

DArt-B 토이프로젝트

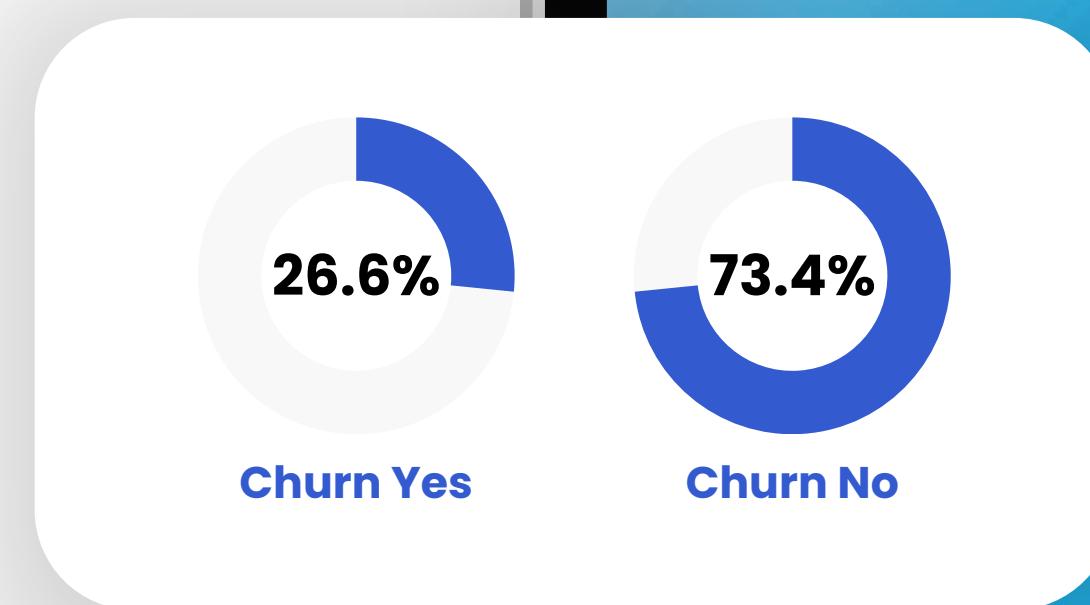
Telco Customer Churn

Explore Now



선지국밥주서영

김선교
장선희
국지희
황서영



Churn Yes

Churn No



Contents

-
- 01 분석 목적
 - 02 데이터 소개
 - 03 결측치 확인 및 처리
 - 04 EDA
 - 05 가설 설정 및 검정
 - 06 모델링
 - 07 결론

분석 목적

01. 분석 목적

통신사 고객 이탈 관련 기사

④ 천지일보

[경제인사이드] 내실 다지는 이통3사... '고객 유치'보단 '이탈 고객 방어'에 온힘

이통사 고객, 알뜰폰으로 이탈 서비스·혜택 강화해 해지 방지 'OTT+통신' 결합 요금제 선
봬 자사 고객 위한 혜택·이벤트도

2024. 9. 5.



신규 고객 유치 + 계약 해지(=이탈)를 피하는 것이 중요

⑤ 세계일보

2024년에만 13만8000명 이탈... SK텔레콤 떠난 고객들, 왜?

올 들어 SK텔레콤 휴대폰 가입자 이탈이 눈에 띄게 늘고 있다. 소비자들이 가성비가 좋은 알뜰폰을 찾는 흐름이 계속되는 가운데 SK텔레콤 서비스에...

2024. 9. 30.



여러 통신 사업자가 서비스를 제공함에 따라

⑥ 에너지경제신문

통신 3사 가입자 300만명 감소... 멤버십 혜택 강화하는 이유

SK텔레콤, KT, LG유플러스 등 국내 이동통신 3사가 올해 들어 멤버십 혜택을 강화하는 전략을 적극 추진하고 있다. 휴대폰 가입자 수 감소 추세 속...

2주 전



 **통신사의 수익성에 큰 영향을 미침**

**이탈 고객을 예측하고, 그들의 특징을 분석하여
그에 맞는 대응 전략을 제시하고자 함**

데이터 소개

02. 데이터 소개 (7043 rows x 21 columns)



고객 정보 관련 변수

1. customerID: 고객 아이디
2. gender: 성별
3. SeniorCitizen: 노인 여부
4. Partner: 배우자 여부
5. Dependents: 부양가족 여부



통신사 서비스 관련 변수

1. tenure: 회사와 함께한 개월 수
2. PhoneService: 전화 서비스 사용 여부
3. MultipleLines: 다중전화선 사용 여부
4. InternetService: 인터넷 서비스 제공업체
5. OnlineSecurity: 온라인 보안 사용 여부
6. OnlineBackup: 온라인 백업 사용 여부
7. DeviceProtection: 장치 보호 사용 여부
8. TechSupport: 기술 지원 사용 여부
9. StreamingTV: TV 스트리밍 사용 여부
10. StreamingMovies: 영화 스트리밍 사용 여부
11. Contract: 계약 기간
12. PaperlessBilling: 전자 청구서 수신 여부
13. PaymentMethod: 결제 방법
14. MonthlyCharges: 월 청구 금액
15. TotalCharges: 청구된 총 금액
16. Churn: 이탈 여부

결측치 확인 및 처리

03. 결측치 확인 및 처리

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   customerID  7043 non-null   object 
 1   gender       7043 non-null   object 
 2   SeniorCitizen 7043 non-null   int64  
 3   Partner      7043 non-null   object 
 4   Dependents   7043 non-null   object 
 5   tenure       7043 non-null   int64  
 6   PhoneService 7043 non-null   object 
 7   MultipleLines 7043 non-null   object 
 8   InternetService 7043 non-null   object 
 9   OnlineSecurity 7043 non-null   object 
 10  OnlineBackup  7043 non-null   object 
 11  DeviceProtection 7043 non-null   object 
 12  TechSupport   7043 non-null   object 
 13  StreamingTV   7043 non-null   object 
 14  StreamingMovies 7043 non-null   object 
 15  Contract      7043 non-null   object 
 16  PaperlessBilling 7043 non-null   object 
 17  PaymentMethod 7043 non-null   object 
 18  MonthlyCharges 7043 non-null   float64
 19  TotalCharges  7043 non-null   object 
 20  Churn        7043 non-null   object 
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

```
customerID 0
gender 0
SeniorCitizen 0
Partner 0
Dependents 0
tenure 0
PhoneService 0
MultipleLines 0
InternetService 0
OnlineSecurity 0
OnlineBackup 0
DeviceProtection 0
TechSupport 0
StreamingTV 0
StreamingMovies 0
Contract 0
PaperlessBilling 0
PaymentMethod 0
MonthlyCharges 0
TotalCharges 11
Churn 0
dtype: int64
```

TotalCharges
object >> float64

TotalCharges 컬럼에 총 11개의 결측값 존재

03. 결측치 확인 및 처리

	tenure	MonthlyCharges	TotalCharges
488	0	52.55	NaN
753	0	20.25	NaN
936	0	80.85	NaN
1082	0	25.75	NaN
1340	0	56.05	NaN
3331	0	19.85	NaN
3826	0	25.35	NaN
4380	0	20.00	NaN
5218	0	19.70	NaN
6670	0	73.35	NaN
6754	0	61.90	NaN

