**Customer Churn Prediction**

이 데이터셋은 고객 이탈(Churn) 여부를 예측하는 문제를 다룹니다. 고객 이탈은 고객이 서비스를 중단하거나 다른 서비스로 전환하는 것을 의미합니다. 이 데이터셋은 다양한 고객 특성을 포함하고 있으며, 이를 통해 고객이 이탈할지 여부를 예측할 수 있습니다. 고객 이탈 예측은 비즈니스에서 매우 중요한 문제로, 고객 유지 전략을 수립하는 데 큰 도움이 됩니다.

customerID: 각 고객에 대한 고유 식별자입니다.

gender: 고객의 성별(Male, Female) 입니다.

SeniorCitizen: 고객이 고령자인지 여부 (1: Yes, 0: No).

Partner: 고객에게 배우자가 있는지 여부(Yes, No).

Dependents: 고객에게 부양가족이 있는지 여부(Yes, No)입니다.

tenure: 고객이 회사에 재직한 개월 수입니다.

PhoneService: 고객에게 전화 서비스가 있는지 여부(Yes, No)입니다.

MultipleLines: 고객이 여러 회선을 보유하고 있는지 여부(Yes, No, No phone service)입니다.

InternetService: 고객이 보유한 인터넷 서비스 유형(DSL, Fiber optic[광섬유], No)입니다.

OnlineSecurity: 고객에게 온라인 보안이 있는지 여부(Yes, No, No internet service)입니다.

OnlineBackup: 고객이 온라인 백업을 사용하는지 여부(Yes, No, No internet service).

DeviceProtection: 고객에게 장치 보호 기능이 있는지 여부(Yes, No, No internet service).

TechSupport: 고객이 기술 지원을 받는지 여부(Yes, No, No internet service).

StreamingTV: 고객이 스트리밍 TV를 사용 중인지 여부(Yes, No, No internet service).

StreamingMovies: 고객에게 스트리밍 영화가 있는지 여부(Yes, No, No internet service).

Contract: 고객의 계약 기간(Month-to-month, One year, Two year).

PaperlessBilling: 고객의 종이 없는 청구서 사용 여부(Yes, No).

PaymentMethod: 고객의 결제 방법(Electronic check[전자 수표], Mailed check[우편 수표], Bank transfer[은행 송금], Credit card[신용카드])입니다.

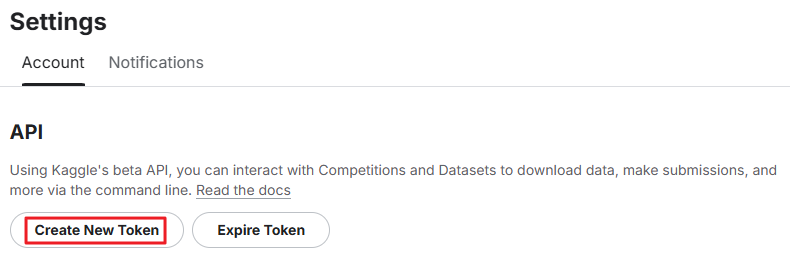
MonthlyCharges: 고객에게 매월 청구되는 금액입니다.

TotalCharges: 고객에게 청구된 총 금액입니다.

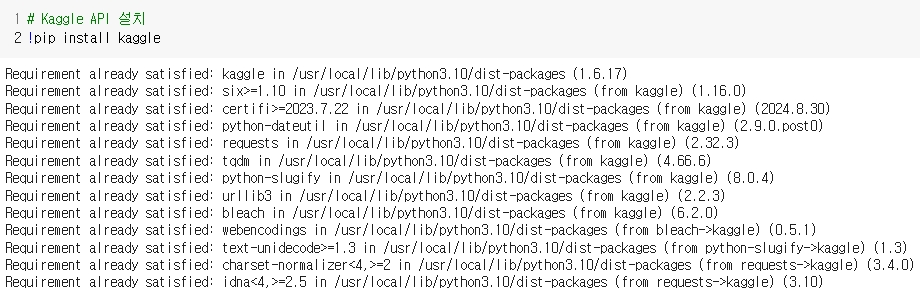
Churn: 고객이 이탈했는지 여부(Yes, No)입니다.

