

고객이탈 예측 분류 모델링

Project
of Section 2

CONTENTS

1. 데이터셋 설명
2. 데이터 전처리 및 특성 공학
3. EDA
4. 모델링
5. 결과

Dataset Information

Dataset Information

01 활용 데이터

가상의 통신회사 고객 데이터셋 (Kaggle)

- 고객 개인정보 (성별, 나이대, 혼인여부, 부양가족여부)
- 고객 가입정보 (가입기간, 계약유형, 결제방법, 청구요금 등)
- 고객 가입서비스 (전화, 인터넷, 기타 부가서비스들)
- 지난달 고객의 이탈 여부

02 선정 이유

Section 2 학습 내용을 적용해볼 수 있는 비즈니스 관련 데이터

Dataset Information

03 데이터 활용 목적 & 문제 정의

: 기존 고객의 이탈 최대한 방지해야한다.

- **WHY?** 기존 고객 유지 <<<<< 신규 고객 확보
- **HOW?** 기존 고객 정보 데이터를 분류 문제로 접근하여 이탈 예측

04 가설

: 머신러닝 모델을 이용하여 고객의 이탈여부를 예측할 수 있다.

- 타겟 : 고객의 이탈여부
- 베이스라인 : 훈련데이터 타겟의 최대 빈도수 이용
- 평가지표 : 재현율

Dataset Information

05 Recall (재현율) 선택한 이유

	Target	
	Count	Percentage
No	5174	73%
Yes	1869	27%

Imbalanced Class

Case 1

서비스 이탈할 고객인데, 유지할 것이라고 잘못 예측해서 어떠한 조치도 하지 않는 오류 발생

▶ **그대로 고객을 잃어버린다.**

Case 2

서비스 유지할 고객인데, 이탈할 것이라고 잘못 예측해서 조치를 취해주는 오류 발생

▶ 그 고객은 어차피 유지할 고객이기 때문에 그대로 남아있는다.

결론 : Case1 의 오류가 더 크리티컬하기때문에 **recall** 을 높이는데 집중해야 한다.

Dataset Processing & Feature Engineering

Data Processing

Raw Data

7042

ID 제거



중복데이터 발생

Customer ID

모든 정보 일치
데이터 제거(22)

7020

데이터 형변환
(object -> float)



결측치 발생

TotalCharges

TotalCharges
공백 데이터 제거(11)

7009

Feature Engineering

새로운 특성 생성

1. BothServices

메인서비스인 '전화'와 '인터넷' 모두 사용하는지 여부 (1 - 모두사용/ 0 - 하나만 사용)

2. NubAdd

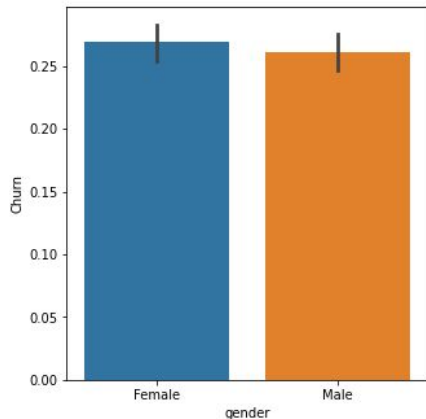
부가서비스 사용 갯수 (0~7)

- **MultipleLines** : 전화 사용시 / 다회선
- **OnlineSecurity** : 인터넷 사용시 / 온라인보안서비스
- **OnlineBackup** : 인터넷 사용시 / 온라인백업서비스
- **DeviceProtection** : 인터넷 사용시 / 장치보호플랜
- **TechSupport** : 인터넷 사용시 / 기술지원플랜
- **StreamingTV** : 인터넷 사용시 / TV 스트리밍 서비스
- **StreamingMovies** : 인터넷 사용시 / Movie 스트리밍 서비스