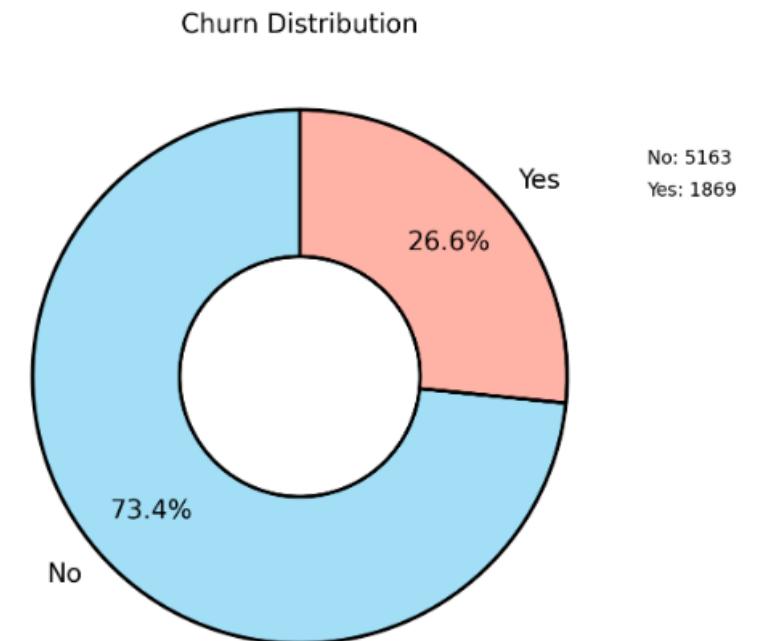
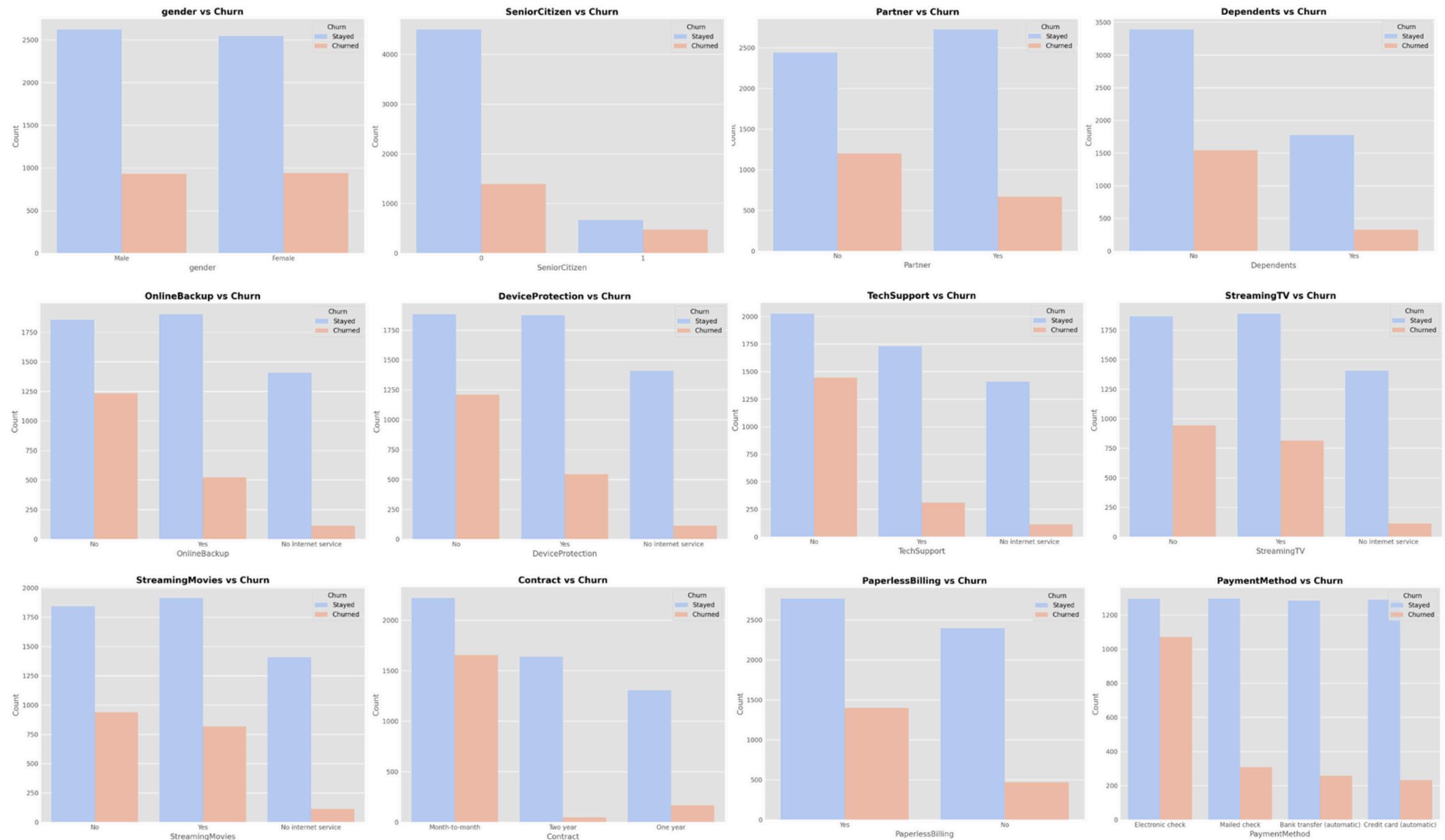
MonthlyCharges(월 청구 요금) 데이터가 있지만
tenure(가입 기간) 값이 0인 것을 확인할 수 있음

