통신사 고객 이탈 예측 ML 모델

Telecom Customer

3차 Project 6팀 1조

김세빈, 김현서, 손준영, 장난영

목차

문제 정의 및 목표

1 데이터 분석을 통해 이탈 고객의 특성을 파악하고 모델링에 적합한 Feature Engineering을 합니다.

2 고객 이탈 예측 모델링과 평가를 통해 이상적인 모델을 찾습니다.

3 성능을 향상시켜 최적의 예측 모델을 만듭니다.

>>>

이탈 고객의 특성을 파악하고
최적의 이탈 예측
ML 모델을 만드는 것

데이터 셋 설명

데이터:

https://www.kaggle.com/datasets/shilongzhuang/telecom-customer-churn-by-maven-analytics

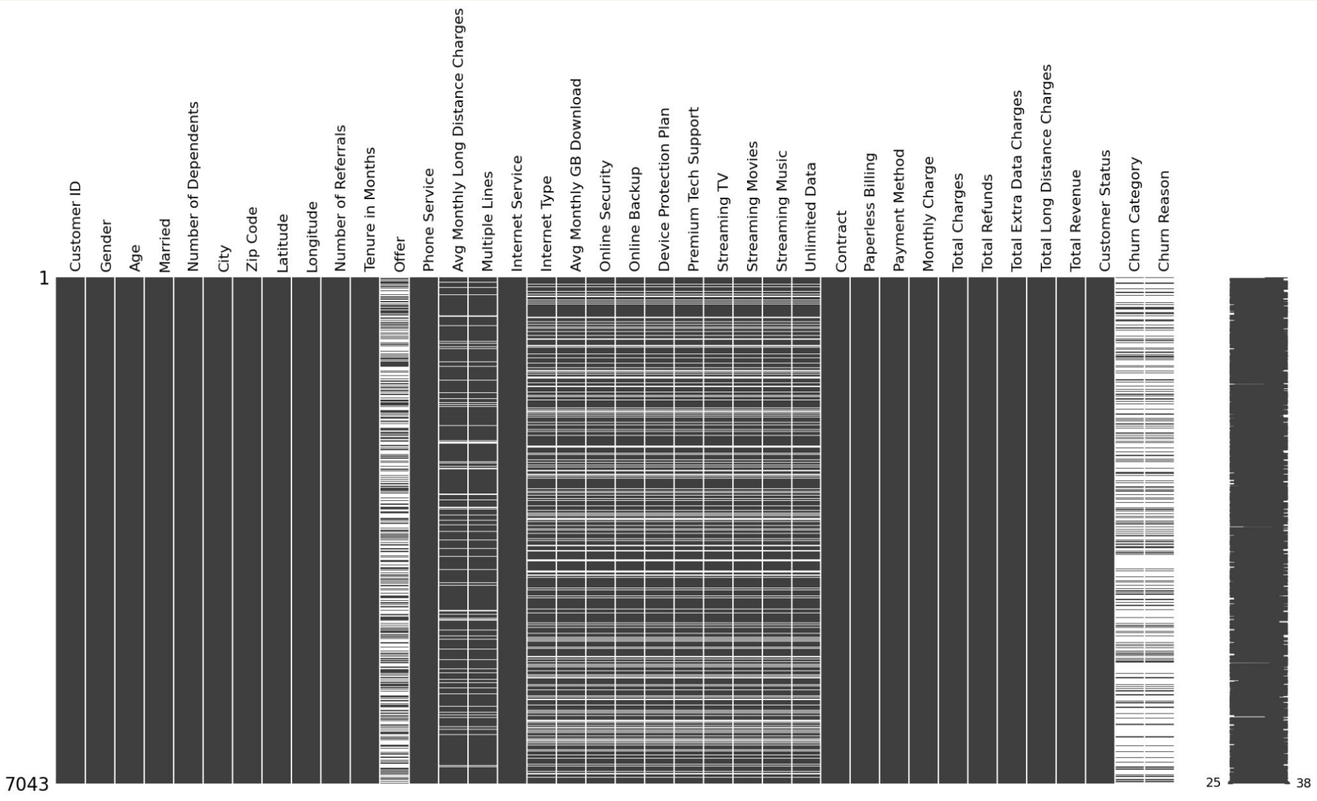
데이터 설명 (7043, 38) :

	Customer ID	Gender	Age	Married	Number of Dependents	City	Zip Code	Latitude	Longitude	Number of Referrals	...	Payment Method	Monthly Charge	Total Charges	Total Refunds	Total Extra Data Charges	Total Long Distance Charges	Total Revenue	Customer Status	Churn Category	Churn Reason
0	0002-ORFBO	Female	37	Yes	0	Frazier Park	93225	34.827662	-118.999073	2	...	Credit Card	65.60	593.30	0.00	0	381.51	974.81	Stayed	NaN	NaN
1	0003-MKNFE	Male	46	No	0	Glendale	91206	34.162515	-118.203869	0	...	Credit Card	-4.00	542.40	38.33	10	96.21	610.28	Stayed	NaN	NaN
2	0004-TLHLJ	Male	50	No	0	Costa Mesa	92627	33.645672	-117.922613	0	...	Bank Withdrawal	73.90	280.85	0.00	0	134.60	415.45	Churned	Competitor	Competitor had better devices
3	0011-IGKFF	Male	78	Yes	0	Martinez	94553	38.014457	-122.115432	1	...	Bank Withdrawal	98.00	1237.85	0.00	0	361.66	1599.51	Churned	Dissatisfaction	Product dissatisfaction
4	0013-EXCHZ	Female	75	Yes	0	Camarillo	93010	34.227846	-119.079903	3	...	Credit Card	83.90	267.40	0.00	0	22.14	289.54	Churned	Dissatisfaction	Network reliability
5	0013-MHZWF	Female	23	No	3	Midpines	95345	37.581496	-119.972762	0	...	Credit Card	69.40	571.45	0.00	0	150.93	722.38	Stayed	NaN	NaN
6	0013-SMEOE	Female	67	Yes	0	Lompoc	93437	34.757477	-120.550507	1	...	Bank Withdrawal	109.70	7904.25	0.00	0	707.16	8611.41	Stayed	NaN	NaN
7	0014-BMAQU	Male	52	Yes	0	Napa	94558	38.489789	-122.270110	8	...	Credit Card	84.65	5377.80	0.00	20	816.48	6214.28	Stayed	NaN	NaN
8	0015-UOCOJ	Female	68	No	0	Simi Valley	93063	34.296813	-118.685703	0	...	Bank Withdrawal	48.20	340.35	0.00	0	73.71	414.06	Stayed	NaN	NaN
9	0016-QLJIS	Female	43	Yes	1	Sheridan	95681	38.984756	-121.345074	3	...	Credit Card	90.45	5957.90	0.00	0	1849.90	7807.80	Stayed	NaN	NaN

- 2022년 제2분기에 캘리포니아의 한 통신 회사의 7,043명의 고객에 관한 정보
- Row: 한 명의 고객 / Column: 고객의 정보, 재직 기간, 구독 서비스, 해당 분기의 상태(가입, 유지, 이탈) 등에 대한 세부 정보
- 종속변수 (목표 변수): Customer Status
- 데이터 타입: int64, float64, object