1. **Kaggle 데이터셋 불러오기**

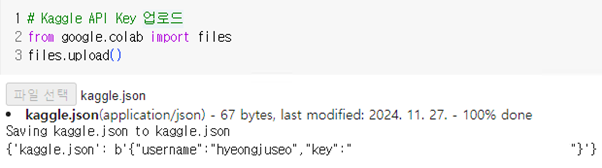
Kaggle의 Settings – Account - API에 들어가서 Create New Token을 클릭하면 “kaggle.json” 파일을 다운로드 받을 수 있습니다.



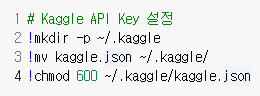
Kaggle API를 설치합니다.



다운로드 받은 “kaggle.json” 파일을 업로드합니다.



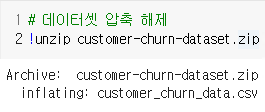
Kaggle API Key를 설정합니다.



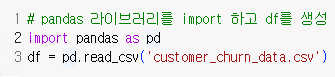
데이터셋을 다운로드합니다.



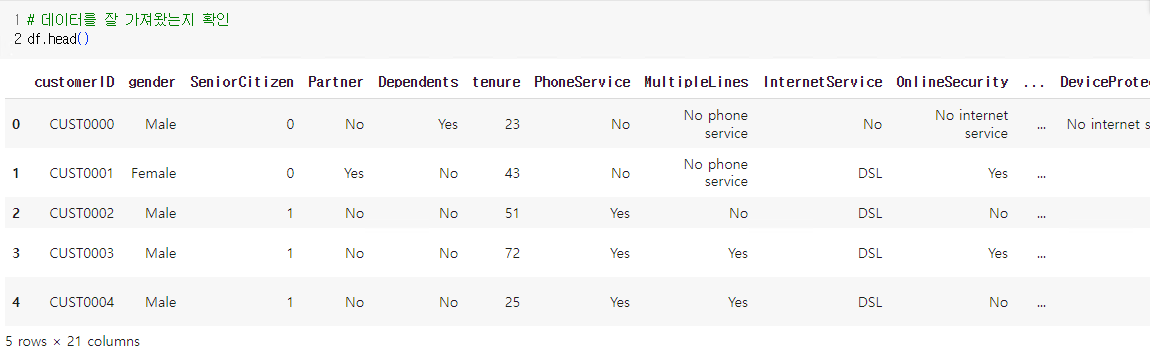
데이터셋의 압축을 해제합니다.



Pandas 라이브러리를 import 하고 df를 생성합니다.

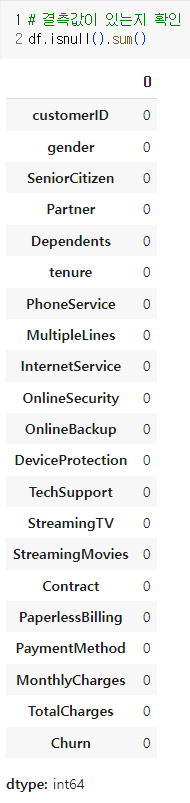


데이터를 잘 가져왔는지 확인합니다.

