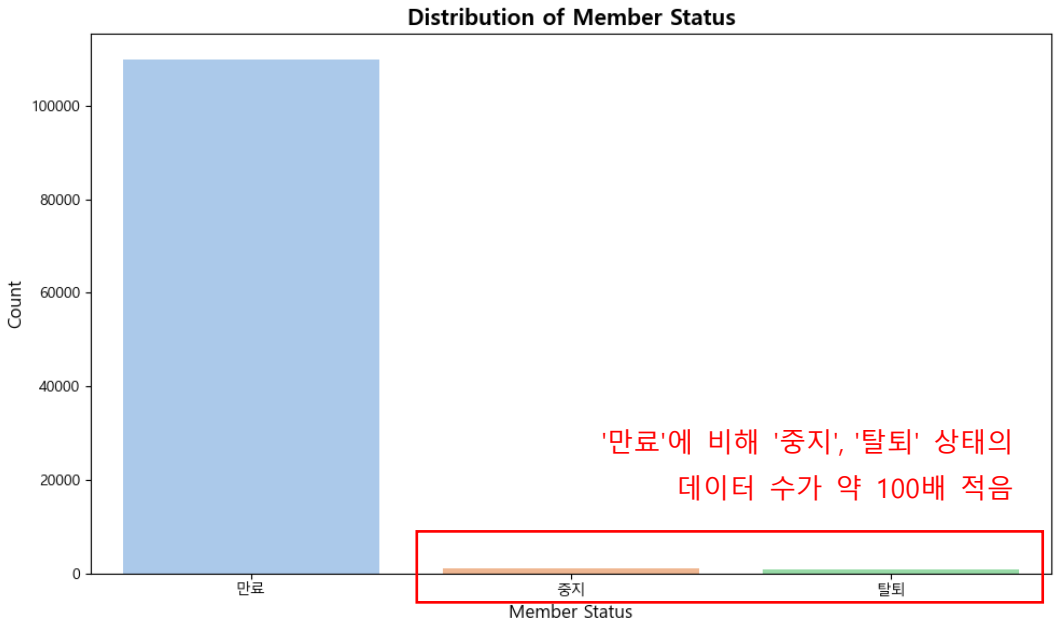


▫ 데이터 분포에 비어있는 곳이 없으며 일관된 분포를 보이는 변수를 독립변수로 채택하였다.

- 회원 상태별 빈도수 확인

▫ 회원 상태 (만료, 미납, 탈퇴)별 빈도수를 Bar Plot 으로 시각화하였다.

▫ Bar Plot



2) 데이터 전처리

- 오버샘플링 : 데이터의 불균형을 해결하기 위해 다수 클래스의 수를 기준으로 소수 클래스 데이터를 오버샘플링(SMOTE; Synthetic Minority Over-sampling Technique)하였

다.

회원상태	데이터 수 (개)	
	오버샘플링 전	**오버샘플링 후**
만료	76920	76920
중지	742	76920
탈퇴	633	76920

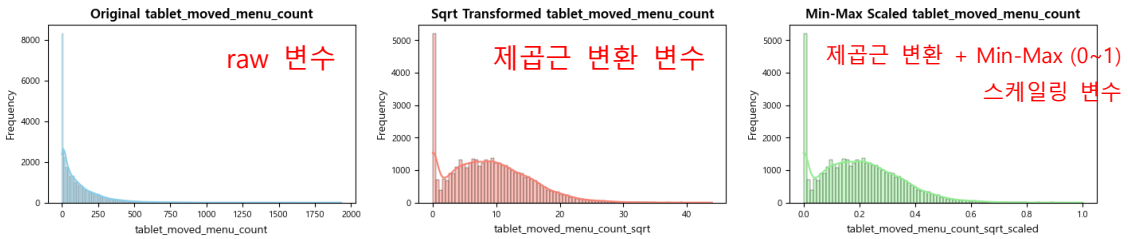
▪ 새로운 Feature 생성 : residual_point(잔여 포인트; 획득 포인트 - 차감 포인트) Feature를 새로 생성하였다.

▪ 데이터 스케일링 : 변수 간의 차이를 줄이고 모델의 성능을 최적화하기 위해 Min-Max 스케일링을 적용하였다.