해당 데이터를 수집할 당시 고객이 가입한 지 한 달이 채 지나지 않아
TotalCharges가 NaN 값일 거라고 추정

이탈한 고객을 예측하는 게 중요한 만큼,
이 행들은 아직 고객 특성을 충분히 반영하지 못해 삭제 처리

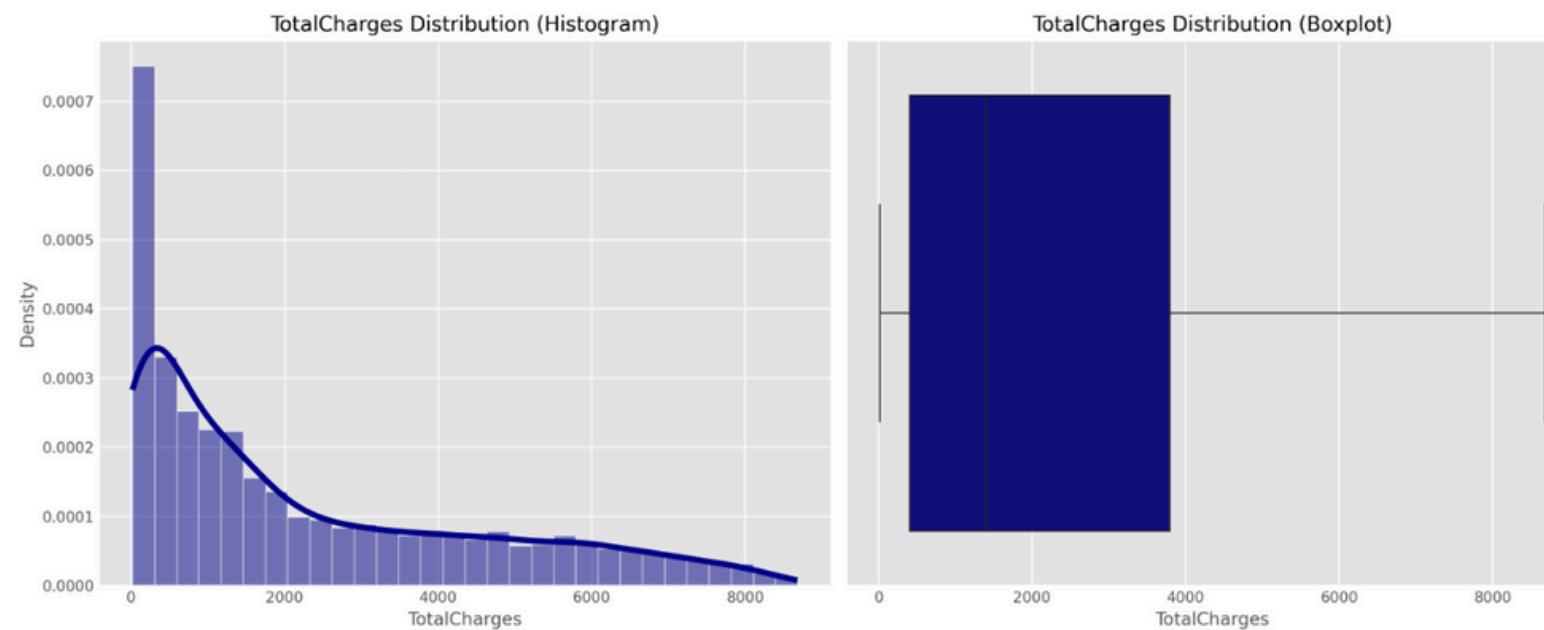
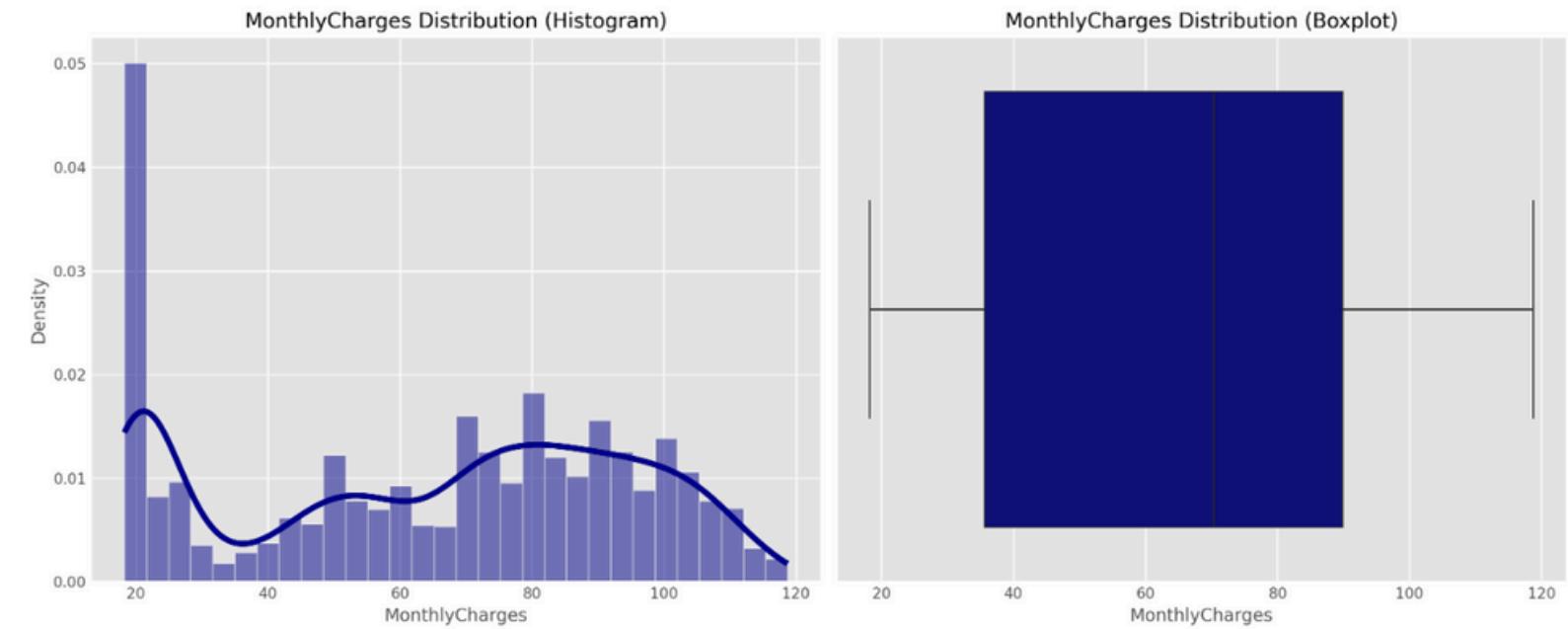
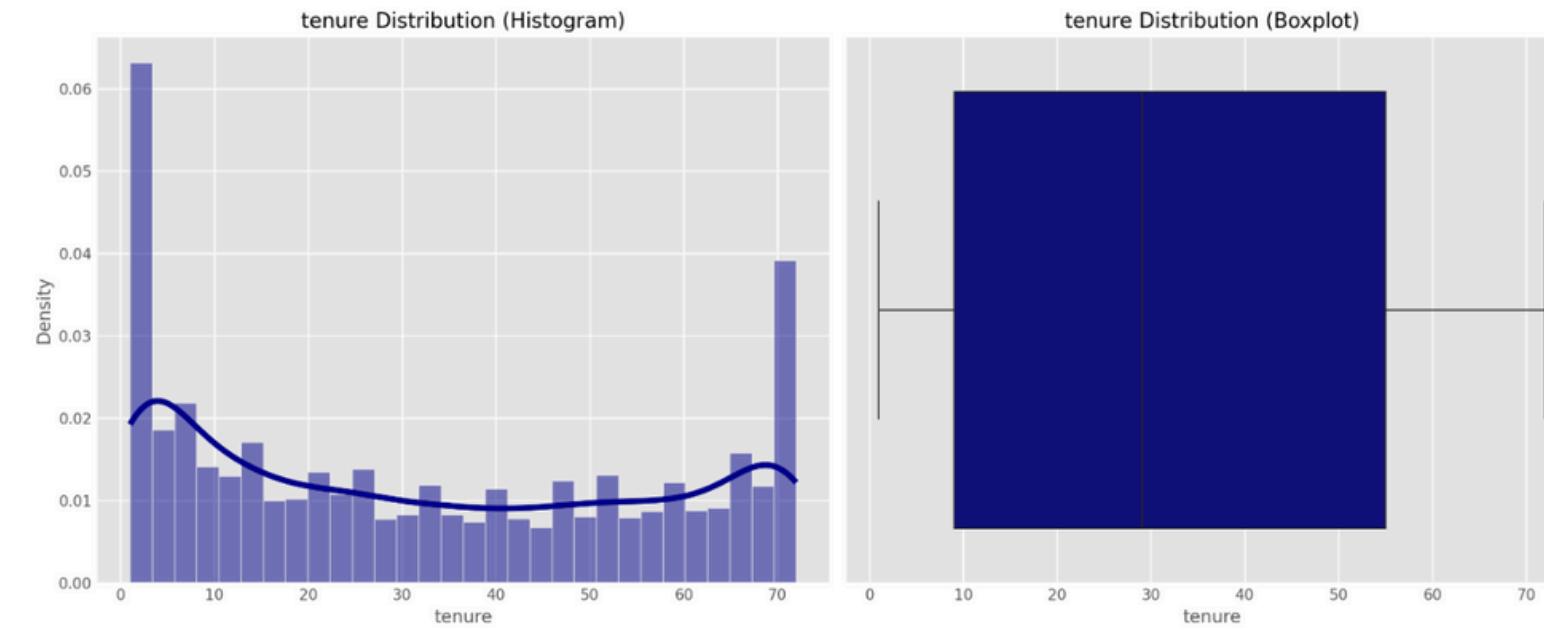
EDA

04. EDA - 범주형 변수



변수별 비율 확인 및
Churn과의 관계 시각화

04. EDA - 연속형 변수



변수별 분포 및 이상치 확인

04. EDA - 파생변수 생성

tenure	MonthlyCharges	TotalCharges
1	29.85	29.85
34	56.95	1889.50
2	53.85	108.15
45	42.30	1840.75
2	70.70	151.65
...
24	84.80	1990.50
72	103.20	7362.90
11	29.60	346.45
4	74.40	306.60
66	105.65	6844.50

TotalCharges의 값이 Tenure x MonthlyCharges의 값과 동일하지 않았음
→ 중간에 요금제를 변경하거나 할인을 받아서 값이 달라졌을 가능성 제시

예상 총 요금

```
df['ExpectedTotalCharges'] = df['tenure'] * df['MonthlyCharges']
```

총 요금 오차(실제-예상)

```
df['TotalChargesDiff'] = df['TotalCharges'] - df['ExpectedTotalCharges']
```

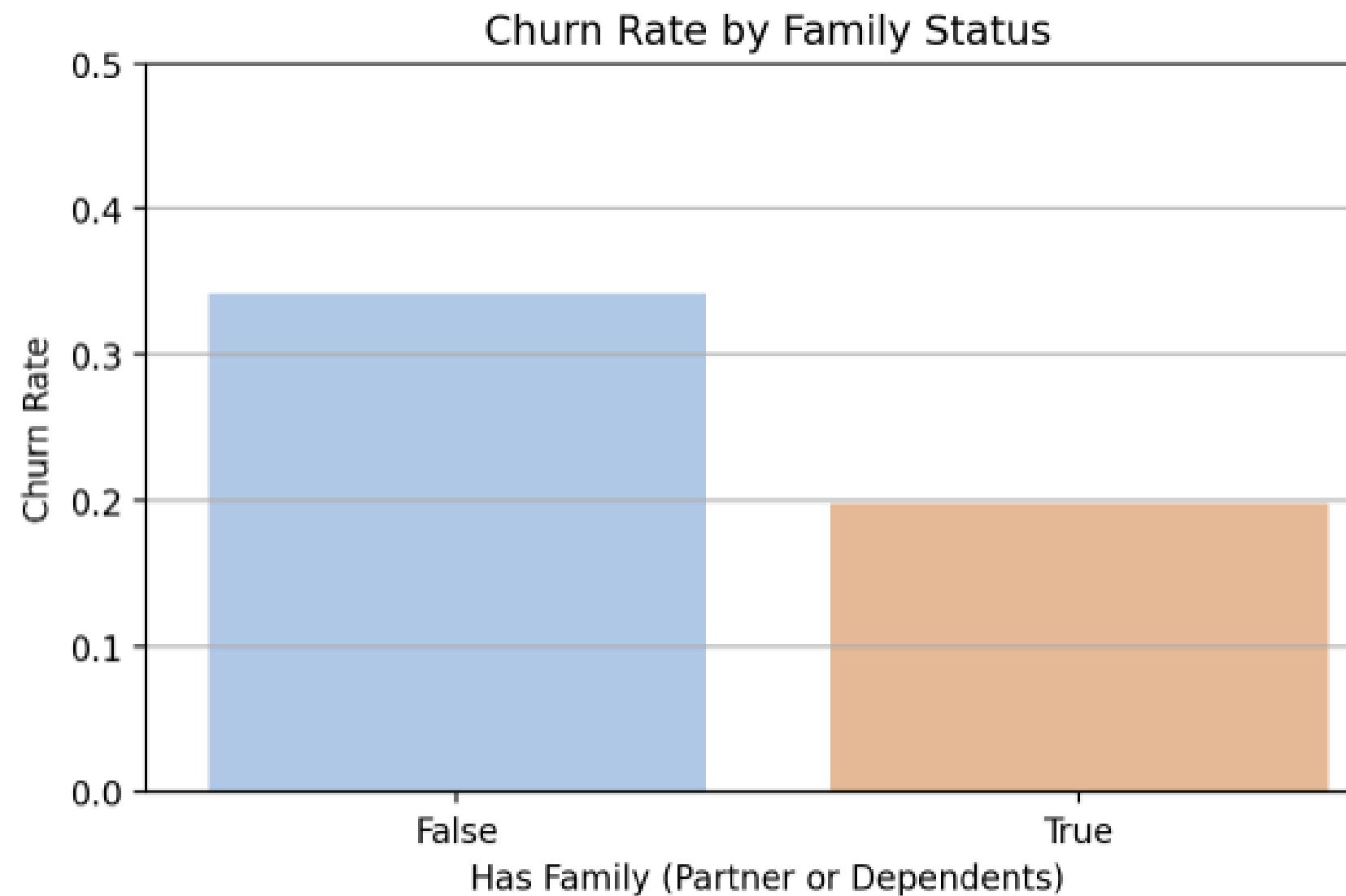
오차 비율

```
df['TotalChargesRatioDiff'] = df['TotalChargesDiff'] / (df['ExpectedTotalCharges'] + 1e-5)
```