EDA 및 분석 – 결측치 처리

✔ Null Columns

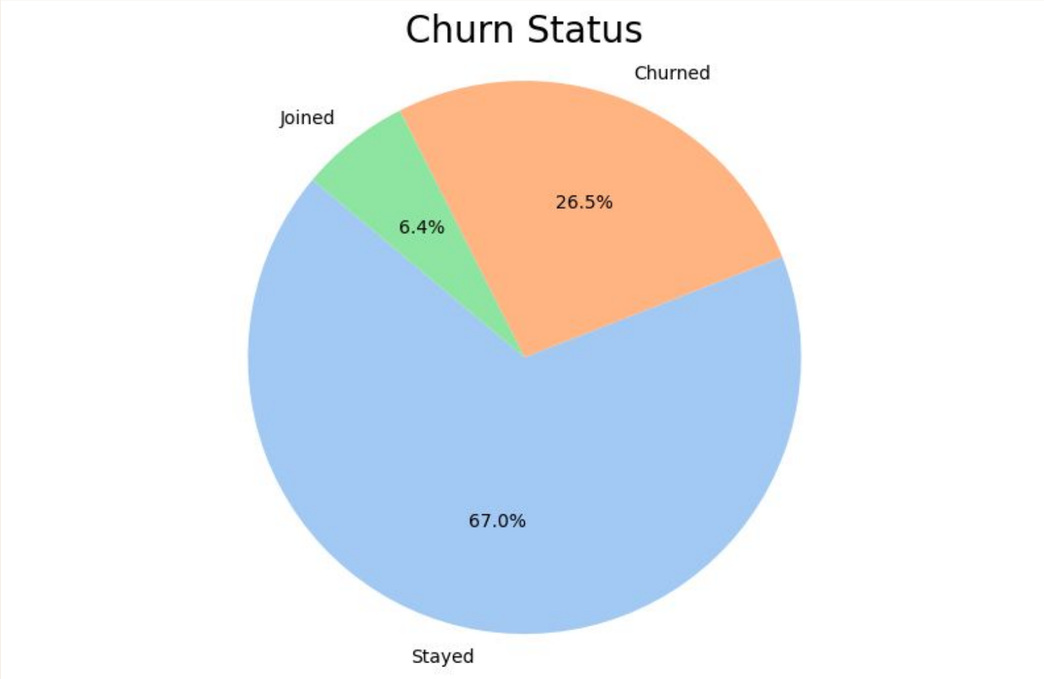


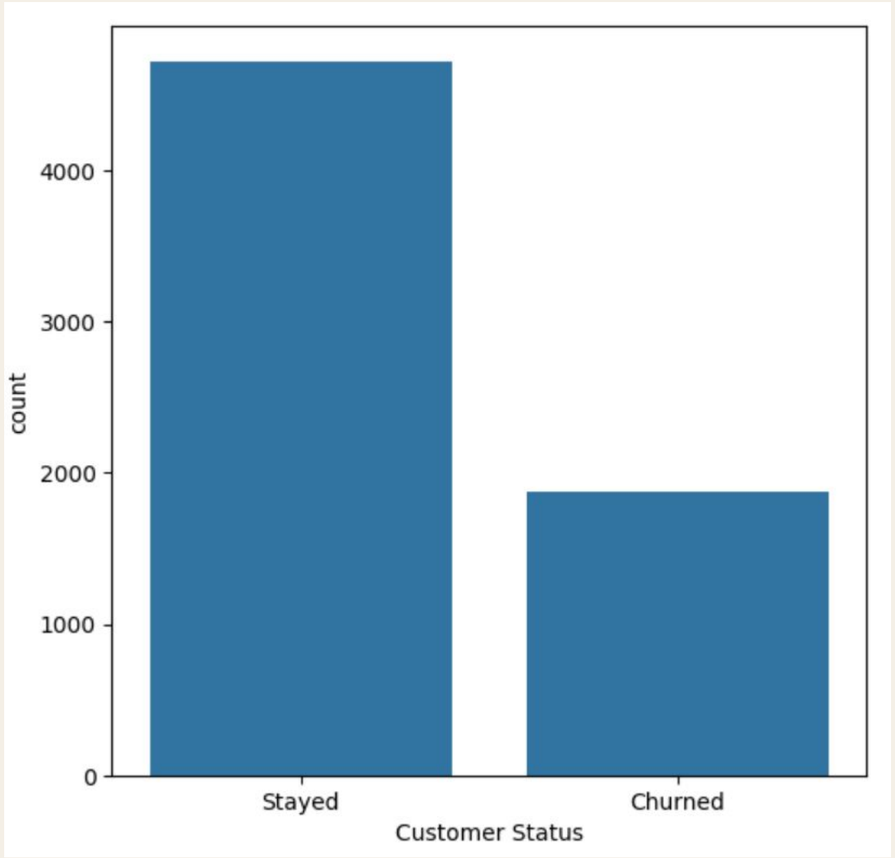
- Offer : None
- Avg Monthly Long Distance Charges : 0
- Avg Monthly GB Download : 0
- Multiple Lines: No Phone Service
- Internet Service – 부가 서비스 :

  No Internet Service
- Churn Category / Churn Reason : None

EDA 및 분석 – 데이터 분석

✔ Bound Variable (Churn Status)

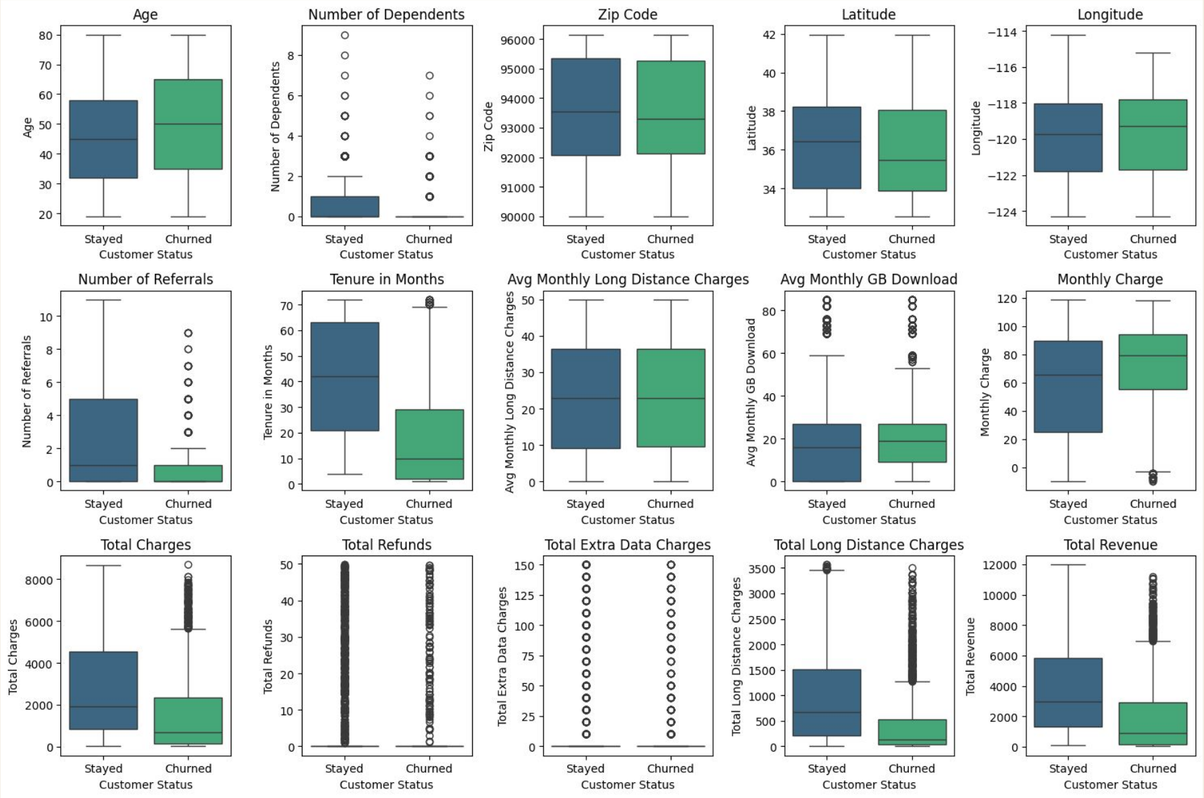




- 'Joined' 상태는 신규 가입자를 나타냄 -> 이탈 여부 불확실
- 'Stayed'와 'Churned' 상태만을 고려하여 데이터 분석 및 모델링