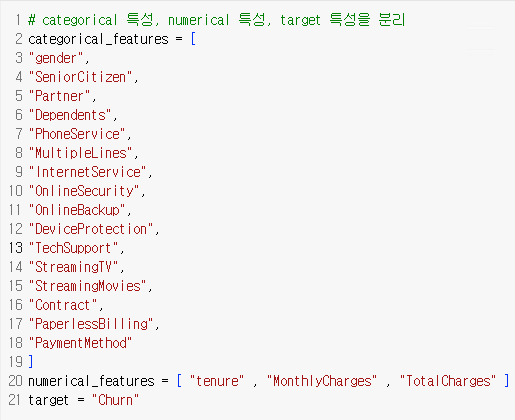


1. **시각화 확인**

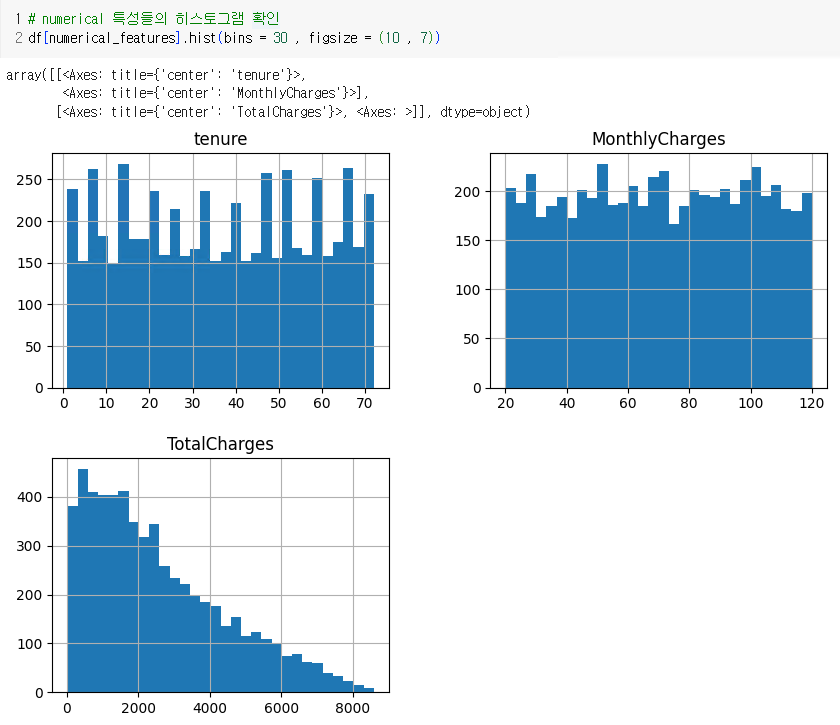
결측값이 있는지 확인합니다.



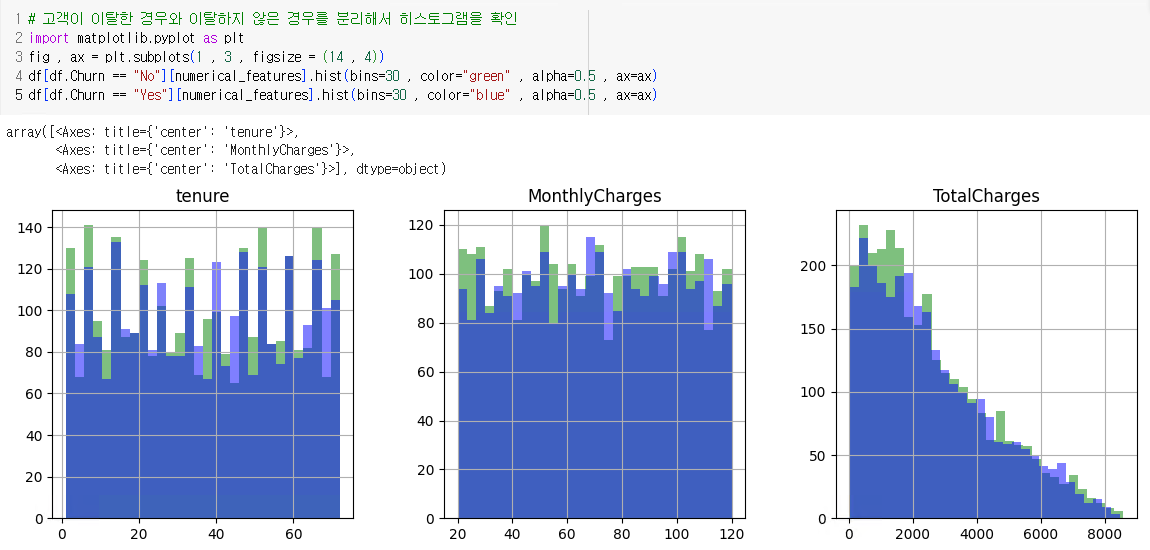
Categorical 특성, Numerical 특성, target 특성을 분리합니다.