가설 설정 및 검정

01. 가설 내용 - 가족 관련 변수와 이탈율

```
df['family'] = ((df['Partner'] == 'Yes') | (df['Dependents'] == 'Yes')).astype(bool)
```



가설: 가족이 있으면 이탈율이 낮아진다

- 이탈 과정이 번거로워 이탈율이 낮아질 것이다.
- 가족끼리 통신사를 둑으면 혜택이 많아 이탈율이 낮을 것이다.

01. 가설 내용 - 가족 관련 변수와 이탈율

```
pd.crosstab(df['family'], df['Churn'], normalize='index')
```

Churn	No	Yes
family		
False	0.657622	0.342378
True	0.801754	0.198246

두 범주형 변수 간의 독립성을 검정

- family: 배우자 또는 부양 가족 유무 (Yes/No)
- Churn: 이탈 여부 (Yes/No)

카이제곱 검정 결과

카이제곱 통계량: 184.07069542119154

p-value : 6.26146434898279e-42

가족이 있는 사람들 중에서 Churn의 여부를 살펴봤을 때 True(이탈)의 비율이 훨씬 낮다.

카이제곱 검정 결과, p-value가 유의수준 0.05보다 현저히 작아, 두 변수 간 유의미한 연관성이 존재함을 통계적으로도 확인함

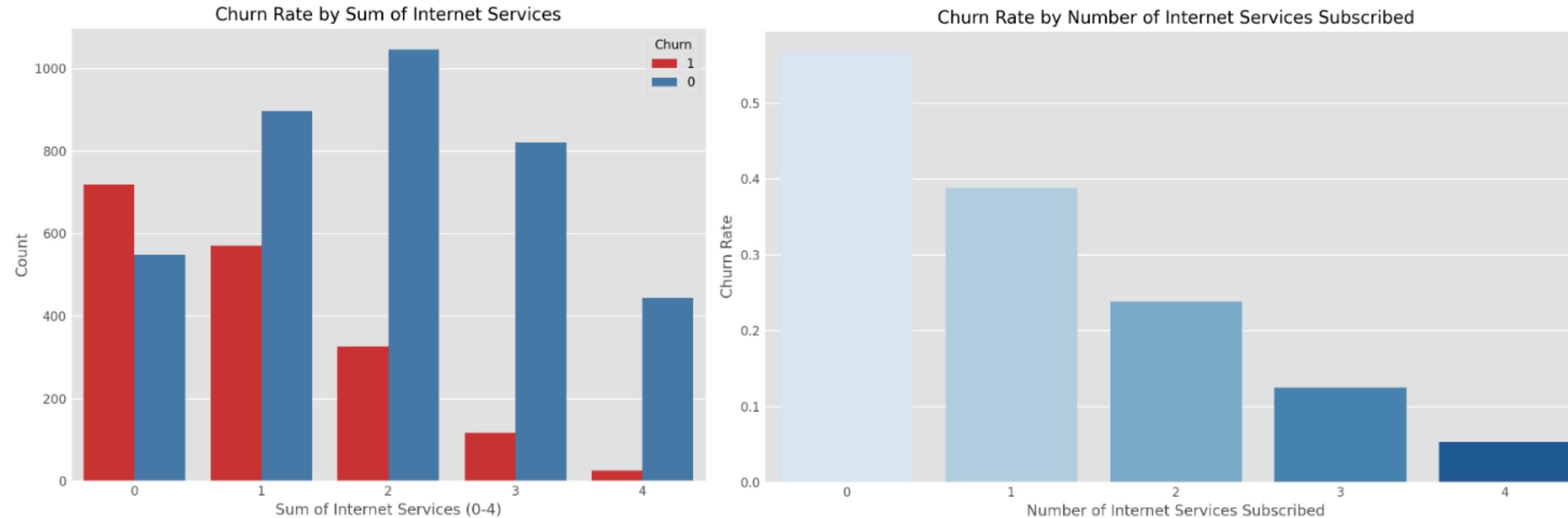
02. 가설 내용 - 인터넷 보안 서비스 관련 변수와 이탈율

```
internet_columns = ['OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport']
df['Sum_of_InternetService'] = df[internet_columns].apply(lambda row: (row == 'Yes').sum(), axis=1)
```

	인터넷 보안 서비스 관련 변수				파생 변수	
	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	Sum_of_InternetService	Churn
0	No	Yes	No	No	1	No
1	Yes	No	Yes	No	2	No
2	Yes	Yes	No	No	2	Yes
3	Yes	No	Yes	Yes	3	No
4	No	No	No	No	0	Yes

가설: 보안 서비스에 더 많이 가입할수록
이탈률이 낮아질 것이다.

02. 가설 내용 - 인터넷 보안 서비스 관련 변수와 이탈율



ANOVA Test Results:

F-statistic: 216.10

P-value: 1.67e-172

Sum_of_InternetService

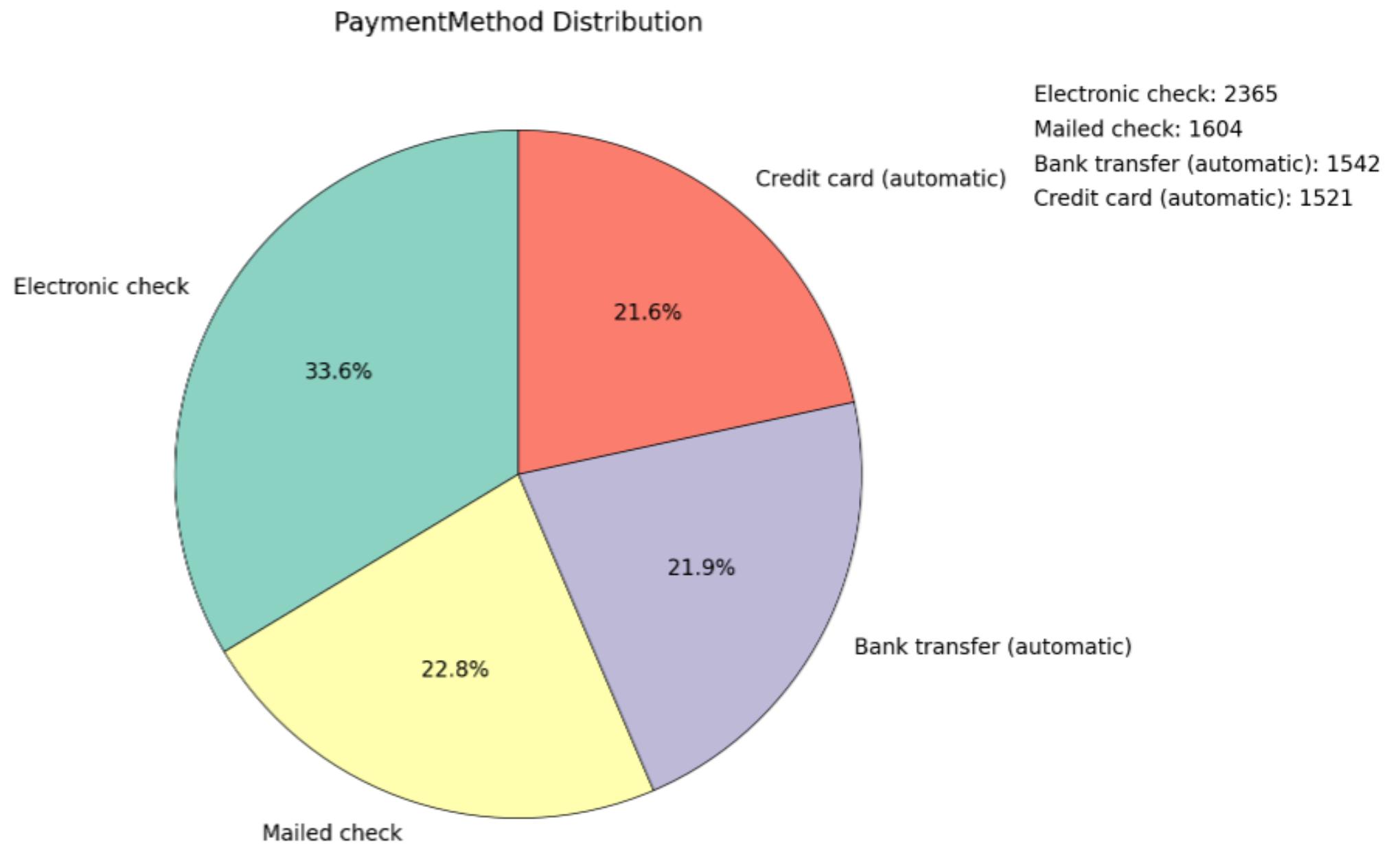
0	0.566693
1	0.388548
2	0.237609
3	0.124867
4	0.053305