EDA 및 분석 – 데이터 분석

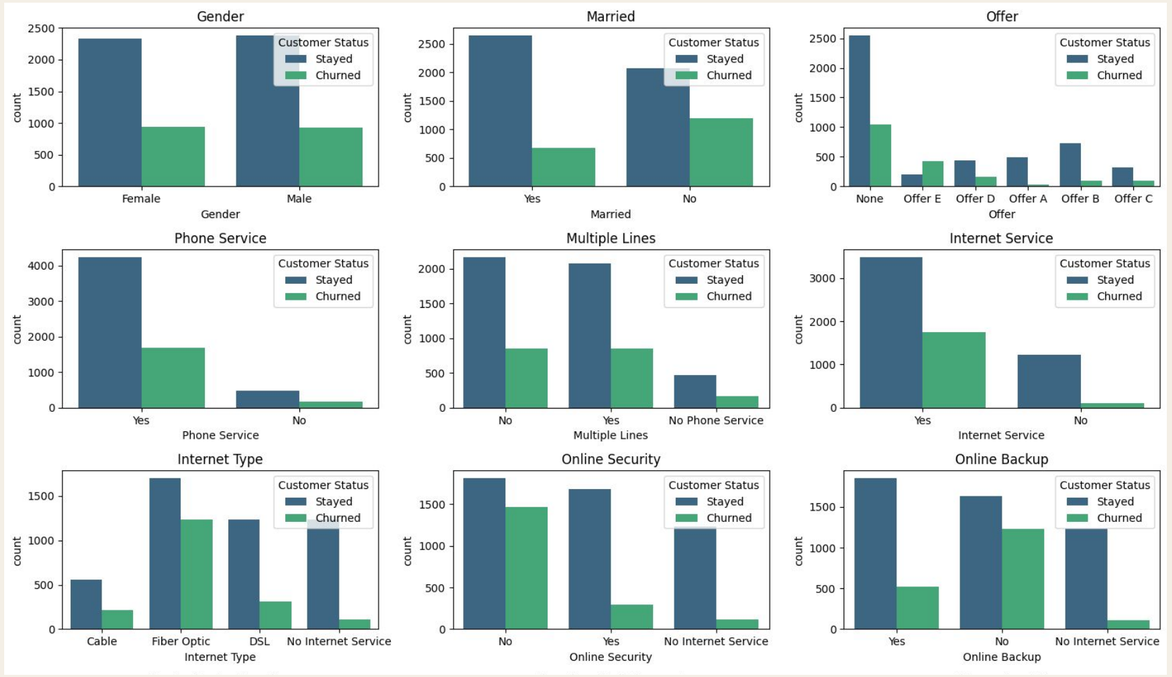
✔ Numeric Columns

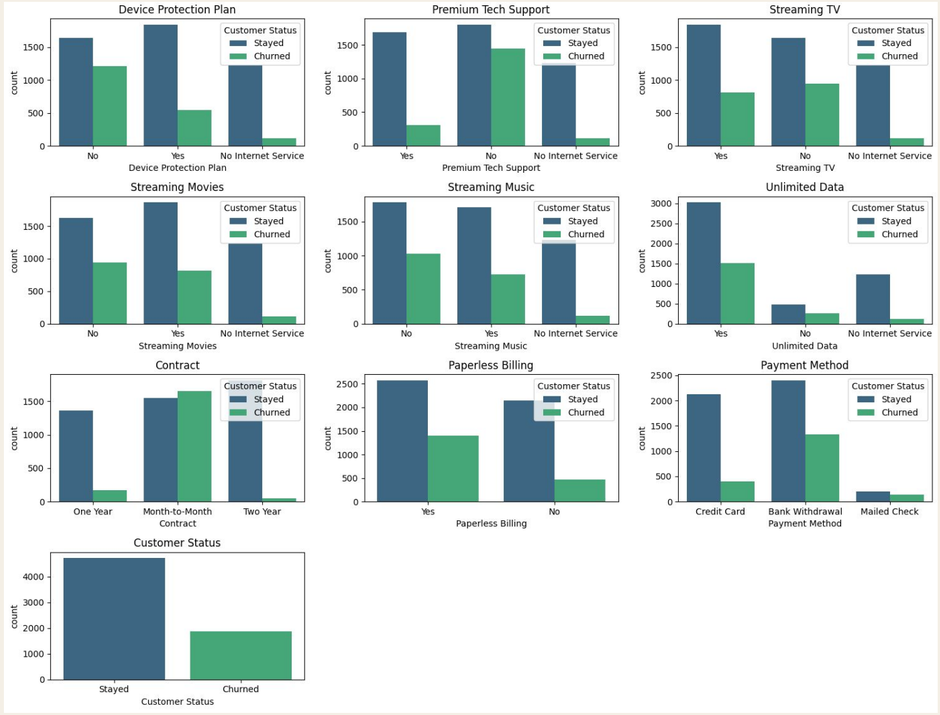


- Tenure in Months: 가입 후 첫 4개월 동안 고객이 이탈할 가능성이 높아짐
- Total Charge/ Total Revenue: 이탈 고객의 총 요금과 총 수익이 체류 고객보다 현저히 낮음

EDA 및 분석 – 데이터 분석

✔ Object Columns

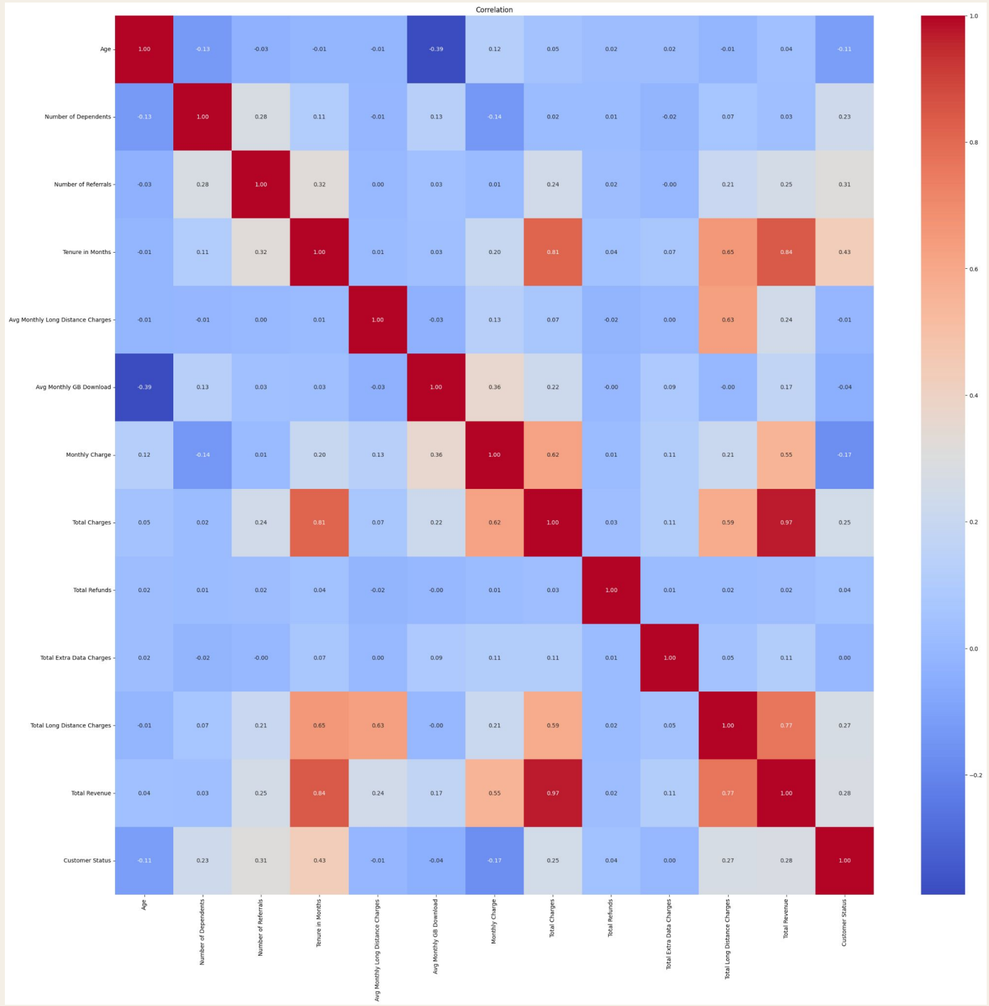




- Offer: 대부분의 고객은 오퍼 유형을 받지 못함. Offer E 이탈률이 좋지 않음
- Service: 전화 서비스나 인터넷 서비스를 이용하는 고객이 이탈할 가능성이 높은 것으로 보이며, 이는 열악한 서비스를 반영
- Contract/Payment Method: 월 구독 및 은행 출금이 있는 고객이 이탈률이 높음

Feature Engineering

✔ Drop Columns



- 불필요한 변수: 'Customer ID', 'City', 'Zip Code', 'Latitude', 'Longitude', 'Churn Category', 'Churn Reason'

- 'Avg Monthly Long Distance Charges', 'Avg Monthly GB Download', 'Total Refunds', 'Total Extra Data Charges',는 'Customer Status'와 거의 또는 전혀 상관 관계가 없는 것으로 나타남. 따라서 이러한 변수들은 종속 변수인 'Customer Status'를 예측하는 데 도움이 되지 않을 것으로 판단되며, 불필요한 변수로 간주하여 제거