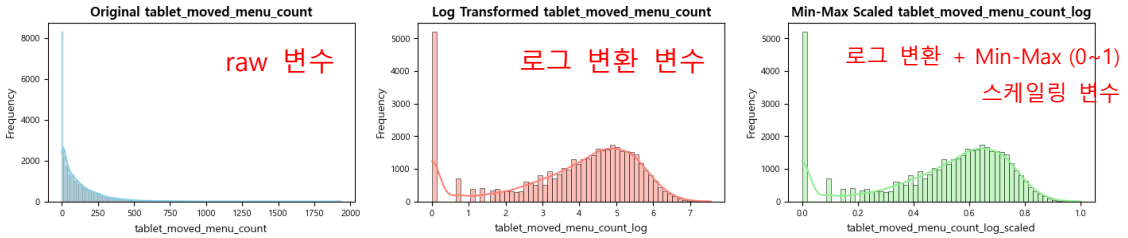
3) 변수 변환

▪ 오른쪽으로 긴 꼬리를 가진 데이터 분포를 완화하기 위해 변수를 제곱근 변환 및 로그 변환을 적용하였다.

▫ 제곱근 변환 결과 히스토그램(대표이미지)



▫ 로그 변환 결과 히스토그램(대표이미지)



4) 머신러닝 모델 학습

▪ 목표 : "만료", "중지", "탈퇴"의 세 가지 회원 상태 중 어느 하나에 속할 확률을 예측하고자 하였다.

▪ 학습 및 예측 모델 : 소프트 맥스 회귀 모델 (지도학습_다중 클래스 분류에 적합)

▪ 두 가지 변수 변환 기법(제곱근/로그 변환)으로 생성된 각각의 경우에 대해 소프트

맥스 회귀 모델을 활용한 머신러닝을 수행하였다.

5) 모델 성능 평가

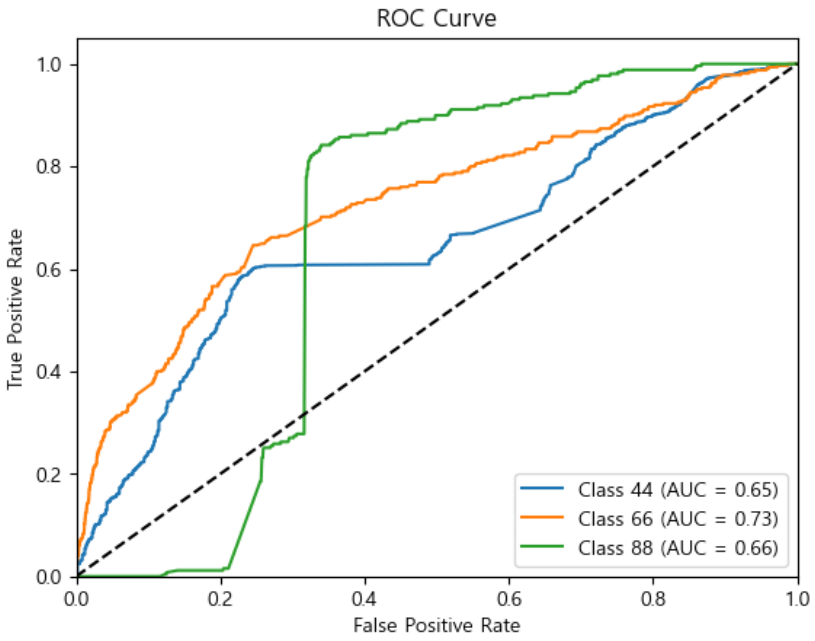
- **모델 A (변수 변환X)**
  - Accuracy : 0.24
  - Precision, Recall, F1-score

회원 상태	Precision	Recall	F1-score	Support
만료	0.99	0.23	0.37	32972
중지	0.03	0.46	0.06	325
탈퇴	0.01	0.94	0.02	259

