

통신사 고객 이탈 예측

기계학습 Team Project

2019204045 윤서환
2019204094 이건희
2019204030 이승훈
2019204083 김효준
2019204023 윤성호

Contents

01. 주제 선정 배경 및 중요성

02. 데이터 설명

03. 데이터 EDA

04. 데이터 전처리, Scaling

05. PCA Modeling

06. Modeling

07. Optimization

08. XAI - SHAP

주제 선정 배경 및 중요성

**씨티 "전기·가스요금 인상에도 5월 물가상승률 3~3.5% 전망"
4인 가구 5G 요금 30만원… “통신비만 잡아도 물가 낮춘다”**

4인 가구 기준, 5G 110GB 요금제 살펴보면

SKT·KT, 27만6000원…LGU+, 30만원

4인 가구 월 전기료 5만원보다 6배 비싸

요금제 구성도 10GB·110GB 두개뿐

“기본 서비스 선택과는 별개로 통신비만 줄여도 물가 낮춘다”

이통 이용자 절반, 가입 통신사에 불만…5G 만족도 LTE보다 낮아

송고시간 | 2022-09-12 07:00



조승한 기자
[기자페이지](#)

| 알뜰폰 이용자 만족도 63%…이통사 이용자보다 높아

정부 “3월 물가 더 오를수도…알뜰폰 요금 추가 인하”

주제 선정 배경 및 중요성

출처: 한국 신용 평가

3) MVNO¹(이하 '알뜰폰') 점유율 상승이 통신사에 미치는 영향과 모니터링 요소는?

당분간 알뜰폰 점유율은 상승흐름을 보일 전망이나, 중간요금제 출시 등을 통한 통신사의 중·저가 요금 수요층 흡수, 알뜰폰 사업자의 제한된 경쟁력 등을 감안할 때, 알뜰폰 점유율 상승에도 통신 3사의 시장지위 및 이익창출력이 크게 훼손되지는 않을 것으로 본다.

다만, 정책지원 등으로 MVNO 사업자의 5G 요금경쟁력이 제고될 경우, 알뜰폰 시장 잠식에 따른 영향은 다소 커질 수 있다.

22년 고객용 휴대폰 시장 내
알뜰폰 점유율 2.1%p 상승

알뜰폰 가입자 수는 2021년 말 최초로 1천만명에 도달하였고, 2022년 말에는 약 1.3천만명까지 증가하였다. 이에 2019년까지 10% 내외에 머물던 점유율도 2022년 말 16.9%까지 상승하였다.

다만, 2019년 이후 알뜰폰 가입자 증가는 단말장치(태블릿PC, 웨어러블 기기 등)와 사물지능통신(차량관제, 원격관제 등) 회선증가에 기인하고 있으며, ARPU가 현저히 높은 고객용 휴대폰 시장 점유율은 통신 3사의 저가 요금제 출시 등으로 2021년까지 오히려 소폭 저하된 모습이었다.

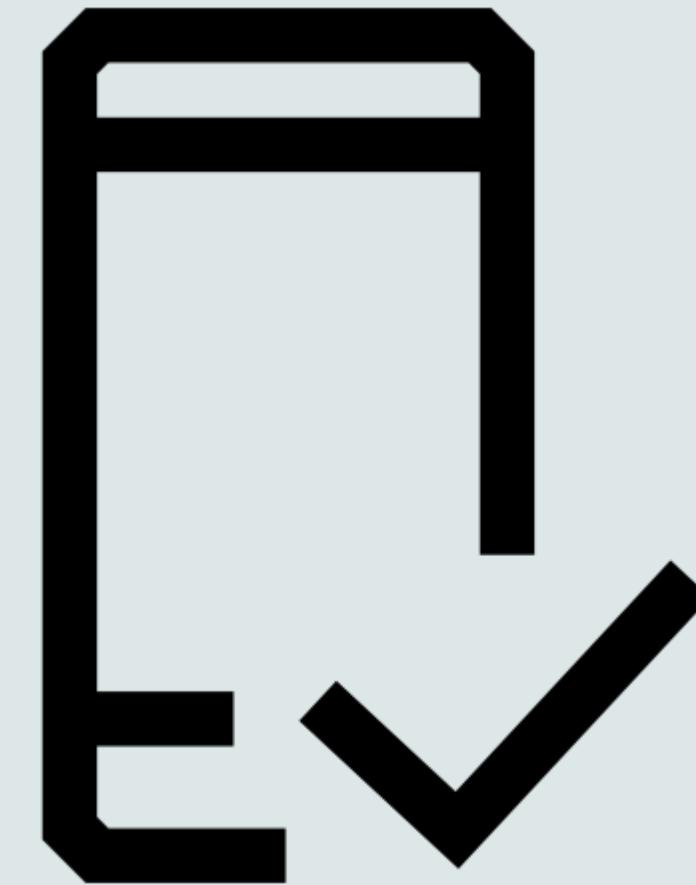
2022년 들어서는 고객용 휴대폰 시장에서도 알뜰폰 점유율이 상승 전환하였다(고객용 휴대폰 기준 MVNO 가입자 수 2021년 말 6.1백만명 → 2022년 말 7.3백만명). 이는 5G 출시로 통신 3사 주력 요금제가 비싸진 가운데, KB국민은행 등 자금력과 서비스 역량을 갖춘 사업자들이 시장에 진입하면서 가입자 유입에 영향을 준 것으로 보인다.