Feature Engineering

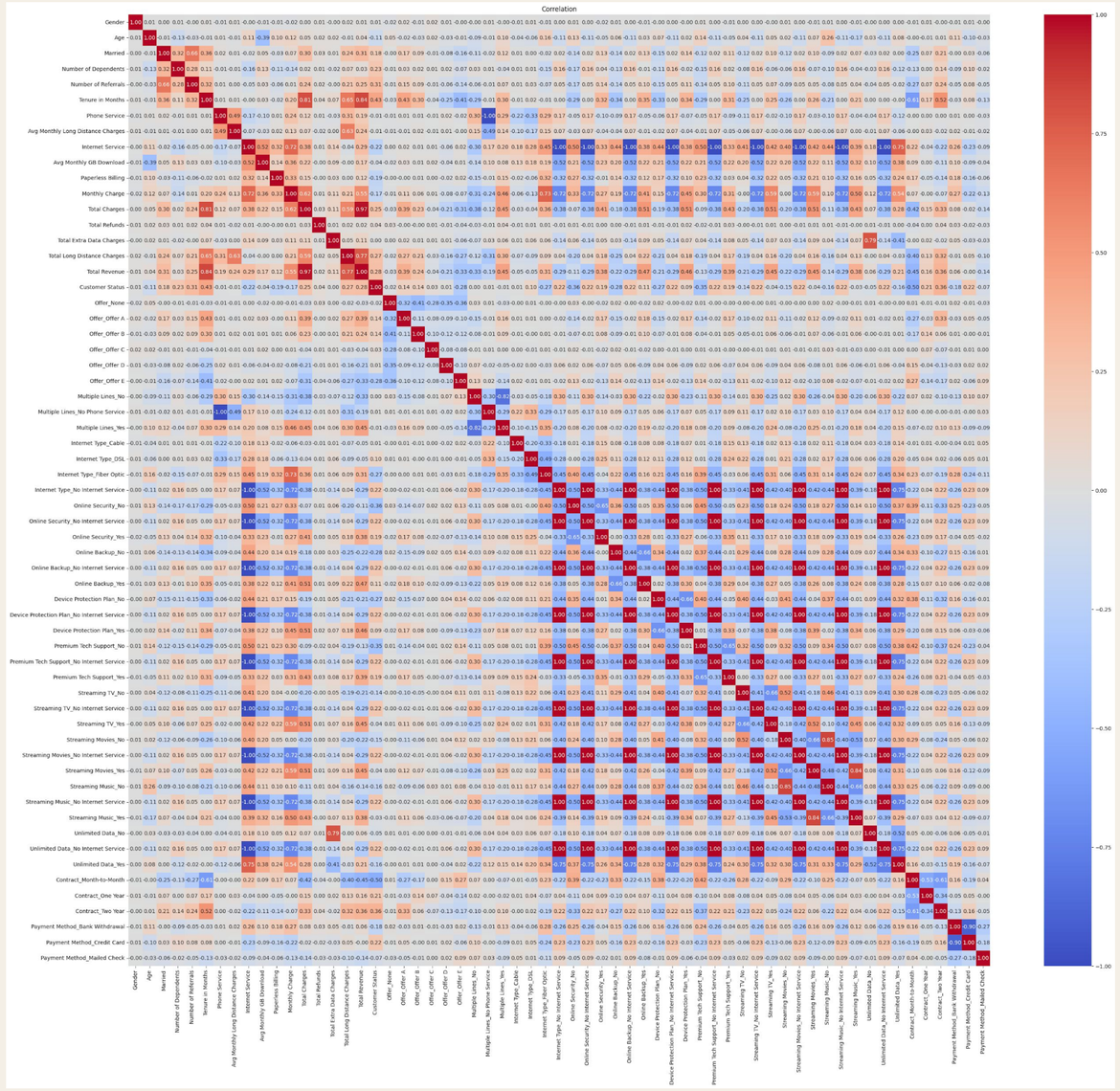
✔ Label Encoding

- 카테고리가 2개인 경우:
- ‘Customer Status’(종속변수)
- ‘Gender’, ‘Married’, ‘Phone Service’, ‘Internet Service’, ‘Paperless Billing’,

✔ OneHot Encoding

- 카테고리가 3개 이상인 경우:
- ‘Offer’, ‘Multiple Lines’, ‘Internet Type’, ‘Online Security’, ‘Online Backup’, ‘Device Protection Plan’, ‘Premium Tech Support’, ‘Streaming TV’, ‘Streaming Movies’, ‘Streaming Music’, ‘Unlimited Data’, ‘Contract’, ‘Payment Method’

Feature Engineering



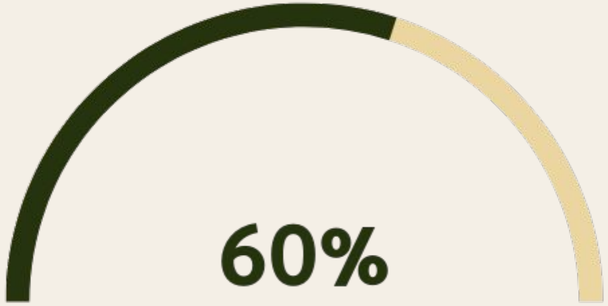
✔ 상관관계가 0.1 이하인 컬럼들 제외

- Offer_None, Offer_Offer C, Offer_Offer D
- Multiple Lines_No, Multiple_Lines_Yes
- Multiple Lines_No Phone Service
- Internet Type_Cable
- Device Protection Plan_Yes
- Streaming TV_Yes
- Streaming Movies_Yes
- Streaming Music_Yes
- Unlimited Data_No
- Payment Method_Mailed Check
- Gender
- Phone Service

모델링 방향 – 사전 작업요약

결측치 처리

데이터 분석



Feature Engineering

결측치를 처리하고 이상치를 확인하여 모델링에 적합한 데이터 셋 구축

Feature Engineering을 통해 불 필요한 칼럼 삭제 및 인코딩

모델링 방향

분류 알고리즘
Scaler 별 성능 평가

선택된 ML 모델의
최적화 시도

결론 및 인사이트

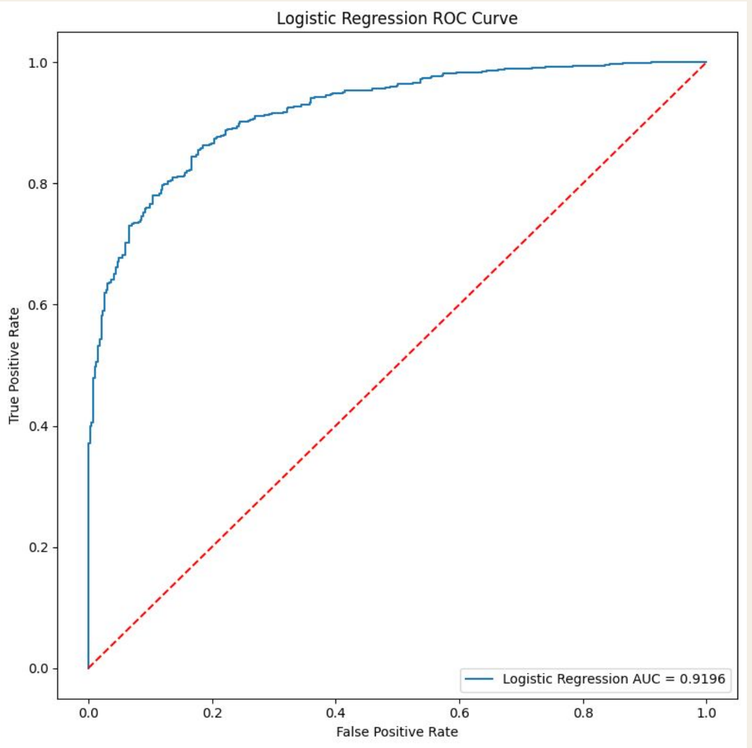
Decision Tree, Logistic Regression, SVM, Naïve Bayes, KNN,

Random Forest, Xgboost, Catboost, LightGBM, Gradient Boost

모델링 및 평가 분석 – Scaler

✔ Logistic Regression

	MinMax	Standard	Robust
Train Accuracy	0.8537	0.8550	**0.8535**
Test Accuracy	0.8535	0.8573	**0.8573**
F1 (Churned / Stayed)	0.74	0.75	**0.75**
	0.90	0.90	**0.90**
AUC (ROC curve)	0.8932	0.9085	**0.9196**



Train vs Test (일반화) : MinMax < Standard = Robust

Accuracy : MinMax < Standard = Robust

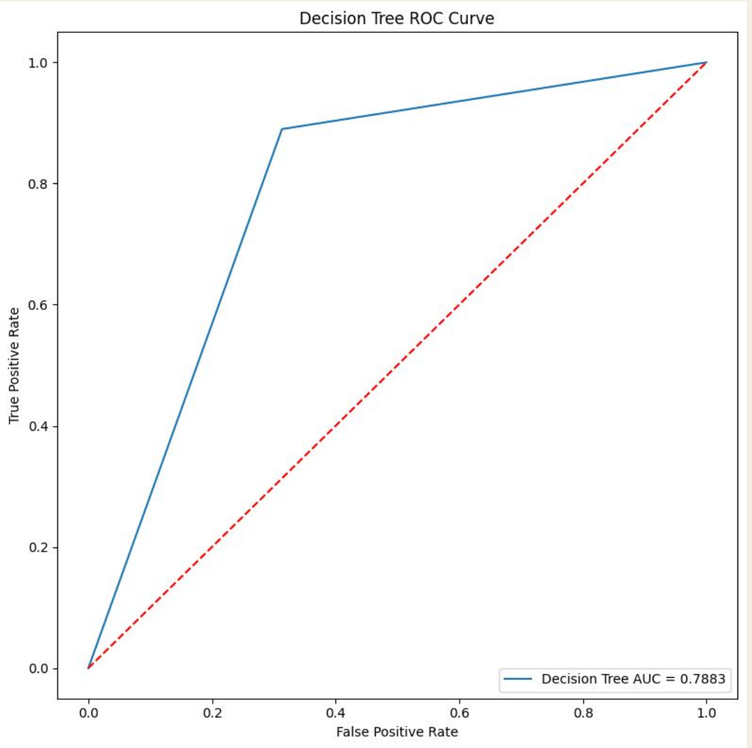
F1 : MinMax < Standard = Robust

AUC : MinMax < Standard < Robust

모델링 및 평가 분석 – Scaler

✔ Decision Tree

	MinMax	Standard	Robust
Train Accuracy	1	1	**1**
Test Accuracy	0.83	0.83	**0.83**
F1 (Churned / Stayed)	0.70	0.70	**0.70**
	0.88	0.88	**0.88**
AUC (ROC curve)	0.7883	0.7883	**0.7883**