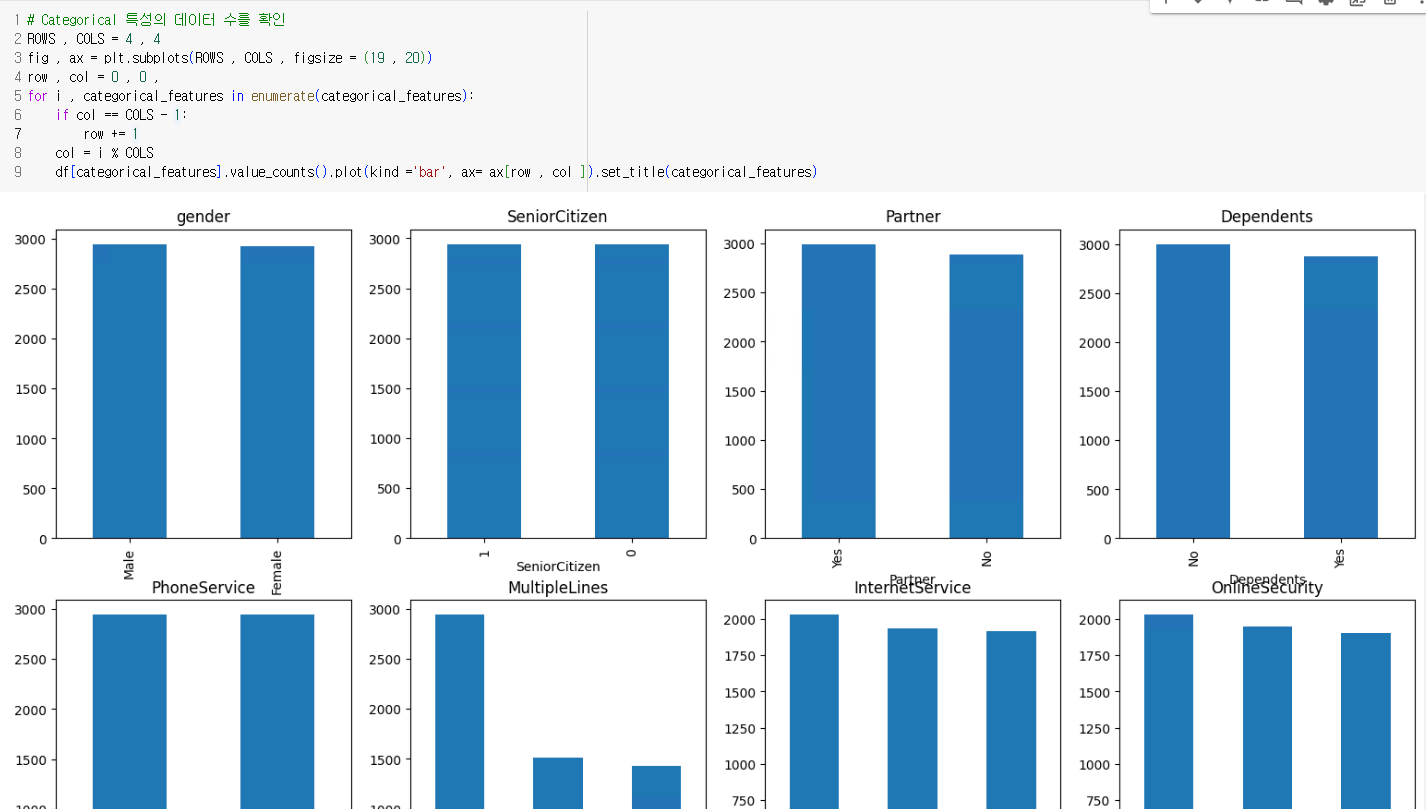
Numerical 특성들의 히스토그램을 시각화로 확인합니다.