주제 선정 배경 및 중요성

이탈 고객 행동 및 동인 이해와 식별 필요

이탈 고객 사전 예방 필요

통신사 고객 유지 및 서비스 향상에 기여



데이터 설명

<https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113>

IBM의 캘리포니아 거주 고객 약 7043명의 전화 및 인터넷 서비스 데이터



IBM Think 제품 솔루션 컨설팅 지원 더 둘러보기

가치 있는 7가지 비즈니스 전략

IBM 전략 보고서를 통해 차세대 비즈니스를 이끌 인사이트를 확보하세요.

7가지 전략 보고서 보기 → Think 보기 →

01 창의적인 혁신 02 컨설팅 03 X-Force 보고서 04 제품 05 Inside IBM 06 고객사례 07 IBM 소개

IBM과 함께하는 창의적인 혁신
하이브리드 클라우드와 AI로 경쟁력을 극대화하여 더 나은 세상을 만드는 법

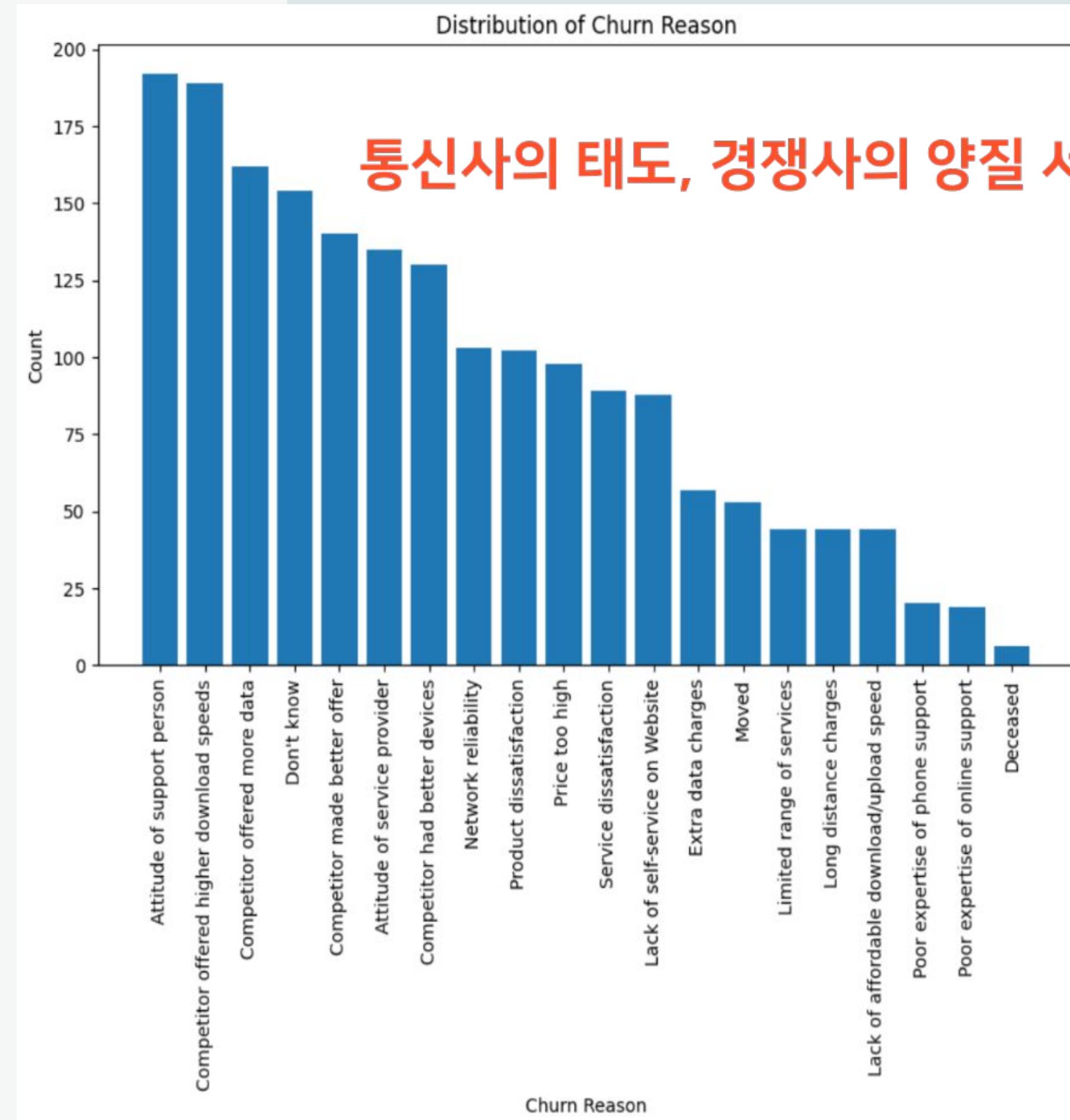
A screenshot of the IBM website's homepage. It features a search bar and navigation menu. The main headline is "가치 있는 7가지 비즈니스 전략". Below it, there's a section for "IBM 전략 보고서" with links to various reports like "창의적인 혁신", "컨설팅", "X-Force 보고서", etc. To the right, there's a graphic of interconnected icons representing business concepts like data, collaboration, and technology. At the bottom, there are three small video thumbnail images showing people working in an office environment.