Train vs Test (일반화) : MinMax = Standard = Robust

Accuracy : MinMax = Standard = Robust

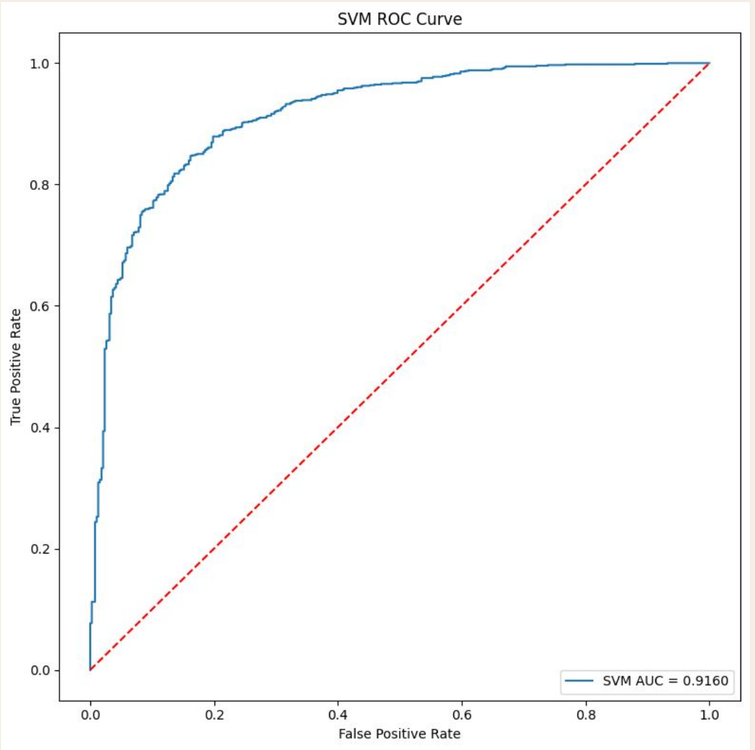
F1 : MinMax = Standard = Robust

AUC : MinMax = Standard = Robust

모델링 및 평가 분석 – Scaler

✔ SVM

	MinMax	Standard	Robust
Train Accuracy	0.8606	**0.8674**	**0.8687**
Test Accuracy	0.8498	**0.8581**	**0.8558**
F1 (Churned / Stayed)	0.72	**0.74**	**0.74**
	0.90	**0.90**	**0.90**
AUC (ROC curve)	0.9062	**0.9098**	**0.9160**



Train vs Test (일반화) : Standard > Robust > MinMax

Accuracy : Standard > Robust > MinMax

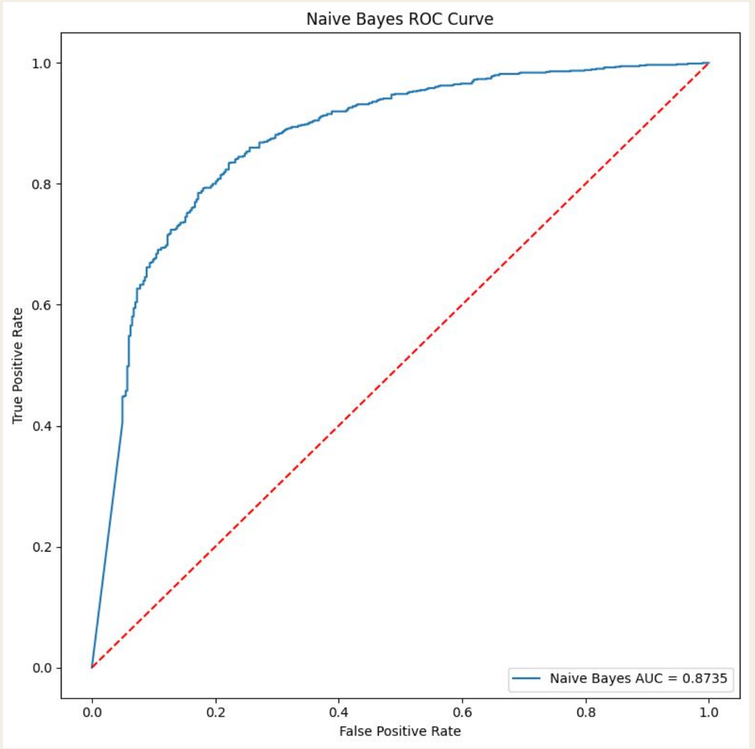
F1 : Standard = Robust > MinMax

AUC : Robust > Standard > MinMax

모델링 및 평가 분석 – Scaler

✔ Naïve Bayes

	MinMax	Standard	Robust
Train Accuracy	0.7663	**0.7662**	0.7663
Test Accuracy	0.7693	**0.7693**	0.7693
F1 (Churned / Stayed)	0.68	**0.68**	0.68
	0.82	**0.82**	0.82
AUC (ROC curve)	0.8735	**0.8736**	0.8735



Train vs Test (일반화) : MinMax = Standard = Robust

Accuracy : MinMax = Standard = Robust

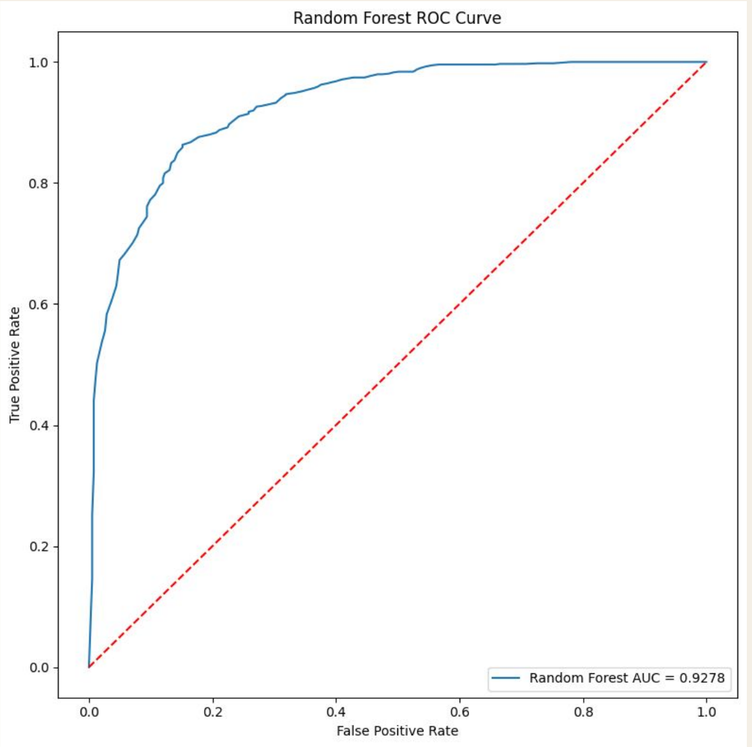
F1 : MinMax = Standard = Robust

AUC : Standard > MinMax = Robust

모델링 및 평가 분석 – Scaler

✔ Random Forest

	MinMax	Standard	Robust
Train Accuracy	1.0	1.0	**1.0**
Test Accuracy	0.8687	0.8687	**0.8695**
F1 (Churned / Stayed)	0.75	0.75	**0.75**
	0.91	0.91	**0.91**
AUC (ROC curve)	0.9279	0.9276	**0.9278**