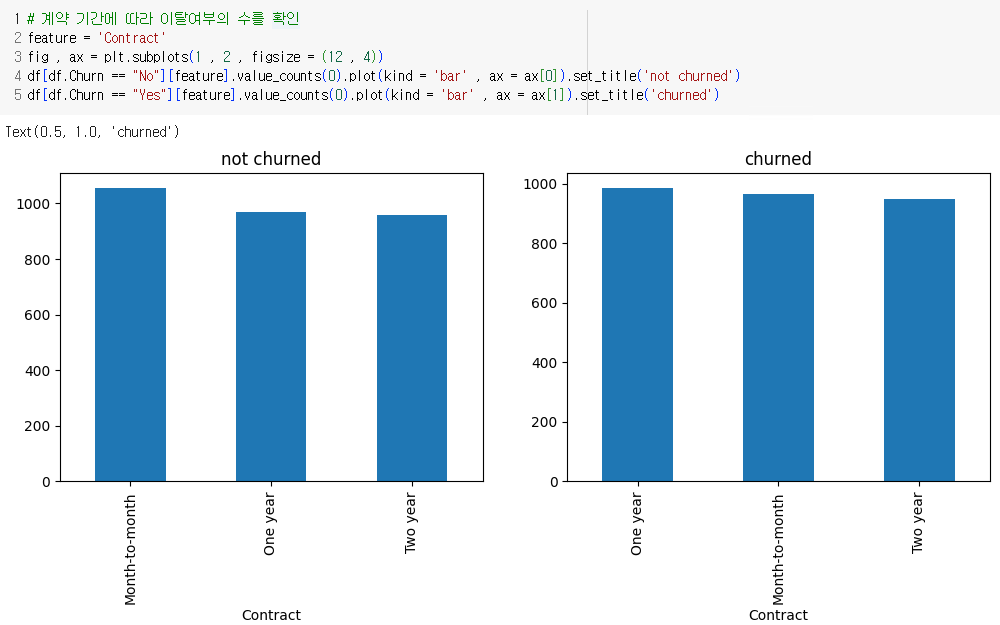
고객이 이탈한 경우와 이탈하지 않은 경우를 분리해서 히스토그램을 확인합니다.



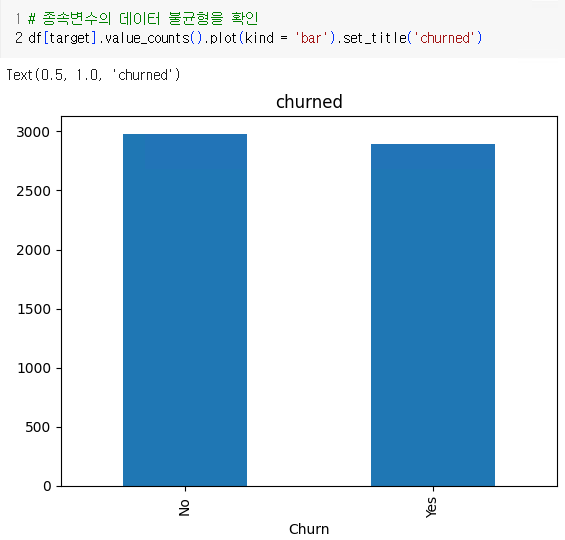
Categorical 특성의 데이터 수를 확인합니다.



계약 기간에 따라 이탈여부의 수를 확인합니다.



종속변수의 데이터 불균형을 확인합니다.



1. **ML 모델 선정**

저는 로지스틱 회귀(Logistic Regression) 모델을 선택하겠습니다. 이 모델을 선택하는 이유는 여러 가지가 있습니다.

첫째, 로지스틱 회귀는 이진 분류 문제에 매우 적합합니다. 종속 변수가 이진(0 또는 1)인 경우에 로지스틱 회귀는 효과적으로 작동합니다. 고객 이탈 예측 문제는 고객이 이탈할지(1) 아니면 이탈하지 않을지(0)를 예측하는 이진 분류 문제이므로, 로지스틱 회귀가 적합한 선택입니다.

둘째, 로지스틱 회귀 모델은 해석 가능성이 높습니다. 모델의 계수(coefficient)는 각 독립 변수(특성)가 종속 변수(결과)에 미치는 영향을 나타내며, 이를 통해 어떤 특성이 고객 이탈에 중요한 역할을 하는지 쉽게 이해할 수 있습니다. 이는 비즈니스 의사 결정에 중요한 정보를 제공할 수 있습니다.