EDA

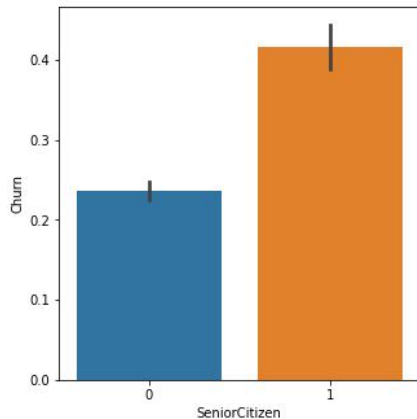
Features(personal) x Target

Gender



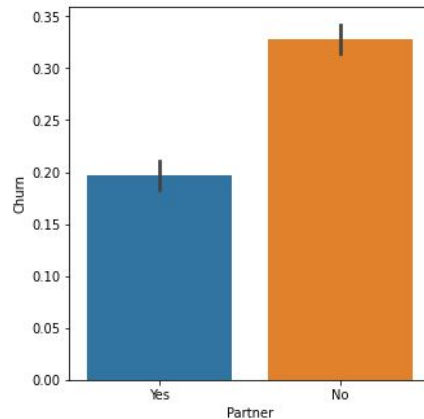
남녀
이탈 =

SeniorCitizen



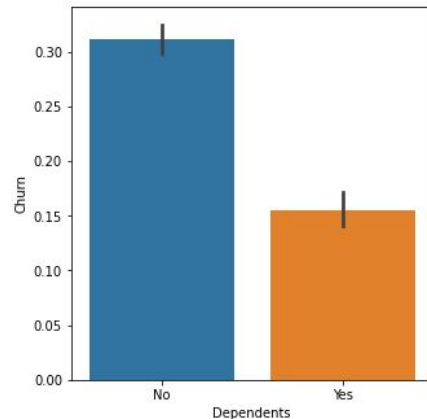
나이가 많은사람
이탈 ↑

Partner



결혼 안한사람
이탈 ↑

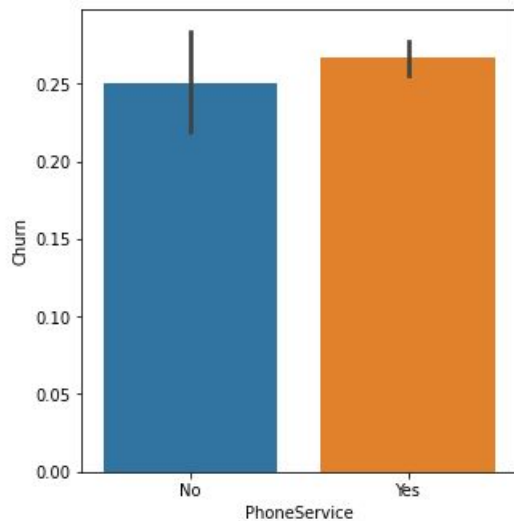
Dependents



부양가족 없는사람
이탈 ↑

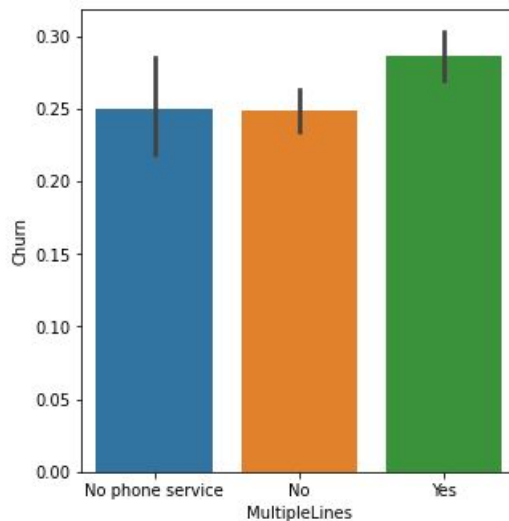
Features(Phone) x Target

Phone Service



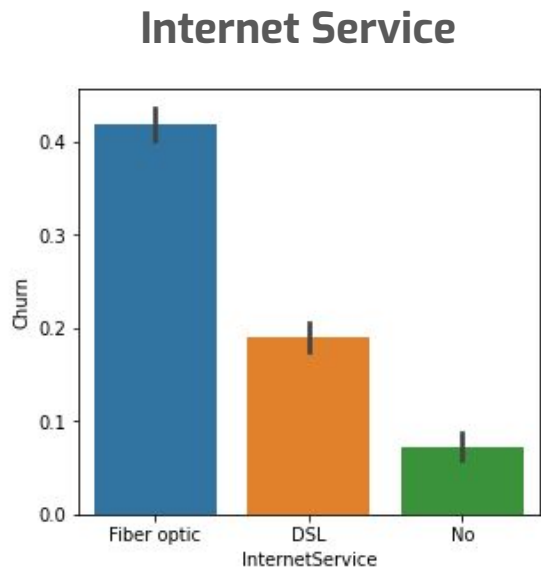
통화서비스 사용여부
이탈 =

Multiple Lines



다회선 사용여부
이탈 =

Features(Internet) x Target



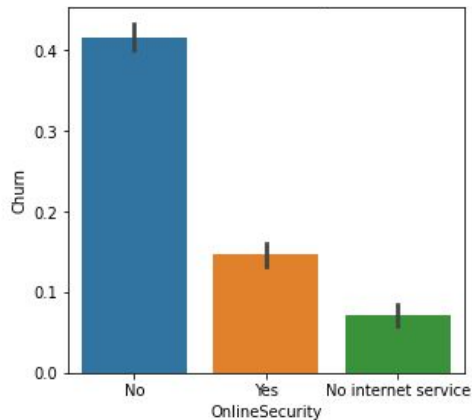
인터넷서비스 사용여부

이탈 비교