데이터 설명 - 컬럼

- CustomerID: 고유 ID
- City: 도시
- Zip Code: 우편번호
- Latitude: 위도
- Longitude: 경도
- Gender: 성별
- Senior Citizen: 65세 이상인지 표시
- Partner: 커플 여부
- Dependents: 가족과 함께 살고 있는지 여부
- Tenure Months: 통신사 가입 기간
- Phone Service: 전화 서비스 가입 여부
- Multiple Lines: 여러 전화 회선에 가입했는지 여부
- Internet Service: 회사에서 인터넷 서비스에 가입했는지 여부
- Online Security: 온라인 보안 서비스에 가입했는지 여부
- Online Backup: 온라인 백업 서비스에 가입된 여부
- Device Protection: 기기 보호 요금제 가입 여부
- Tech Support: 회사의 추가 기술 지원 계획에 가입 여부
- Streaming TV: TV 서비스 여부
- Streaming Movies: 영화를 구독 서비스 여부
- Contract: 계약 기간
- Paperless Billing: 인터넷 영수증 발급 여부
- Payment Method: 결제하는 방법
- Monthly Charge: 월 통신 이용 비용
- Total Charges: 분기 말까지 계산된 고객 총 요금
- Churn Label: 이번 분기에 회사를 떠났는지 여부(Yes, No)
- Churn Value: 이번 분기에 회사를 떠났는지 여부(1: Yes, 2: No)
- Churn Score: 이탈할 가능성을 나타내는 점수
- CLTV: 고객가치
- Churn Reason: 이탈하는 사유

데이터 설명

| | |
|-------------------|-------|
| Customer ID | 0 |
| Count | 0 |
| Country | 0 |
| State | 0 |
| City | 0 |
| Zip Code | 0 |
| Lat Long | 0 |
| Latitude | 0 |
| Longitude | 0 |
| Gender | 0 |
| Senior Citizen | 0 |
| Partner | 0 |
| Dependents | 0 |
| Tenure Months | 0 |
| Phone Service | 0 |
| Multiple Lines | 0 |
| Internet Service | 0 |
| Online Security | 0 |
| Online Backup | 0 |
| Device Protection | 0 |
| Tech Support | 0 |
| Streaming TV | 0 |
| Streaming Movies | 0 |
| Contract | 0 |
| Paperless Billing | 0 |
| Payment Method | 0 |
| Monthly Charges | 0 |
| Total Charges | 0 |
| Churn Label | 0 |
| Churn Value | 0 |
| Churn Score | 0 |
| CLTV | 0 |
| Churn Reason | 5174 |
| dtype: | int64 |