셋째, 로지스틱 회귀는 결과가 특정 클래스에 속할 확률을 예측할 수 있습니다. 예를 들어, 고객이 이탈할 확률을 예측할 수 있으며, 이를 통해 더 정교한 의사 결정을 할 수 있습니다. 예측된 확률을 기반으로 특정 임계값(threshold)을 설정하여 이탈 여부를 결정할 수 있습니다.

넷째, 로지스틱 회귀는 계산이 효율적입니다. 대규모 데이터셋에서도 빠르게 학습할 수 있으며, 과적합(overfitting)을 방지하기 위해 정규화(regularization) 기법을 쉽게 적용할 수 있습니다. 이는 모델의 성능을 안정적으로 유지하는 데 도움이 됩니다.

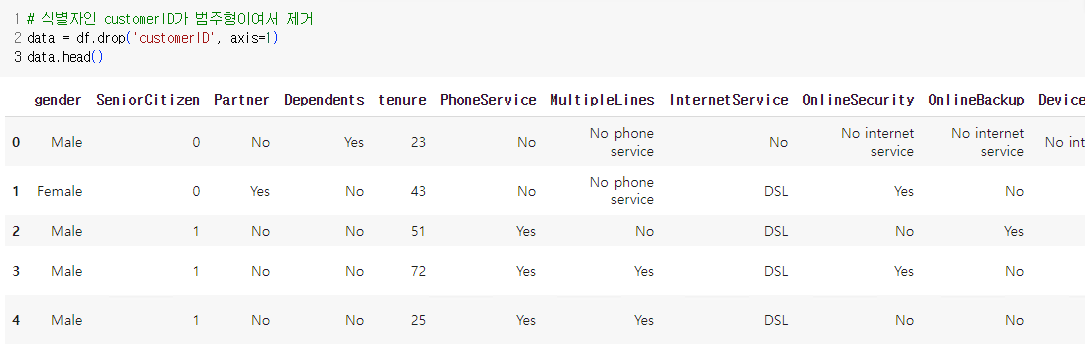
다섯째, 로지스틱 회귀는 독립 변수와 종속 변수 사이의 선형 관계를 가정합니다. 이 가정이 만족되는 경우, 로지스틱 회귀는 매우 효과적인 모델이 될 수 있습니다. 또한, 독립 변수 간의 다중공선성(multicollianearity)에 민감하지 않으며, 다중공선성이 있는 경우에도 비교적 안정적인 성능을 보입니다.

마지막으로, 로지스틱 회귀는 비교적 단순한 모델로, 많은 가정을 필요로 하지 않습니다. 이는 모델을 이해하고 해석하는 데 도움이 되며, 복잡한 모델에 비해 과적합의 위험이 적습니다. 또한, 로지스틱 회귀는 다중 클래스 분류 문제로 확장할 수 있어 다양한 분류 문제에 적용할 수 있습니다.

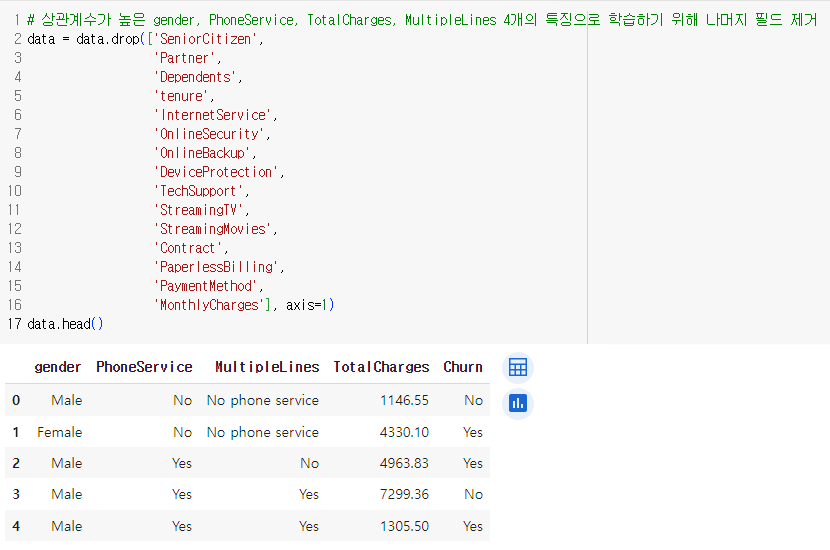
다중공선성은 독립 변수들 간에 강한 상관관계가 존재하는 경우를 말합니다. 이는 모델의 성능과 신뢰성에 부정적인 영향을 미칠 수 있습니다.