  - Confusion Matrix

		예측		
		만료	**중지**	**탈퇴**
실제	**만료**	7513	4786	20673
	중지	43	151	131
	탈퇴	16	0	243

  - ROC 커브



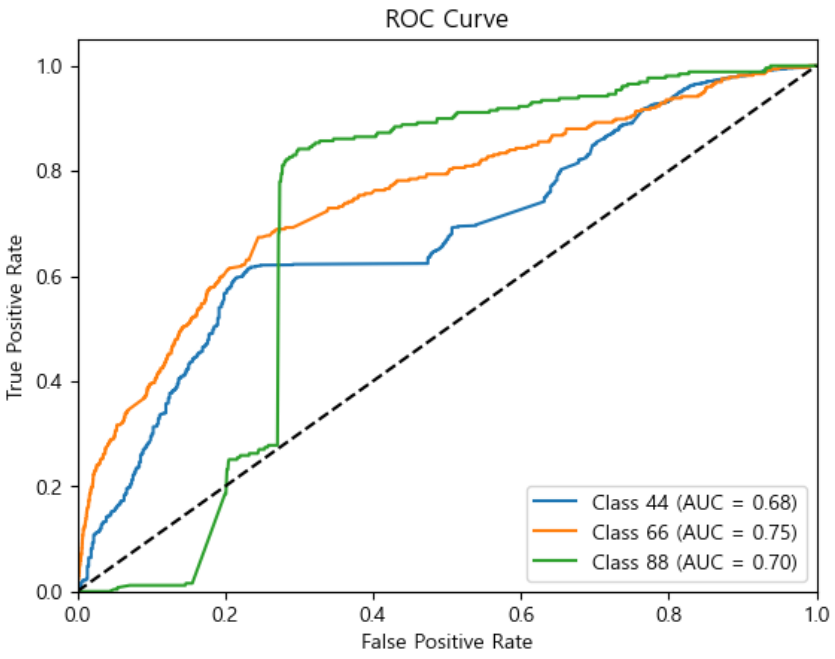
- **모델 B (제곱근 변수 변환)**
  - Accuracy : 0.45
  - Precision, Recall, F1-score

회원 상태	Precision	Recall	F1-score	Support
만료	0.99	0.45	0.62	32972
중지	0.03	0.52	0.06	325
탈퇴	0.02	0.88	0.03	259

  - Confusion Matrix

		예측		
		만료	**중지**	**탈퇴**
실제	**만료**	14714	5056	13202
	중지	68	168	89
	탈퇴	31	0	228

  - ROC 커브



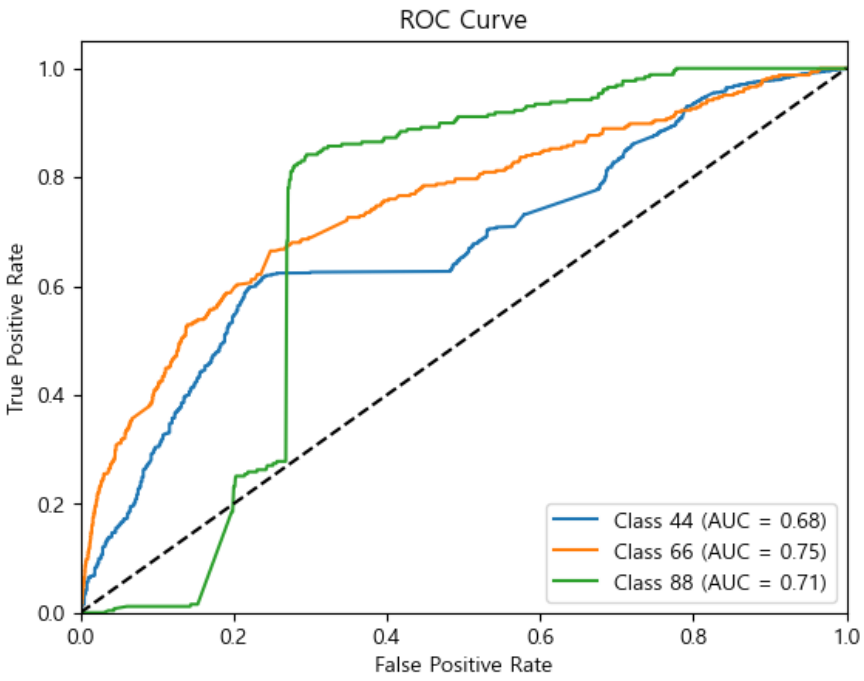
- **모델 C (로그 변수 변환)**
  - Accuracy : 0.49
  - Precision, Recall, F1-score

회원 상태	Precision	Recall	F1-score	Support
만료	0.99	0.48	0.65	32972
중지	0.03	0.54	0.06	325
탈퇴	0.02	0.86	0.04	259