데이터 설명 - 컬럼

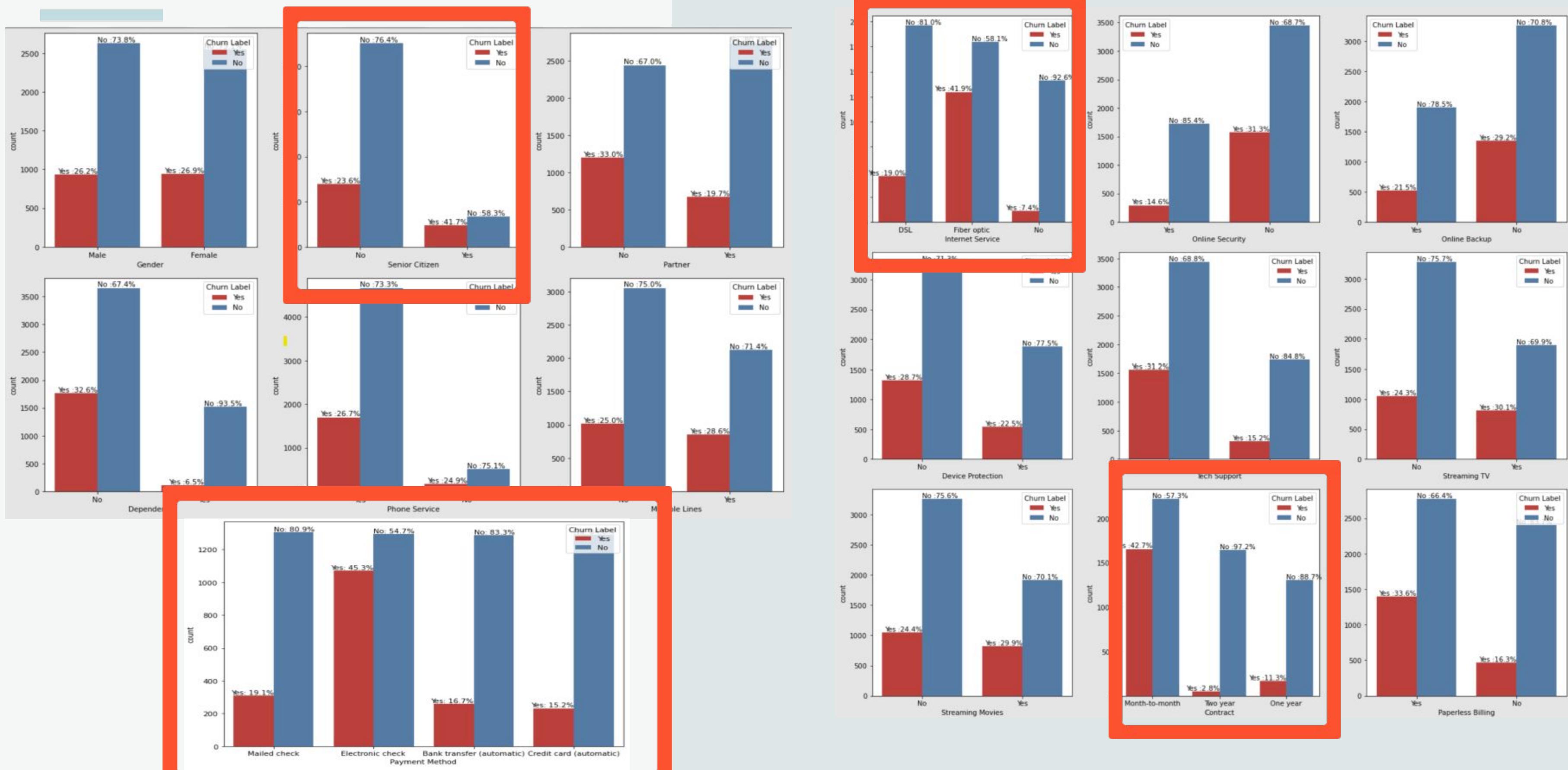
전체 데이터 type과 cardinality 확인

| | Count | Zip Code | Latitude | Longitude | Tenure Months | Monthly Charges | Churn Value | Churn Score | CLTV |
|-------|--------|--------------|-------------|-------------|---------------|-----------------|-------------|-------------|-------------|
| count | 7043.0 | 7043.000000 | 7043.000000 | 7043.000000 | 7043.000000 | 7043.000000 | 7043.000000 | 7043.000000 | 7043.000000 |
| mean | 1.0 | 93521.964646 | 36.282441 | -119.798880 | 32.371149 | 64.761692 | 0.265370 | 58.699418 | 4400.295755 |
| std | 0.0 | 1865.794555 | 2.455723 | 2.157889 | 24.559481 | 30.090047 | 0.441561 | 21.525131 | 1183.057152 |