1. **데이터셋 분리**

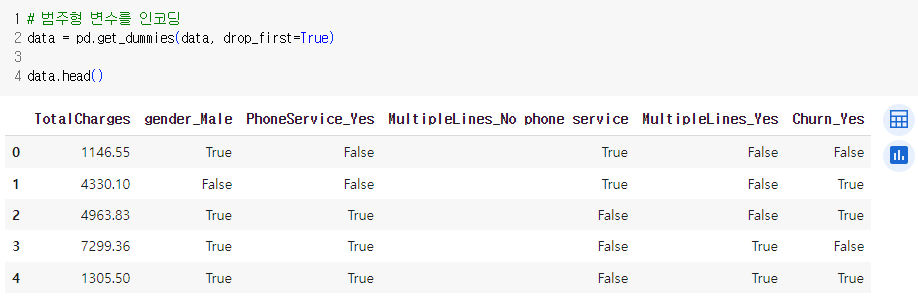
데이터셋을 분리하기 전에 전처리 먼저 진행합니다. 식별자인 customerID를 제거



상관계수가 높은 gender, PhoneService, TotalCharges, MultipleLines 4개의 특징으로 학습하기 위해 나머지 필드를 제거합니다.



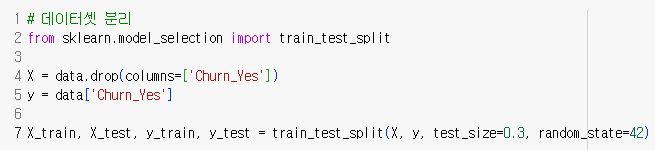
범주형 변수를 인코딩합니다. (drop\_first=True는 첫번째 카테고리를 제거합니다. 다중공선성을 방지하는 기능을 합니다.)



TotalCharges를 StandardScaler를 이용해서 스케일링합니다. (모델의 성능을 향상시킵니다.)

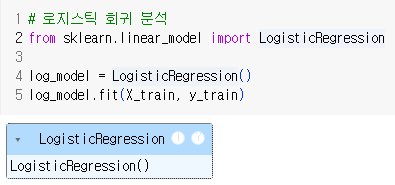


데이터셋을 학습용과 검증용으로 분리합니다.



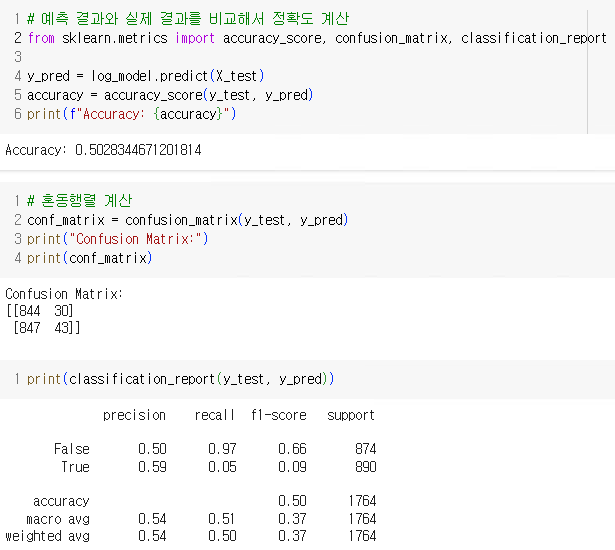
1. **학습 실행**

로지스틱 회귀 분석 모델로 학습합니다.



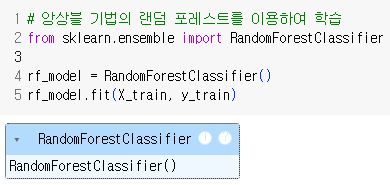
1. **새로운 데이터로 예측 후 평가결과 확인**

예측 결과와 실제 결과를 비교해서 정확도 및 혼동행렬을 계산합니다.