Name: Churn, dtype: float64

시각화 결과 보안 관련 서비스를 많이 가입할수록 이탈 확률이 낮아짐을 확인할 수 있음

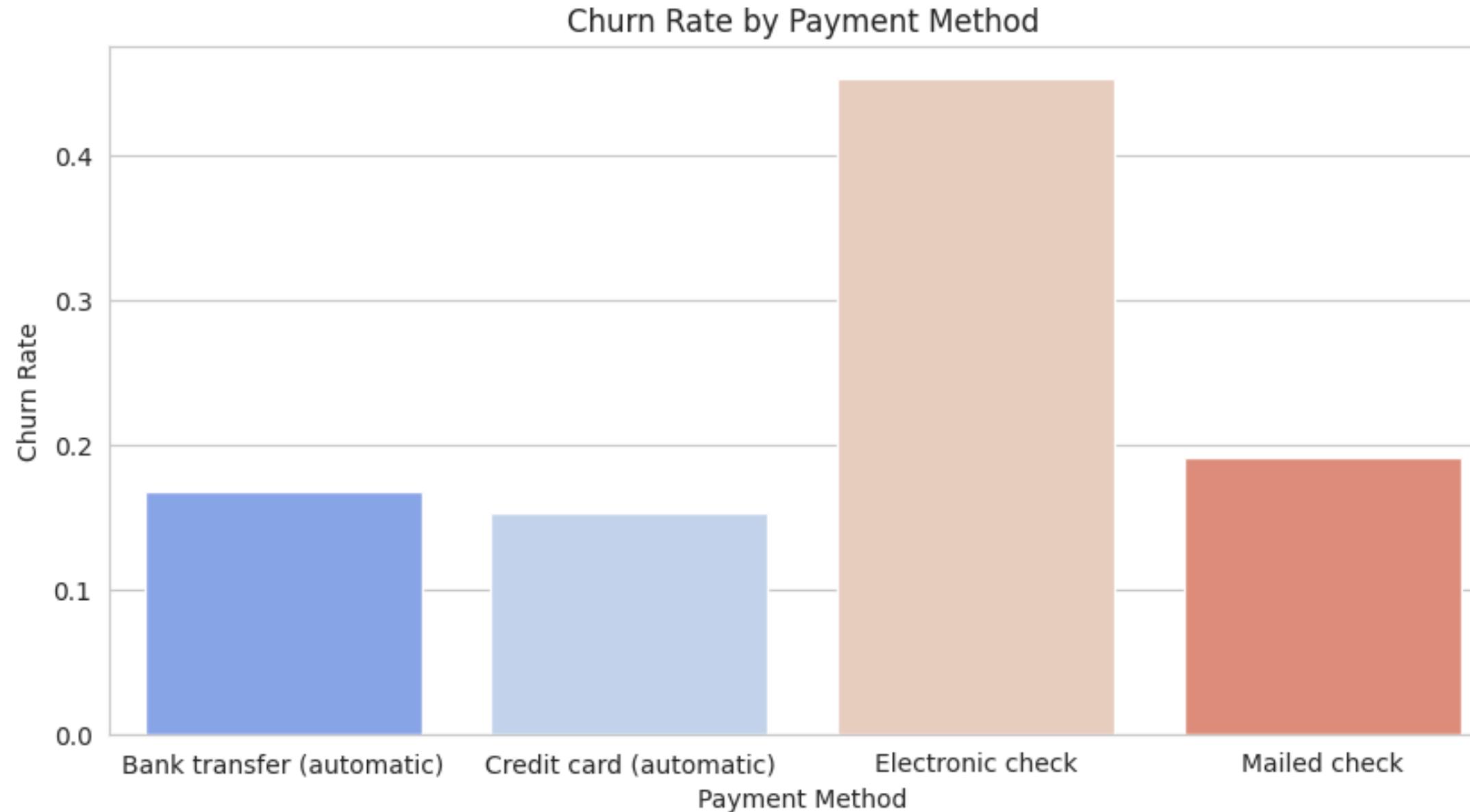
보안 관련 서비스를 가입한 사람들 중에서 Churn의 여부를 살펴봤을 때 4개 서비스를 모두 가입한 사람들의 이탈 비율이 가장 낮다.
ANOVA 분산 분석 결과, p-value가 유의수준 0.05보다 현저히 작아, 각 그룹 간 이탈률 차이가 통계적으로도 유의미함을 확인함

03. 가설 내용 - 결제 방식 변수와 이탈률



가설: 결제 방식과 이탈률 간에는 통계적으로 유의미한 관계가 있다

03. 가설 내용 - 결제 방식 변수와 이탈률



카이제곱 검정 결과:

Chi-Square Statistic: 645.4299001234638
p-value: 1.4263098511063342e-139

전자 결제 방식의 이탈률이 가장 높고, 자동화 방식의 두 가지 결제방식이 이탈률이 낮음을 볼 수 있다.

카이제곱 검정 결과, p-value가 유의수준 0.05보다 현저히 작아, 결제 방식과 이탈률 유의미한 연관성이 존재함을 통계적으로도 확인함

04. 가설 내용 - 계약 기간 변수와 이탈율

```
df["is_monthly_customer"] = (df["Contract"] == 0).astype(int)
pd.crosstab(df['is_monthly_customer'], df['Churn'], normalize='index')
```

Churn 0 1

가설: 한달 단위로 계약하는 소비자들은 이탈율이 높을 것이다.