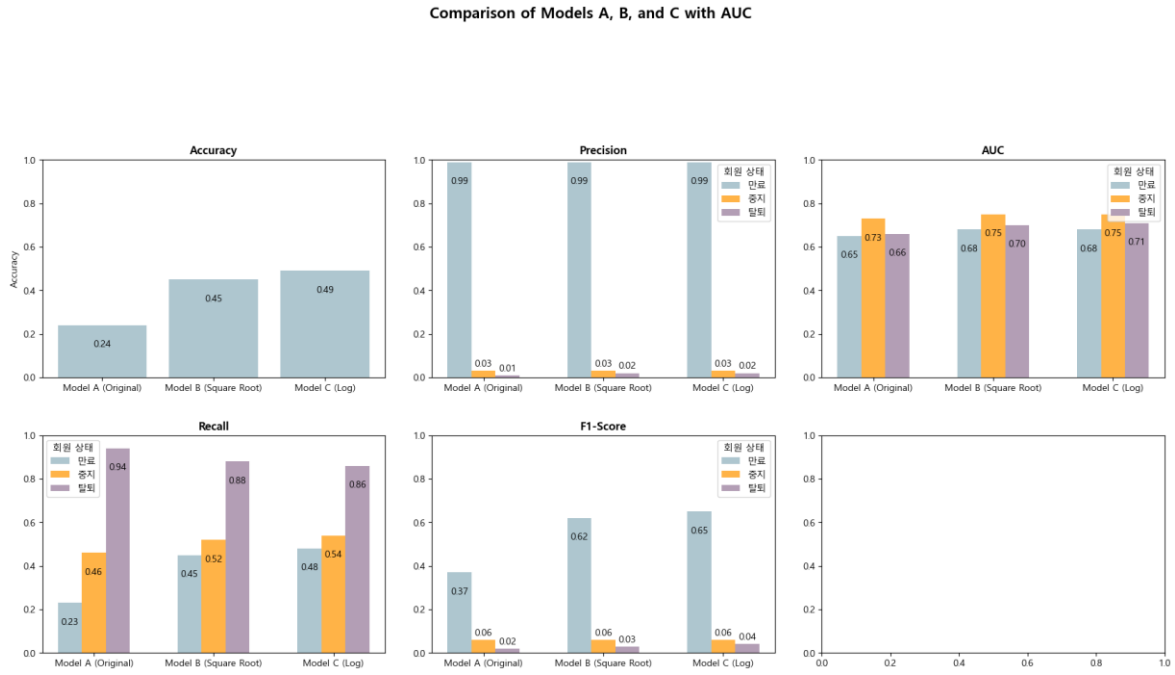
  - Confusion Matrix

		예측		
		만료	**중지**	**탈퇴**
실제	**만료**	15890	5014	12068
	중지	71	177	77
	탈퇴	35	0	224

  - ROC 커브



- A, B, C 모델 성능 평가 비교 시각화



7. Discussion

본 보고서에서는 만료, 중지, 탈퇴 상태로 분류된 회원 데이터를 기반으로, 다양한 변수 변환 방법을 통해 각 상태를 예측하는 머신러닝 모델의 성능을 비교 분석하였다. 모델 A(변수 변환X), 모델 B(제곱근 변환), 모델 C(로그 변환)의 성능을 각각 평가한 결과, 데이터의 변수 변환 방법이 모델의 예측 성능에 미치는 영향을 확인할 수 있었다.

3가지 모델의 Accuracy, Precision, Recall, F1-Score, 그리고 AUC 값을 통해 각 모델의 성능을 분석하였다. 분석 결과, 변수 변환이 일부 지표에서는 성능 향상에 긍정적인 영향을 미친 반면, 다른 지표에서는 큰 차이가 없거나 성능이 저하되는 모습을 보였다.

- **정확도(Accuracy)** : 모델 A의 Accuracy는 0.24로, 모델 B(0.45)와 모델 C(0.49)에 비해 상대적으로 낮았다. 이는 변수의 제곱근•로그 변환이 전반적인 모델 Accuracy 향상에 기여했음을 보여준다.

- **정밀도(Precision)** : 모든 모델에서 "만료" 상태의 Precision은 0.99로 매우 높게 유지되었지만, "중지"와 "탈퇴" 상태의 Precision은 매우 낮게(0.01~0.03) 나타났다. 따라서 변수의 제곱근•로그 변환이 Precision에는 큰 영향을 미치지 않았음을 알 수 있다.