Train vs Test (일반화) : Standard = Robust = MinMax

Accuracy : Robust > MinMax = Standard

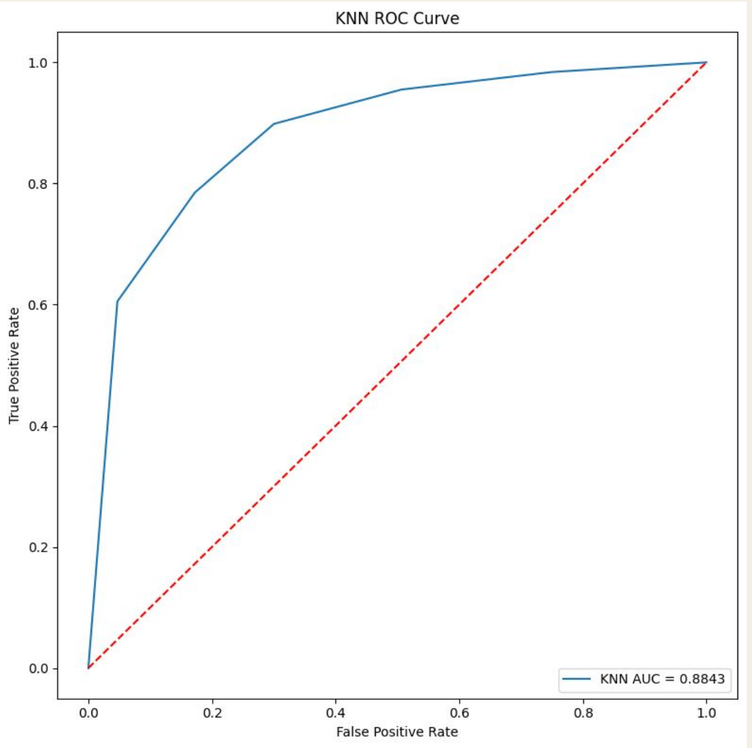
F1 : Standard = Robust = MinMax

AUC : MinMax > Robust > Standard

모델링 및 평가 분석 – Scaler

✔ KNN

	MinMax	Standard	Robust
Train Accuracy	0.8731	0.8793	**0.8810**
Test Accuracy	0.8316	0.8247	**0.8407**
F1 (Churned / Stayed)	0.70	0.69	**0.72**
	0.88	0.88	**0.89**
AUC (ROC curve)	0.8586	0.8712	**0.8843**



Train vs Test (일반화) : Robust > MinMax > Standard

Accuracy : Robust > MinMax > Standard

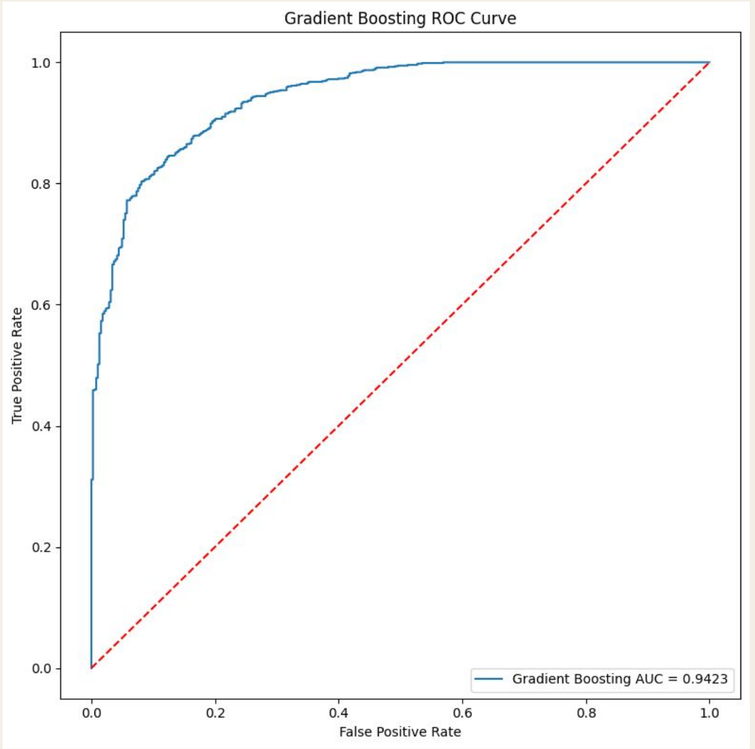
F1 : = Robust > MinMax > Standard

AUC : Robust > Standard > MinMax

모델링 및 평가 분석 – Scaler

✔ Gradient Boosting

	MinMax	Standard	Robust
Train Accuracy	0.8979	0.8979	**0.8979**
Test Accuracy	0.8786	0.8786	**0.8786**
F1 (Churned / Stayed)	0.77	0.77	**0.77**
	0.92	0.92	**0.92**
AUC (ROC curve)	0.9423	0.9423	**0.9423**



Train vs Test (일반화) : MinMax = Standard = Robust

Accuracy : MinMax = Standard = Robust

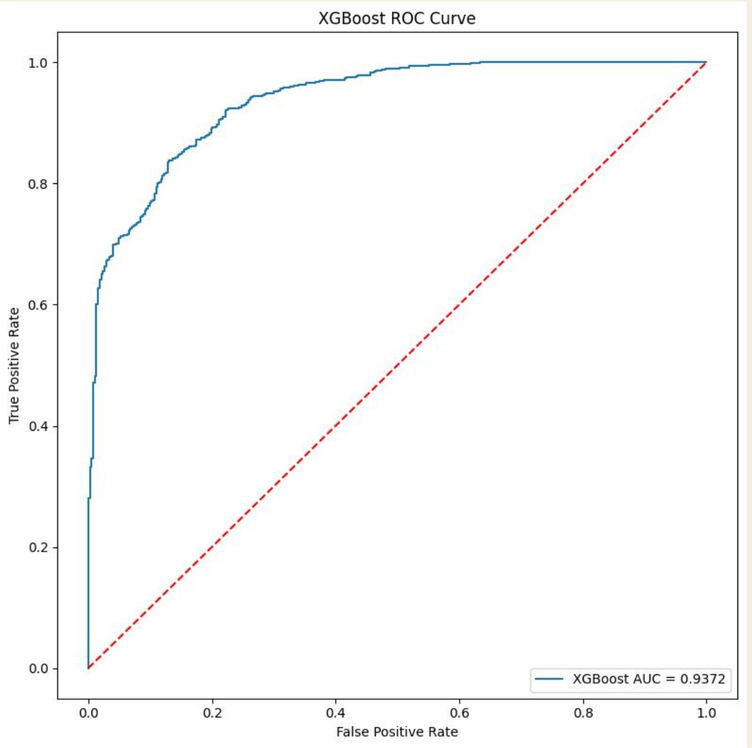
F1 : MinMax = Standard = Robust

AUC : MinMax = Standard = Robust

모델링 및 평가 분석 – Scaler

✔ XGBoost

	MinMax	Standard	Robust
Train Accuracy	0.9913	0.9913	**0.9913**
Test Accuracy	0.8824	0.8824	**0.8824**
F1 (Churned / Stayed)	0.78	0.78	**0.78**
	0.92	0.92	**0.92**
AUC (ROC curve)	0.9372	0.9372	**0.9372**



Train vs Test (일반화) : MinMax = Standard = Robust

Accuracy : MinMax = Standard = Robust

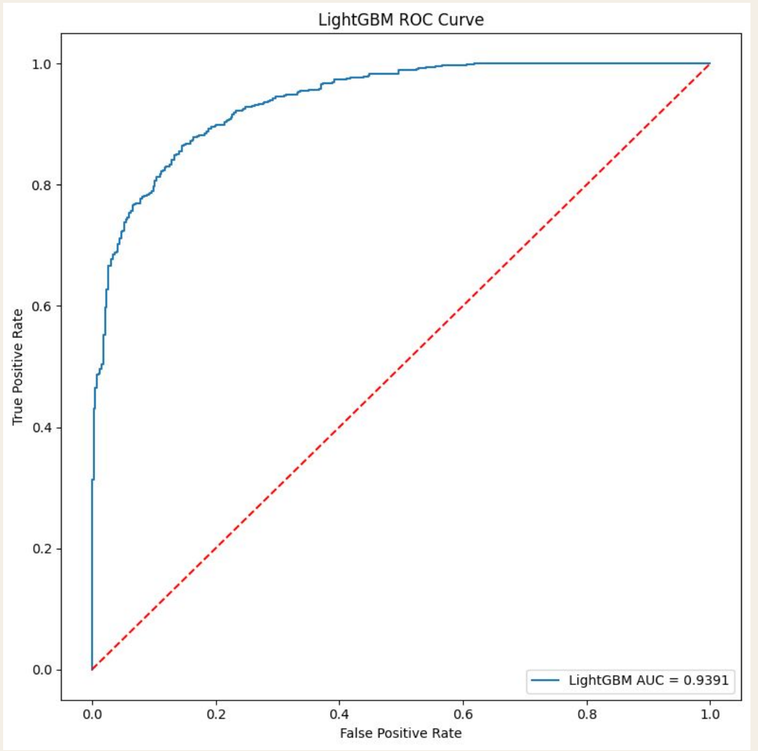
F1 : MinMax = Standard = Robust

AUC : MinMax = Standard = Robust

모델링 및 평가 분석 – Scaler

✔ LightGBM

	MinMax	Standard	Robust
Train Accuracy	0.9624	0.9621	**0.9598**
Test Accuracy	0.8718	0.8687	**0.8741**
F1 (Churned / Stayed)	0.76	0.76	**0.76**
	0.91	0.91	**0.91**
AUC (ROC curve)	0.9398	0.9398	**0.9391**