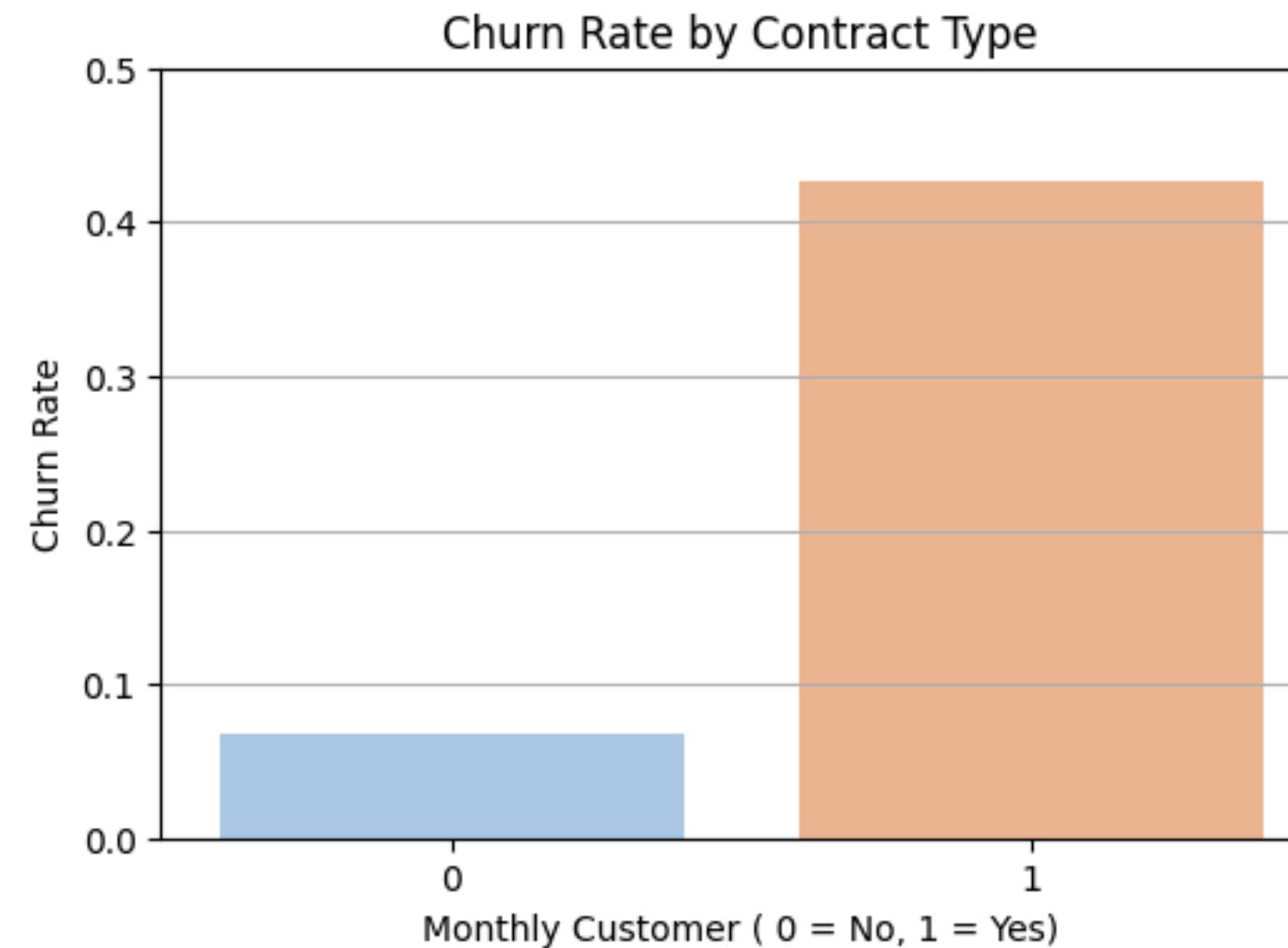
		is_monthly_customer	
		0	1
Churn	0	0.932214	0.067786
	1	0.572903	0.427097

- Churn (0 = No, 1 = Yes)
- Monthly Customer (0 = No, 1 = Yes)

- 월별 단위 계약 고객은 장기 계약(1년/2년) 고객보다 해지 부담이 적음
- 연간 계약 고객은 위약금이나 할인 혜택으로 인해 쉽게 해지하지 않지만, 월 단위 고객은 언제든지 손쉽게 해지 가능
- 서비스가 필요 없을 때 바로 해지할 가능성이 높음

04. 가설 내용 - 계약 기간 변수와 이탈율

```
contingency_table = pd.crosstab(df["is_monthly_customer"], df["Churn"])
chi2_stat, p_value, dof, expected = stats.chi2_contingency(contingency_table)
```



카이제곱 검정 결과

Chi-square Statistic: 1149.1043152026427

P-value: 7.023467851331667e-252

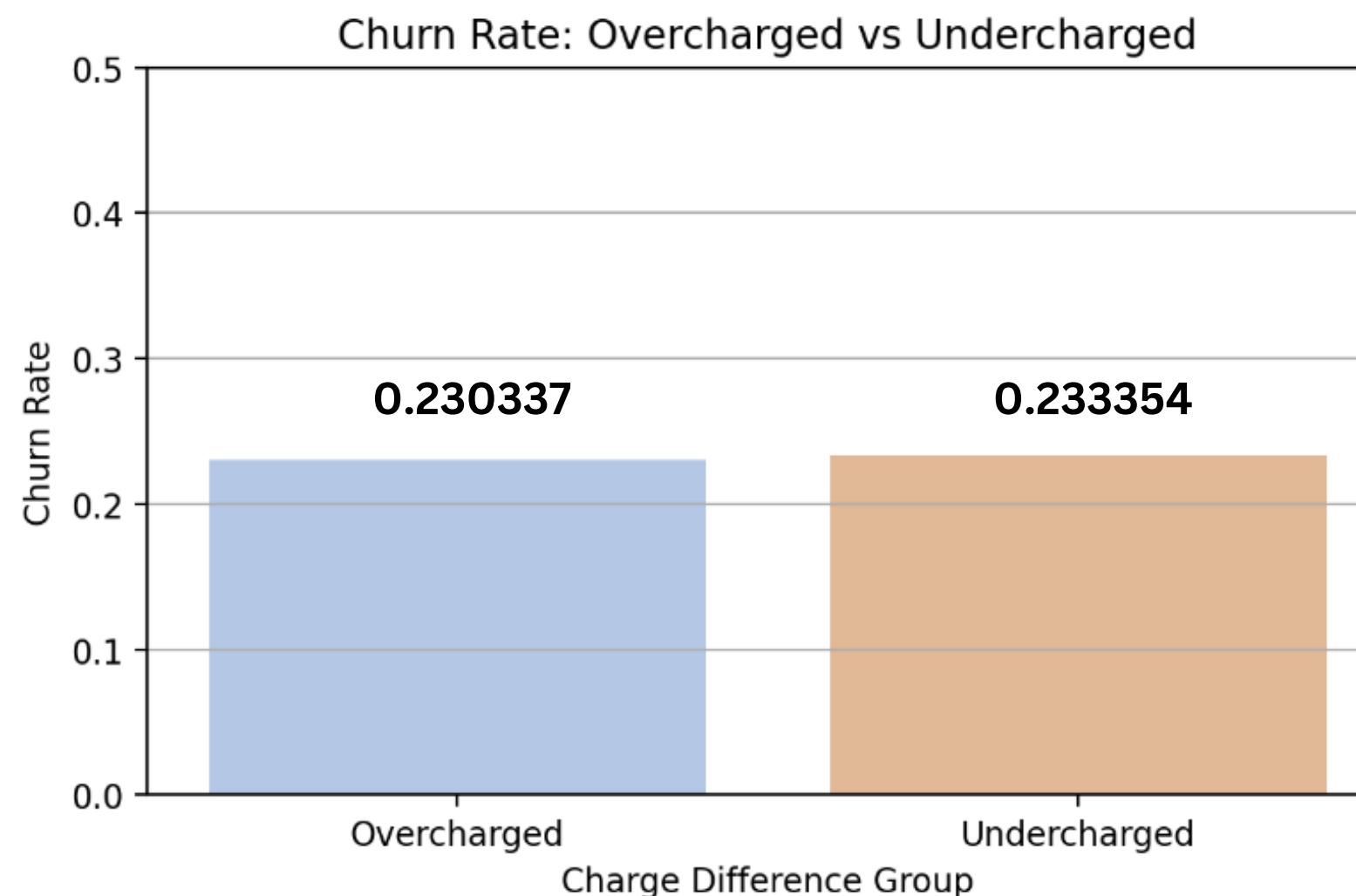
Churn Rate (Monthly Customers): 0.4271

Churn Rate (Non-Monthly Customers): 0.0676

매달 결제하는 사람들의 Churn의 여부를 살펴봤을 때 이탈율이 훨씬 높다.

카이제곱 검정 결과, p-value가 유의수준 0.05보다 현저히 작아, 두 변수 간 유의미한 연관성이 존재함을 통계적으로도 확인함

05. 가설 내용 - 이상 가격과 이탈율



가설: 예상 값보다 많은 요금을 지불하는 고객일수록
이탈 가능성이 높아질 것이다.