- **재현율(Recall)** : 모델 A에서 "만료" 상태의 Recall은 0.23, 모델 B는 0.45, 모델 C는 0.48로, 변수 변환 후 성능이 약 2배 향상되었다. 또한 "중지" 상태의 Recall도 모델 A(0.46)에서 모델 B(0.52), 모델 C(0.54)로 증가했다. 이는 변수의 제곱근•로그

변환이 Recall에서 긍정적인 영향을 미친 것을 보여준다. 반면, "탈퇴" 상태의 Recall은 세 모델 모두에서 매우 높게 유지되었으며, 모델 A에서 0.94, 모델 B에서 0.88, 모델 C에서 0.86으로, 오히려 약간(0.06~0.08)의 성능 저하가 관찰되었다.

- **F1-Score**: 모델 A에서 "만료" 상태의 F1-Score는 0.37이었지만, 모델 B에서는 0.62, 모델 C에서는 약 1.7배 향상되었다. 그러나 "중지"와 "탈퇴" 상태에서는 F1-Score가 모든 모델에서 낮게 나타났으며, 큰 차이는 없었다.
- **AUC**: 모델 A의 AUC 값은 "만료" 상태에서 0.65, "중지" 상태에서 0.73, "탈퇴" 상태에서 0.66이었으며, 모델 B와 C에서는 전반적으로 AUC 값이 약간(0.02~0.08) 향상되었다. 특히 모델 C는 "중지" 상태에서 0.75로 가장 높은 AUC 값을 기록했다. 이는 로그 변환이 일부 클래스에서 예측 성능을 개선하는 데 기여했음을 시사한다.

이러한 결과는 변수 변환이 특정 상황에서 유용할 수 있음을 시사하지만, 모든 경우에 유리한 것은 아니며, 데이터의 특성과 문제의 성격에 따라 신중하게 적용해야 함을 보여준다. 특히, 만료 상태와 같은 클래스가 압도적으로 많은 불균형 데이터셋에서는 다수 클래스에 대한 높은 Precision을 유지하는 반면, 소수 클래스의 예측 성능은 향상되지 않을 수 있음을 확인하였다.

8. Conclusion

본 프로젝트에서는 만료, 중지, 탈퇴 상태로 분류된 회원 데이터를 대상으로 다양한 변수 변환 방법을 적용하여 머신러닝 모델의 성능을 비교 분석하였다. 분석 결과, 데이터의 변수 변환이 모델의 예측 성능에 다양한 영향을 미칠 수 있음을 확인할 수 있었다.

모델 A(변수 변환X)의 경우, 전반적인 성능이 모델 B(제곱근 변환)와 모델 C(로그 변환)에 비해 낮게 나타났다. 특히, 모델 C의 로그 변환은 "만료"와 "중지" 상태에서의 재현율(Recall)과 AUC를 개선하는 데 긍정적인 영향을 미쳤다. 이는 로그 변환이 데이터의 분포를 정규화하여 모델이 더 나은 예측 성능을 발휘하도록 도운 것으로 볼 수 있다.

다만, 모든 지표에서 변수 변환이 일관된 성능 향상을 보이지는 않았다. 예를 들어, Precision과 F1-Score의 경우, "중지"와 "탈퇴" 상태에서 변수 변환이 큰 영향을 미치지 않았으며, 일부 경우 오히려 성능 저하가 관찰되었다.

결론적으로, 변수 변환은 모델의 예측 성능을 향상시키는 데 중요한 역할을 할 수 있으며, 특히 불균형 데이터 문제를 해결하는 데 유용할 수 있다. 그러나 변수 변환이 모든 상황에서 효과적이지 않을 수 있으므로, 다양한 변환 방법을 시도하고 그 결과를 면밀히

분석하여 가장 적합한 방법을 선택하는 것이 중요하다.

향후 프로젝트에서는 회원이 언제 만료, 중지, 탈퇴 상태로 바뀔지 그 시기를 예측하기 위해 시계열 데이터를 추가로 분석할 필요가 있다. 이를 통해 단순히 상태 변화만 예측하는 것이 아니라, 그 시점까지도 정확히 예측할 수 있는 모델을 개발하는 데 집중할 수 있을 것이다.

9. 소스코드

- 프로젝트의 소스코드는 별첨 자료('만료및탈퇴회원_소스코드_240822_신유라_수정.ipynb')로 제공되며, 다음의 주요 섹션으로 구성되어 있다:
  - 데이터 전처리
  - 모델 학습
  - 성능 평가 및 시각화

10. 약어

- **SMOTE** : Synthetic Minority Over-sampling Technique (합성 소수 클래스 오버샘플링 기법)
- **EDA** : Exploratory Data Analysis (탐색적 데이터 분석)