Fiber optic > DSL > No

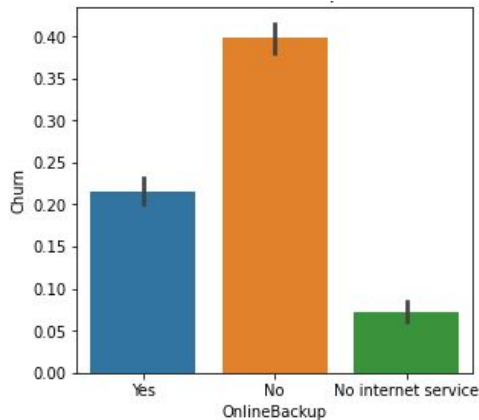
Features(Internet) x Target

OnlineSecurity



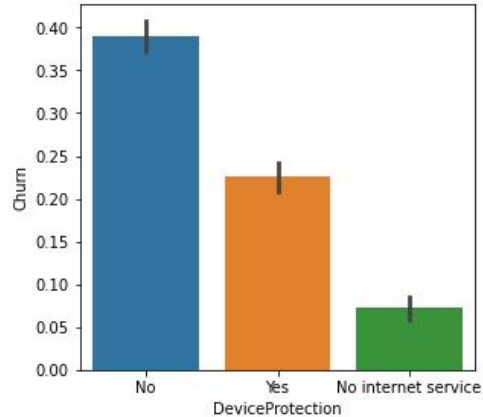
온라인보안서비스 No
이탈 ↑

OnlineBackup



온라인백업서비스 No
이탈 ↑

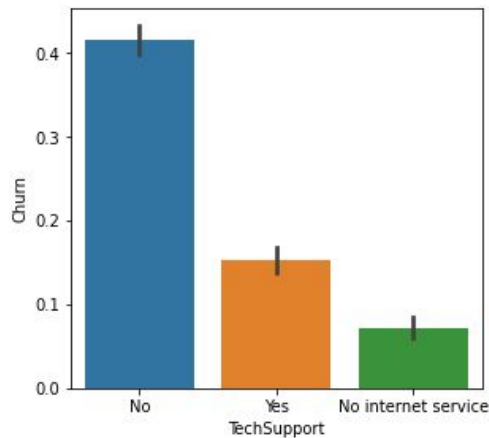
DeviceProtection



장치보호플랜 No
이탈 ↑

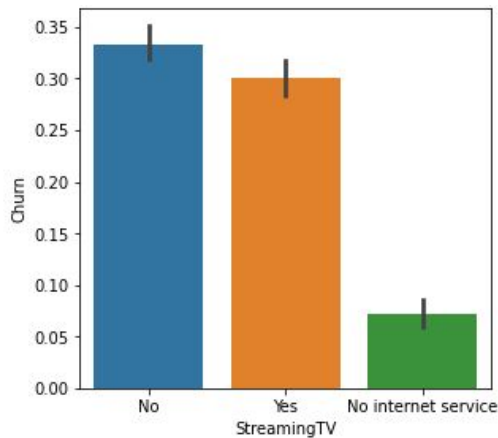
Features(Internet) x Target

TechSupport



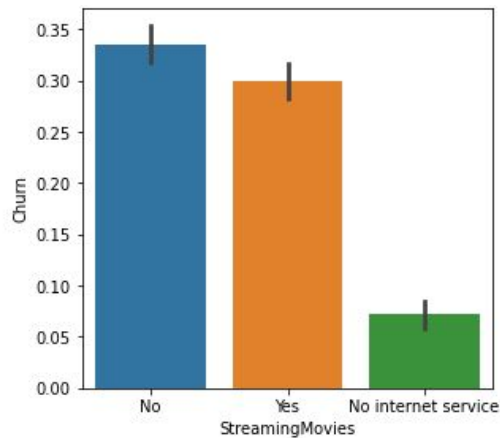
기술지원플랜 No
이탈 ↑

StreamingTV



스트리밍 TV No
이탈 ↑

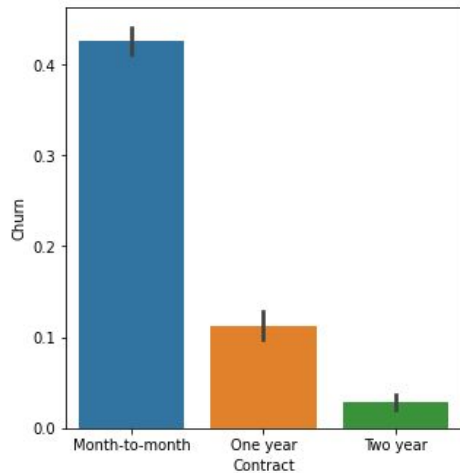
StreamingMovies



스트리밍 영화 No
이탈 ↑

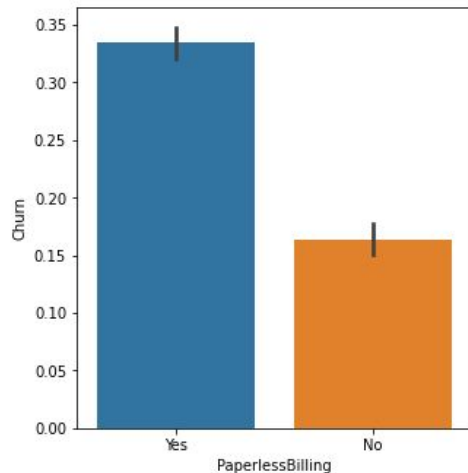