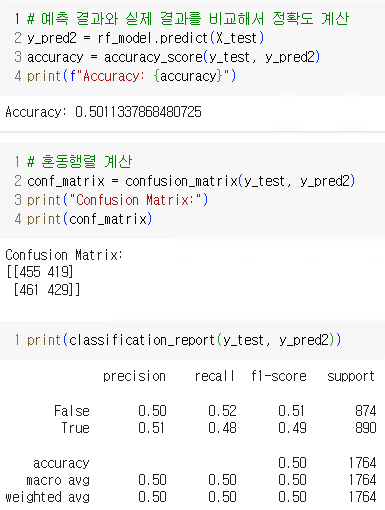
1. **최적화 실행하여 학습 재실행**

랜덤 포레스트 모델로 학습합니다.



1. **새로운 데이터로 예측 후 평가결과 확인**

예측 결과와 실제 결과를 비교해서 정확도 및 혼동행렬을 계산합니다.



1. **최종 분석결과**

이번 분석에서는 고객 이탈 예측 문제를 해결하기 위해 로지스틱 회귀(Logistic Regression)와 랜덤 포레스트(Random Forest) 두 가지 모델을 사용하여 성능을 비교하였습니다. 각 모델의 성능을 평가하기 위해 “classification\_report”를 사용하여 정밀도(precision), 재현율(recall), F1-스코어(f1-score), 및 정확도(accuracy)를 계산하였습니다.

로지스틱 회귀 모델의 결과를 보면, False클래스에 대한 재현율(recall)이 0.97로 매우 높은 반면, True클래스에 대한 재현율은 0.05로 매우 낮습니다. 이는 모델이 대부분의 경우 False클래스를 예측하는 경향이 있음을 나타냅니다. True클래스에 대한 F1-스코어는 0.09로 매우 낮아, 모델이 True클래스를 잘 예측하지 못하고 있음을 알 수 있습니다. 전체 정확도는 0.50으로, 모델이 무작위로 예측하는 것과 큰 차이가 없습니다.

랜덤 포레스트 모델의 결과를 보면, False클래스와 True클래스에 대한 정밀도(precision)와 재현율(recall)이 각각 0.50과 0.51로 거의 동일합니다. 이는 모델이 두 클래스를 균형 있게 예측하고 있음을 나타냅니다. True클래스에 대한 F1-스코어는 0.50으로, 로지스틱 회귀 모델보다 훨씬 높은 값을 보입니다. 전체 정확도는 0.50으로, 로지스틱 회귀 모델과 동일합니다.

이번 분석을 통해 로지스틱 회귀와 랜덤 포레스트 모델의 성능을 비교해본 결과, 두 모델 모두 전체 정확도는 0.50으로 동일하였으나, 세부 지표에서 차이가 있었습니다. 로지스틱 회귀 모델은 False클래스를 매우 잘 예측하는 반면, True클래스를 거의 예측하지 못하는 경향이 있었습니다. 반면 랜덤 포레스트 모델은 두 클래스를 균형 있게 예측하였으며, True클래스에 대한 F1-스코어가 로지스틱 회귀 모델보다 훨씬 높았습니다.

랜덤 포레스트 모델은 앙상블 학습 기법을 사용하여 여러 개의 결정 트리를 결합함으로써, 단일 모델보다 더 나은 성능을 보일 수 있습니다. 이번 분석에서도 랜덤 포레스트 모델이 True클래스를 예측하는 데 더 나은 성능을 보였음을 확인할 수 있었습니다.

결론적으로, 고객 이탈 예측 문제에서는 랜덤 포레스트 모델이 로지스틱 회귀 모델보다 더 균형잡힌 예측 성능을 보였으며, 특히 True클래스를 예측하는 데 더 우수한 성능을 보였습니다. 따라서, 고객 이탈 예측 문제를 해결하기 위해 랜덤 포레스트 모델을 사용하는 것이 더 적합하다고 판단됩니다.