- $\text{TotalChargesRatioDiff} > 0$, 이탈율 증가
- $\text{TotalChargesRatioDiff} < 0$. 이탈율 감소

로지스틱 회귀 분석 결과 (수치형 → 범주형)

	coef	std err	z	P> z	[0.025	0.975]
const	-1.0161	0.027	-37.638	0.000	-1.069	-0.963
TotalChargesRatioDiff	-0.5332	0.531	-1.004	0.315	-1.574	0.507

점진적인 추세나 영향을 발견할 수 있지 않을까해서 로지스틱 회귀분석 진행
계수만 확인했을 때는 오차비율이 증가할수록 이탈확률이 감소함을 시사하지만, p-value > 0.05로, 귀무가설이 채택

모델링

06. 모델링 - 피처 엔지니어링

선택한 변수

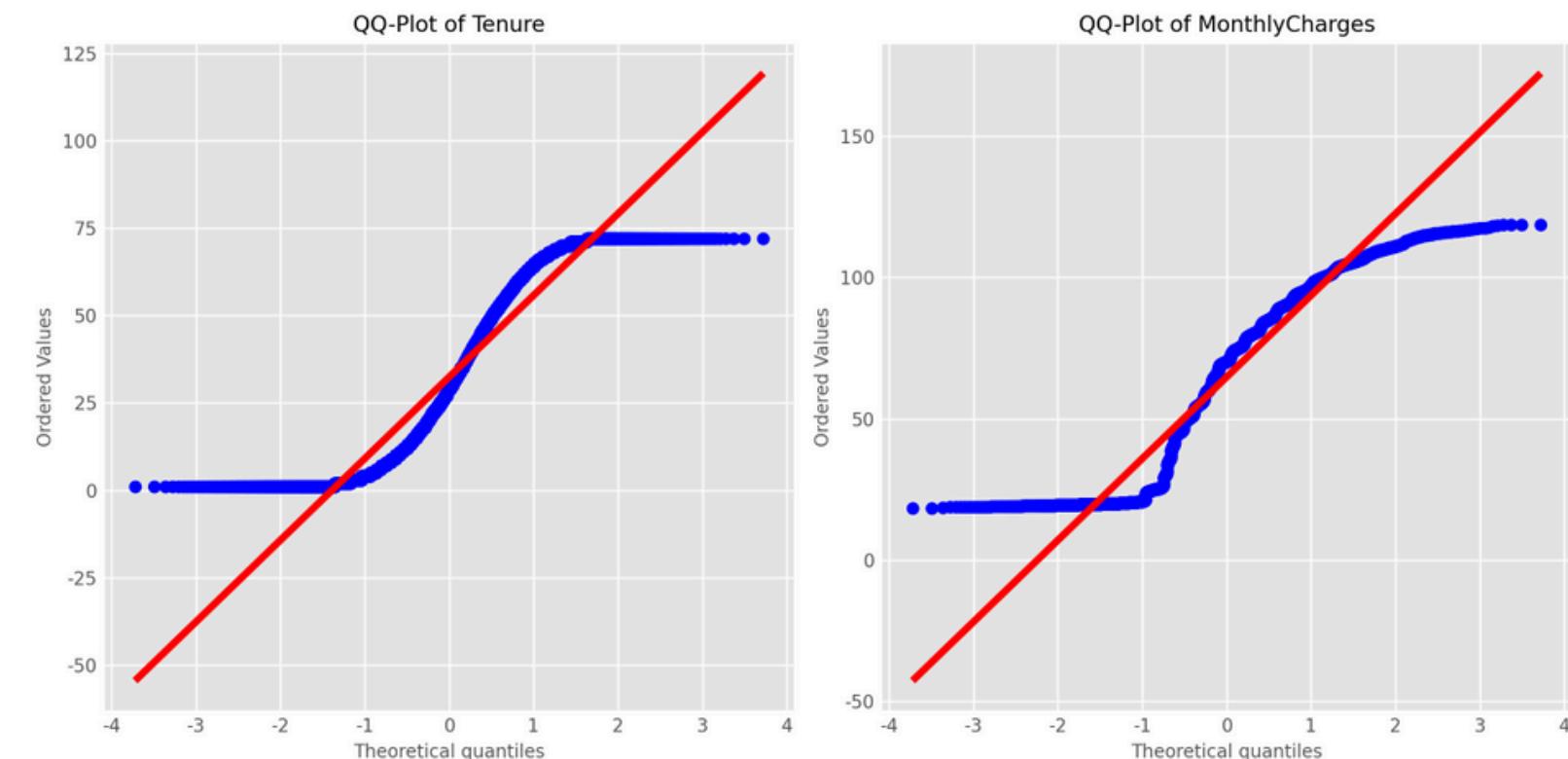
```
Index(['SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService',  
       'MultipleLines', 'InternetService', 'StreamingTV', 'StreamingMovies',  
       'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges',  
       'Churn', 'Sum_of_InternetService'],  
      dtype='object')
```

```
['Partner', 'Dependents', 'PhoneService', 'MultipleLines',  
 'StreamingTV', 'StreamingMovies', 'PaperlessBilling']
```

→ 이진 변수로 변환

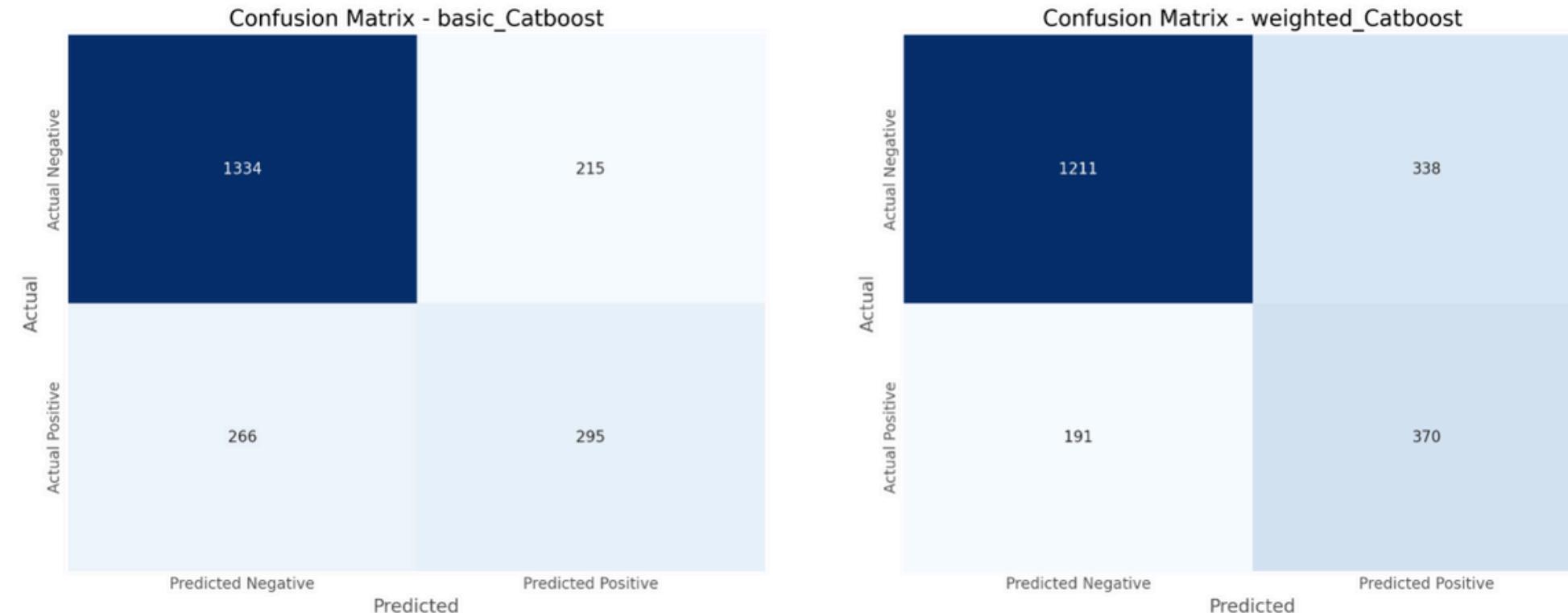
```
['InternetService', 'Contract', 'PaymentMethod']
```

→ 더미화 진행 (One-hot Encoding)



연속형 변수인 'tenure'과 'MonthlyCharges'는 분포 확인 결과 정규 분포를 따르지 않아 **Min-Max Scailing**으로 처리

06. 모델링 - 모델 비교



- TN(Churn No): 1334 → 1211
- TP(Churn Yes): 295 → 370
- FN(이탈했으나 이탈하지 않았다고 분류): 266 → 191
- FP(이탈하지 않았으나 이탈했다고 분류): 215 → 338

기본 모델이 전체적인 정확도는 높지만
가중치를 적용한 모델이 이탈 고객을 더 잘 찾아냄

	Accuracy	Recall	Roc_Auc	Precision
basic_Catboost	0.772038	0.525847	0.693524	0.578431
weighted_Catboost	0.749289	0.659537	0.720666	0.522599