Train vs Test (일반화) : Robust > MinMax > Standard

Accuracy : MinMax = Standard = Robust

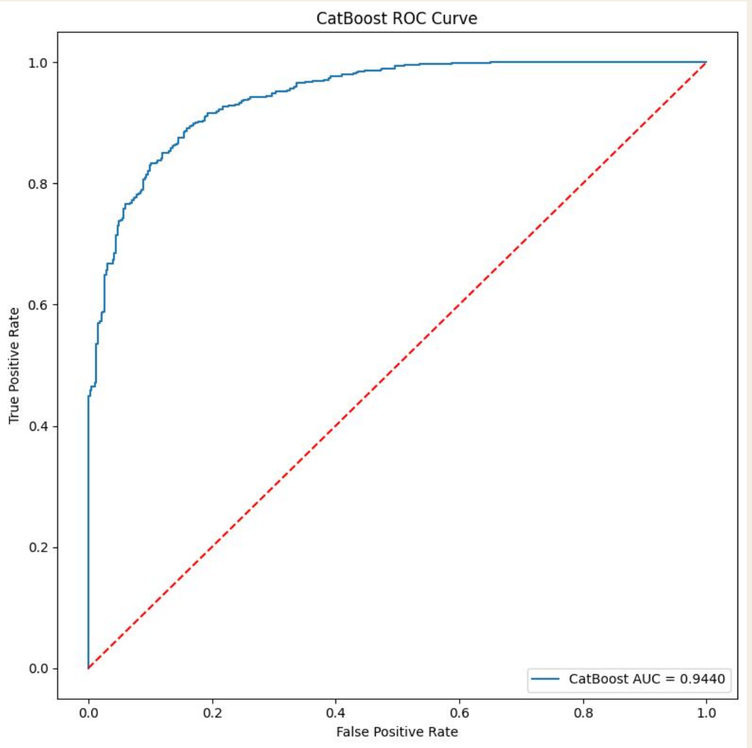
F1 : MinMax = Standard = Robust

AUC : MinMax = Standard = Robust

모델링 및 평가 분석 – Scaler

✔ CatBoost

	MinMax	Standard	Robust
Train Accuracy	0.9461	0.9461	**0.9461**
Test Accuracy	0.8756	0.8756	**0.8756**
F1 (Churned / Stayed)	0.77	0.77	**0.77**
	0.92	0.92	**0.92**
AUC (ROC curve)	0.9440	0.9440	**0.9440**



Train vs Test (일반화) : MinMax = Standard = Robust

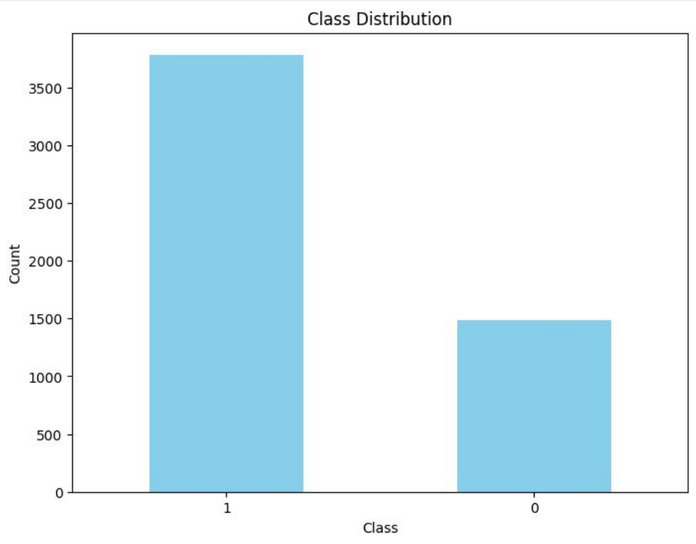
Accuracy : MinMax = Standard = Robust

F1 : MinMax = Standard = Robust

AUC : MinMax = Standard = Robust

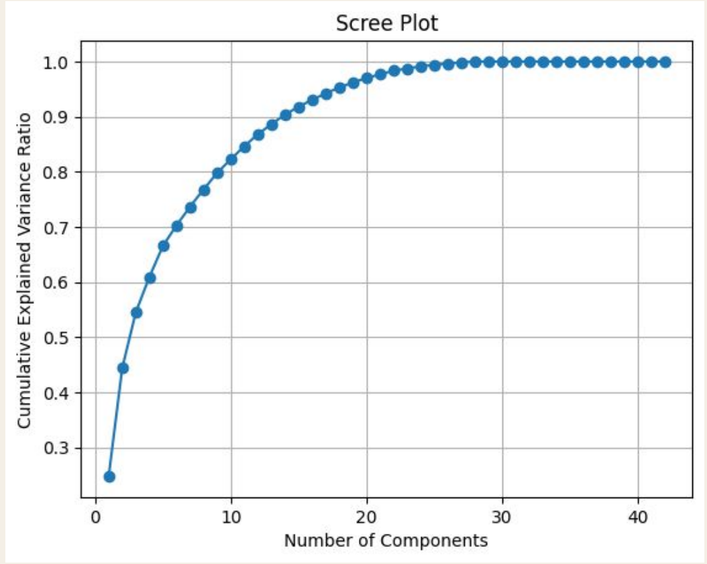
모델 선정

✔ Data Imbalance 처리



- Over Sampling: 성능이 오히려 저하됨

✔ 다중 공선성 처리 - PCA



PC 1, explained_variance: 24.77
PC 2, explained_variance: 44.40
PC 3, explained_variance: 54.51
PC 4, explained_variance: 60.90
PC 5, explained_variance: 66.54
PC 6, explained_variance: 70.26
PC 7, explained_variance: 73.65
PC 8, explained_variance: 76.81
PC 9, explained_variance: 79.78
PC10, explained_variance: 82.25
PC11, explained_variance: 84.62
PC12, explained_variance: 86.79
PC13, explained_variance: 88.62
PC14, explained_variance: 90.35
PC15, explained_variance: 91.79
PC16, explained_variance: 93.02
PC17, explained_variance: 94.21
PC18, explained_variance: 95.27
PC19, explained_variance: 96.19
PC20, explained_variance: 96.99
.
.

- 18 개로 차원축소(95% 데이터): 성능 저하

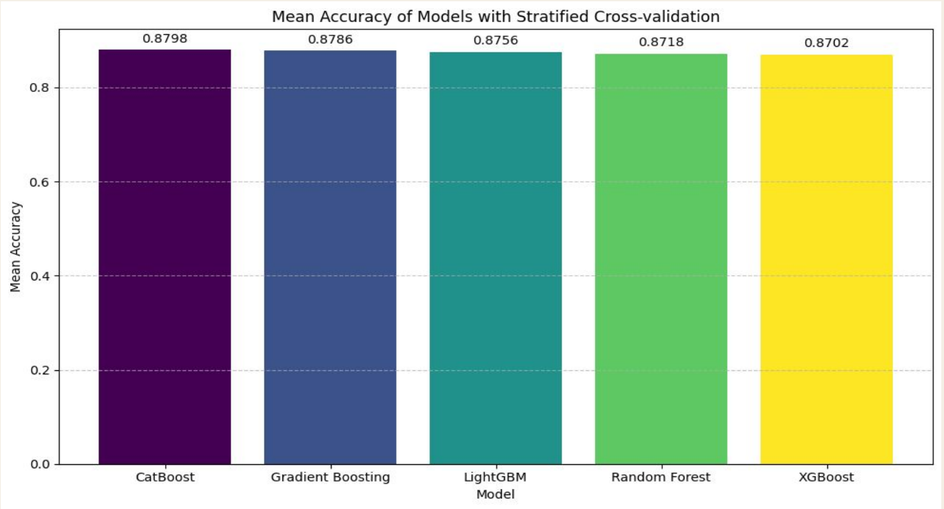
💡원 데이터 유지!!!