Features(Account) x Target

Contract



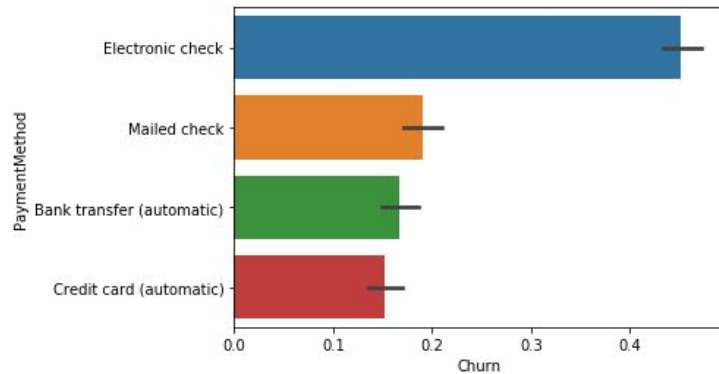
계약기간 짧은수록
이탈 ↑

PaperlessBilling



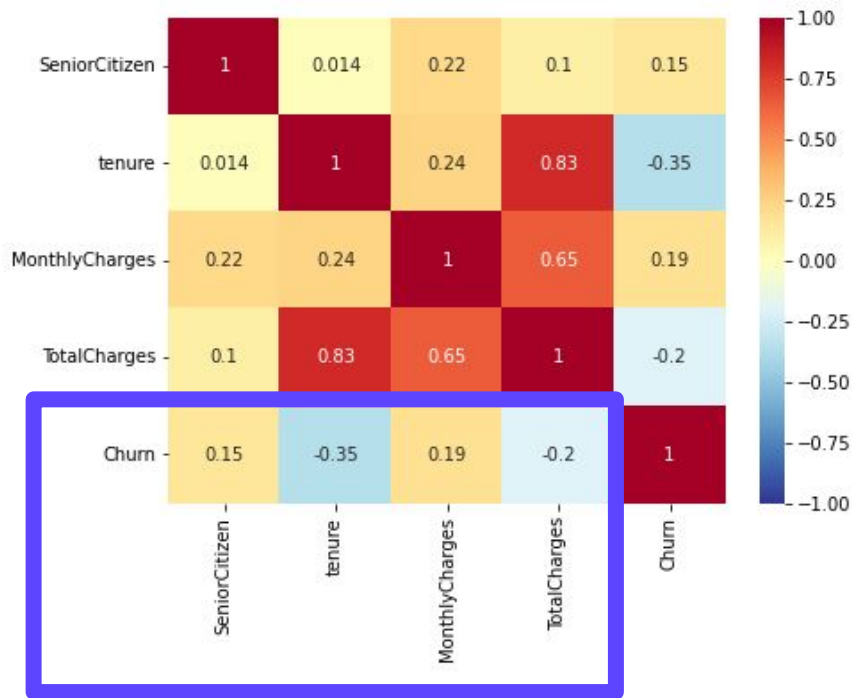
전자청구서 Yes
이탈 ↑

PaymentMethod







전자수표
이탈 ↑

Corr heatmap



✓ 수치형 변수들과 타겟간에 상관관계는 거의 없음

1. 'SeniorCitizen' : 양의 상관관계. (노인들이 이탈 )
2. 'MonthlyCharges' : 양의 상관관계. (월요금이 높을수록 이탈 )
3. 'TotalCharges' : 음의 상관관계. (총요금이 높을수록 이탈 )
4. 'tenure' : 음의 상관관계. (가입기간 오래될수록 이탈 )

✓ 'tenure'와 'TotalCharges'와 강한 양의 상관관계
(가입된지 오래될수록 총요금이 많다)

ML Modeling

ML Modeling

모델 학습하기 전에

1. 범주형 Features 오디널 인코딩 ▶ 파이프라인 내에서 실행
2. Imbalanced Target Class ▶ **Class weight** 조절
3. 교차 검증 ▶ **RandomizedSearchCV**
4. 하이퍼파라미터 튜닝 ▶ **RandomizedSearchCV**

Model Comparison

#	Classifier Model	accuracy	recall	f1	roc_auc
1	RandomForest	0.826	0.874	0.718	0.921
2	XGBoost	0.820	0.890	0.716	0.925
3	LightGBM	0.822	0.902	0.720	0.921

Test Score

Baseline

LightGBM

Accuracy

Accuracy

Recall

F1

0.735

0.761

0.749

0.624

Result

LightGBM

더 탐구해야 할 부분:

- 고객 이탈에 영향을 미치는 특성 파악 필요
- 성능을 더 높일 수 있는 모델로 보완 필요

모델이 완성된 후 활용 방안:

- 이탈을 하려는 고객의 유형을 예측
- 더 이상 서비스를 이용하지 않고 이탈하려는 이유 파악
- 이탈을 막기위한 계획 수립 (ex 집중적인 고객 유지 프로그램을 개발 등)

-> 고객 유지 개선