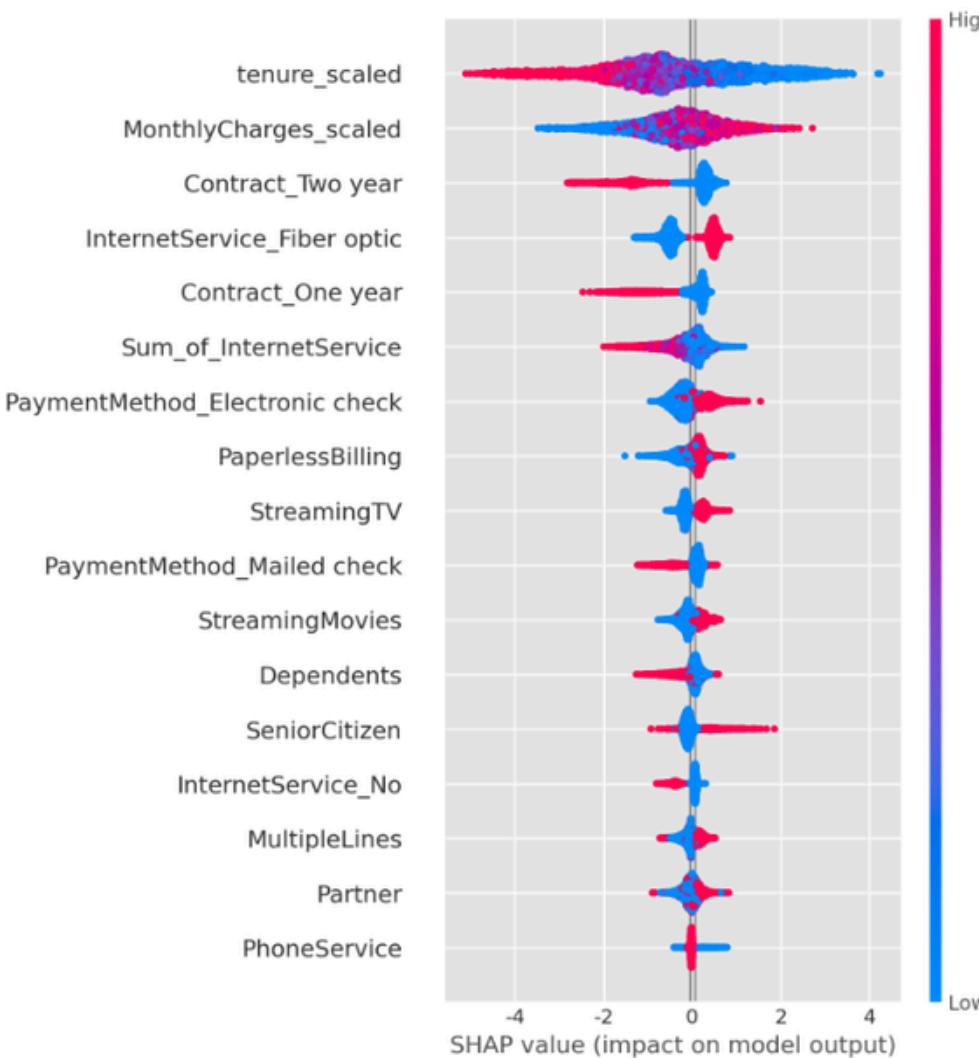
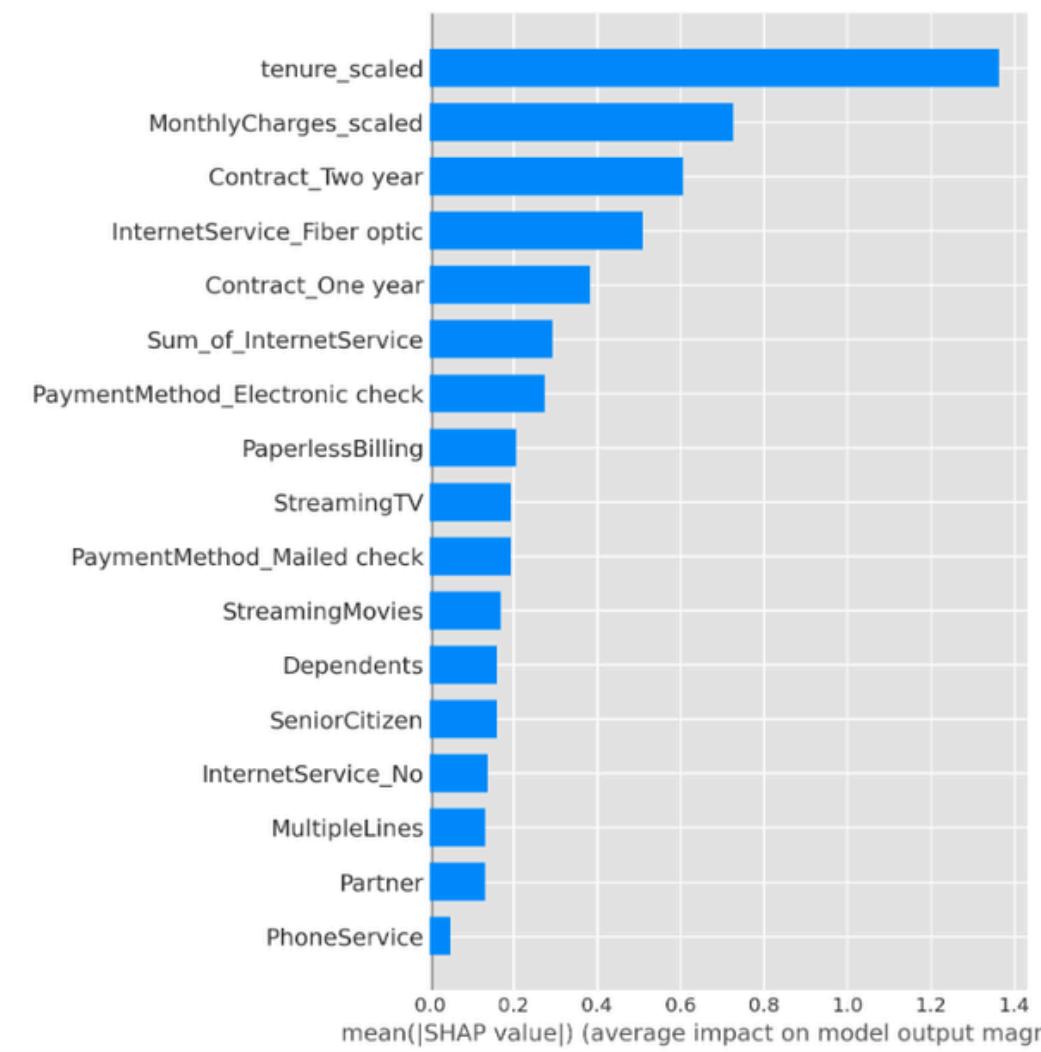
이탈 고객(소수 클래스)을 잘 예측해 이에 맞는 대응으로
비즈니스 손실을 줄이고자 함이 목표

→ 가중치를 적용한 모델이 더 효과적

06. 모델링 - 중요 피쳐 시각화

SHAP(SHapley Additive exPlanations):

- 모델의 예측 결과를 설명할 수 있게 해주는 모델 해석 기법
- 각 특성이 모델 예측에 얼마나 기여했는지를 분배하는 방식



x축: SHAP value

: 각 특성이 모델 예측에 미치는 영향을 나타냄

- 값이 클수록 예측에 더 강한 영향을 미침
- SHAP value이 **양수**면 **이탈 확률이 높이는** 영향을 의미하고 반대로 **음수**면 **이탈 확률을 낮추는** 영향을 의미함

ex.

tenure의 경우 고객이 장기 계약을 유지할수록 이탈 확률이 낮아짐

y축: Feature value

: 특성의 실제 값

- 파란색은 낮은 값, 빨간색은 높은 값을 의미

결론

07. 결론

이탈율이 낮은 고객들의 특징

장기이용

장기결제

인터넷 서비스
가입 개수

이탈 방지 전략

(1) 장기계약 유도

- 장기 계약 프로모션 제공 (예: 1년/2년 계약 시 할인)

(2) 다양한 서비스 세트 상품

- 여러 인터넷 서비스 결합 상품을 통한 고객 락인 유도
- 가격보다 서비스를 중시하는 고객들을 타겟으로 캠페인 전개
 - 타경쟁사의 일시적인 프로모션에 쉽게 해지하지 않는 소비자
→ 장기적으로 긍정적인 효과를 얻을 수 있음

이탈율이 높은 고객들의 특징

월별 요금

Fiber 인터넷
사용

전자 결제

전자 청구서

고객 이탈 선제 대응 전략

(1) 서비스 개선

- 이탈하는 고객들의 만족도 조사 결과를 기반으로 서비스 개선 및 프로모션 진행
- 전자 청구서의 UX/UI에 프로모션이나 혜택 정보를 함께 추가하여 해당 요금이 합리적임을 함께 보여줌

(2) 가격에 민감한 고객에 프로모션 제공

- 가격 부담으로 이탈하는 고객에게 저렴한 요금제 추천 및 요금제 컨설팅 진행
- 해지 시점 개인별 가격 측면 프로모션을 제공하여 지속적 사용을 유도

Thank You