모델 선정

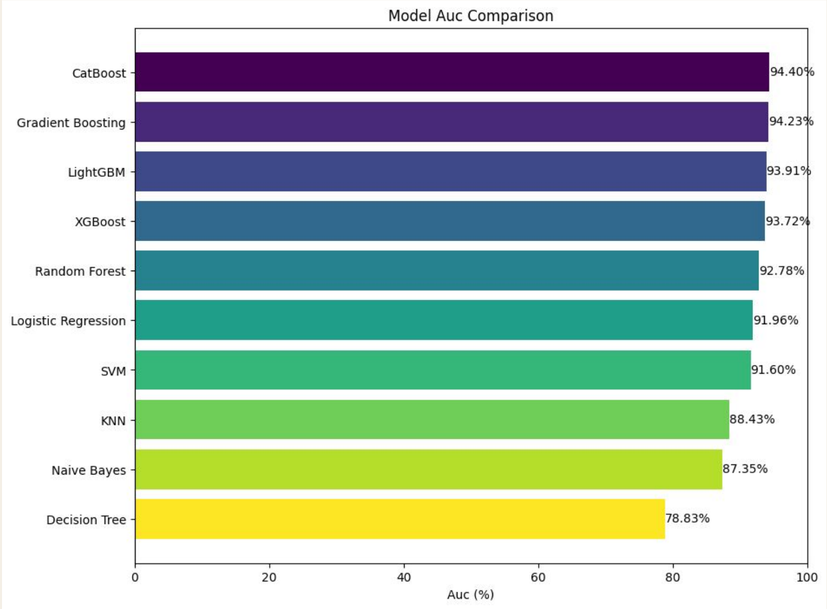
✔ Scaler : Robust

✔ Model:

	Catboost	Gradient	Xgboost
Train Accuracy	0.9461	0.8979	0.9913
Test Accuracy	0.8756	0.8786	0.8824
F1 (Churned / Stayed)	0.77	0.77	0.78
	0.92	0.92	0.92
AUC (ROC curve)	0.9440	0.9423	0.9372



Stratified K-Fold 교차 검증 결과



AUC Ranking

모델링 및 평가 분석 – 최적화

✔ CatBoost

	초기모델	**Grid Search**	Random Search	Feature Selection
Train Accuracy	0.9461	**0.8943**	0.8938	0.8924
Test Accuracy	0.8756	**0.8816**	0.8741	0.8809
F1 (Churned / Stayed)	0.77	**0.77**	0.76	0.77
	0.92	**0.92**	0.91	0.92
AUC (ROC curve)	0.9440	**0.9452**	0.9436	0.9447

- Grid Search 는 depth= 4, iterations=100, l2_leaf_reg=1, learning_rate=0.1일 때가 최적의 값
- Random Search는 depth=6, iterations=187, l2_leaf_reg=4.337086111390218, learning_rate=0.024286681792194 08일 때가 최적의 값
- 특성 중요도가 0인 변수는 제거, depth=4, iterations=100, l2_leaf_reg=1, learning_rate=0.1일 때가 최적의 값

💡 Grid Search 모델이 테스트 데이터에 대한 정확도가 가장 높기 때문에 가장 일반화 능력이 좋음.

모델링 및 평가 분석 – 최적화

✔ Gradient Boosting

	초기 모델	GridSearchCV	SMOTE-ENN	**Feature Selection**
Train Accuracy	0.8979	0.9087	0.9751	**0.8977**
Test Accuracy	0.8786	0.8725	0.8232	**0.8832**
F1 (Churned / Stayed)	0.77	0.76	0.75	**0.78**
	0.92	0.91	0.86	**0.92**
AUC (ROC curve)	0.9423	0.9423	0.9262	**0.9431**

- Grid Search : [n_estimators=100, learning_rate=0.05, max_depth=5] 일 때가 최적의 값 이지만 성능 하락
- SMOTE-ENN : 초기 모델보다 전체적으로 성능 하락
- Feature Selection : 특성 중요도 상위 25개를 사용했을 때 성능 향상

💡 특성 중요도 상위 25개를 선택 후 사용했을 때 일반화와 성능 향상했으며 F1 score가 증가 -> 가장 우수함

모델링 및 평가 분석 – 최적화

✔ XGBoost

	초기모델	Grid Search	**Optuna**	Feature Selection
Train Accuracy	0.9913	0.8939	**0.9218**	0.9224
Test Accuracy	0.8824	0.8763	**0.8862**	0.8771
F1 (Churned / Stayed)	0.78	0.76	**0.79**	0.77
	0.92	0.92	**0.92**	0.92
AUC (ROC curve)	0.9372	0.9426	**0.9437**	0.9422

- Grid Search: 'learning_rate': 0.2, 'max_depth': 3, 'min_child_weight': 2, 'n_estimators': 50 -> 성능 하락
- Optuna : 'max_depth': 5, 'learning_rate': 0.0817512560423265, 'n_estimators': 137 -> 최적!
- Feature Selection(특성중요도 상위 25개) -> 성능 하락

💡 Optuna 로 최적화 했을때 과적합도 어느 정도 감소 하였고 모델의 성능도 향상 되었다. Optuna 채택!

결론

✔ XGBoost 모델 Optuna 최적화

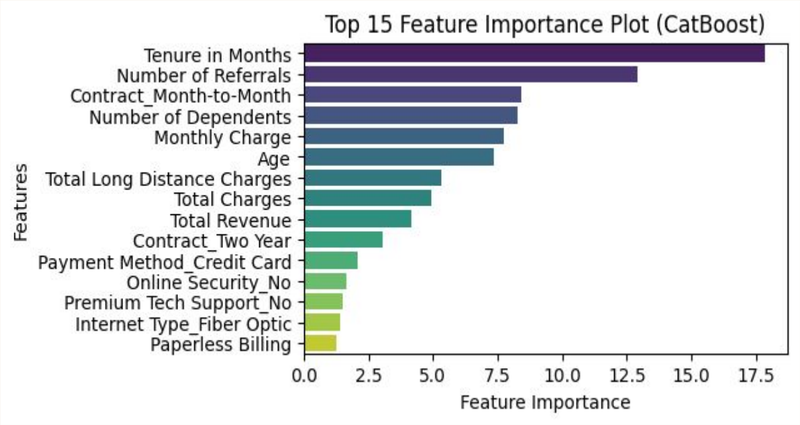
	Catboost (Grid Search)	**Gradient (상위 25개 특성)**	**XGboost (Optuna)**
Train Accuracy	0.8943	0.8977	**0.9218**
Test Accuracy	0.8816	0.8832	**0.8862**
F1 (Churned / Stayed)	0.77	0.78	**0.79**
	0.92	0.92	**0.92**
AUC (ROC curve)	**0.9452**	0.9431	0.9437

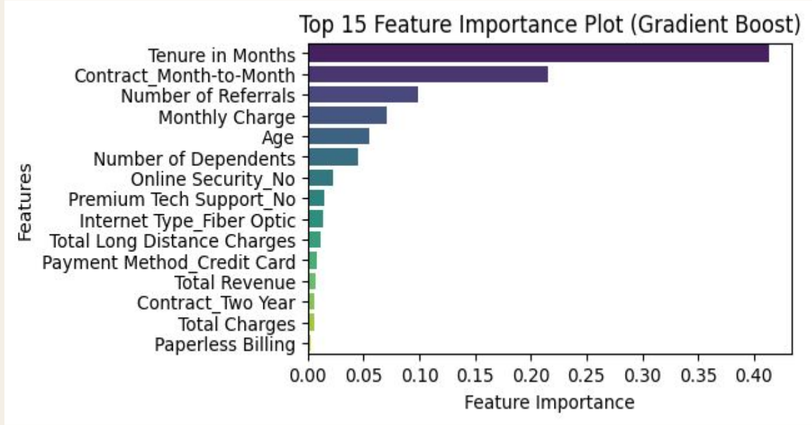
- XGBoost 모델은 정확도와 F1 Score에서 우수한 성능을 보임
- AUC 또한 경쟁 모델과 비교해 충분히 높은 성능을 제공

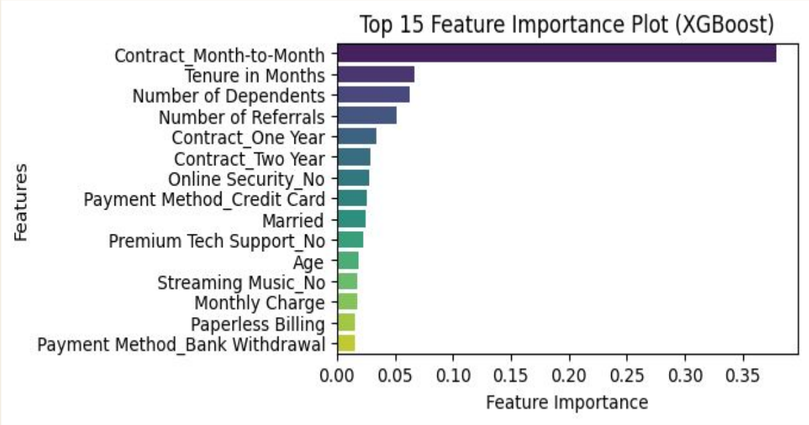
💡 하이퍼 파라미터가 'learning_rate': 0.08175125604232 65, 'max_depth': 5, 'n_estimators': 137인 XGBoost를 추천

결론

✔ Feature Importance (Top 15)







- 공통적으로 중요한 특성: Tenure in Months, Contract Month-to-Month, Number of Referrals, Number of Dependents
- 가입 기간: 장기 고객은 이탈 가능성이 낮으며, 이들을 위한 보상 제도
- 추천인 수: 추천인이 많은 고객은 서비스 만족도가 높고 이탈 가능성이 낮으므로, 추천 프로그램 강화가 유리
- 월 단위 계약: 유연성이 높아 이탈률이 증가할 수 있으므로 장기 계약을 유도
- 부양 가족 수: 부양 가족이 많을수록 서비스 유지 가능성이 높으므로, 가족 중심의 서비스 패키지 제공

감사합니다

End of Document

2024.05.15

6팀 1조