| min | 1.0 | 90001.000000 | 32.555828 | -124.301372 | 0.000000 | 18.250000 | 0.000000 | 5.000000 | 2003.000000 |
| 25% | 1.0 | 92102.000000 | 34.030915 | -121.815412 | 9.000000 | 35.500000 | 0.000000 | 40.000000 | 3469.000000 |
| 50% | 1.0 | 93552.000000 | 36.391777 | -119.730885 | 29.000000 | 70.350000 | 0.000000 | 61.000000 | 4527.000000 |
| 75% | 1.0 | 95351.000000 | 38.224869 | -118.043237 | 55.000000 | 89.850000 | 1.000000 | 75.000000 | 5380.500000 |
| max | 1.0 | 96161.000000 | 41.962127 | -114.192901 | 72.000000 | 118.750000 | 1.000000 | 100.000000 | 6500.000000 |

수치형 데이터 describe

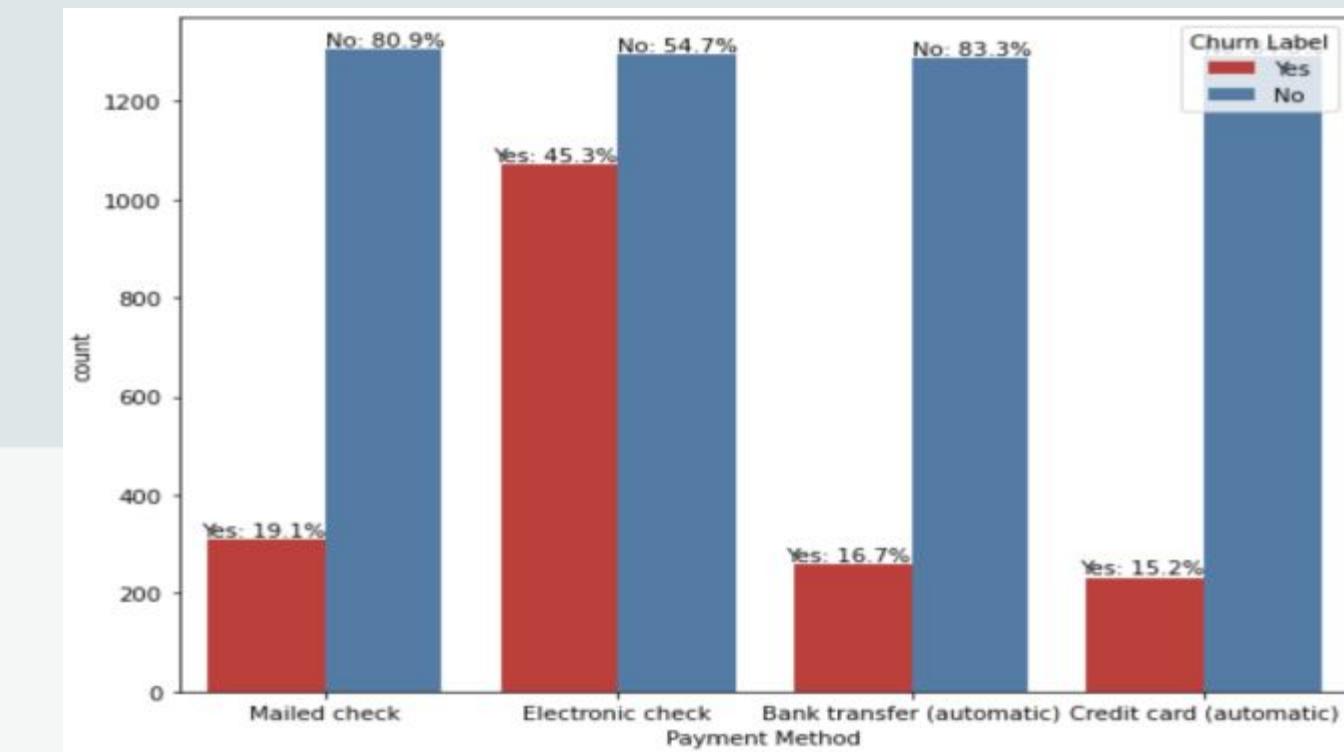
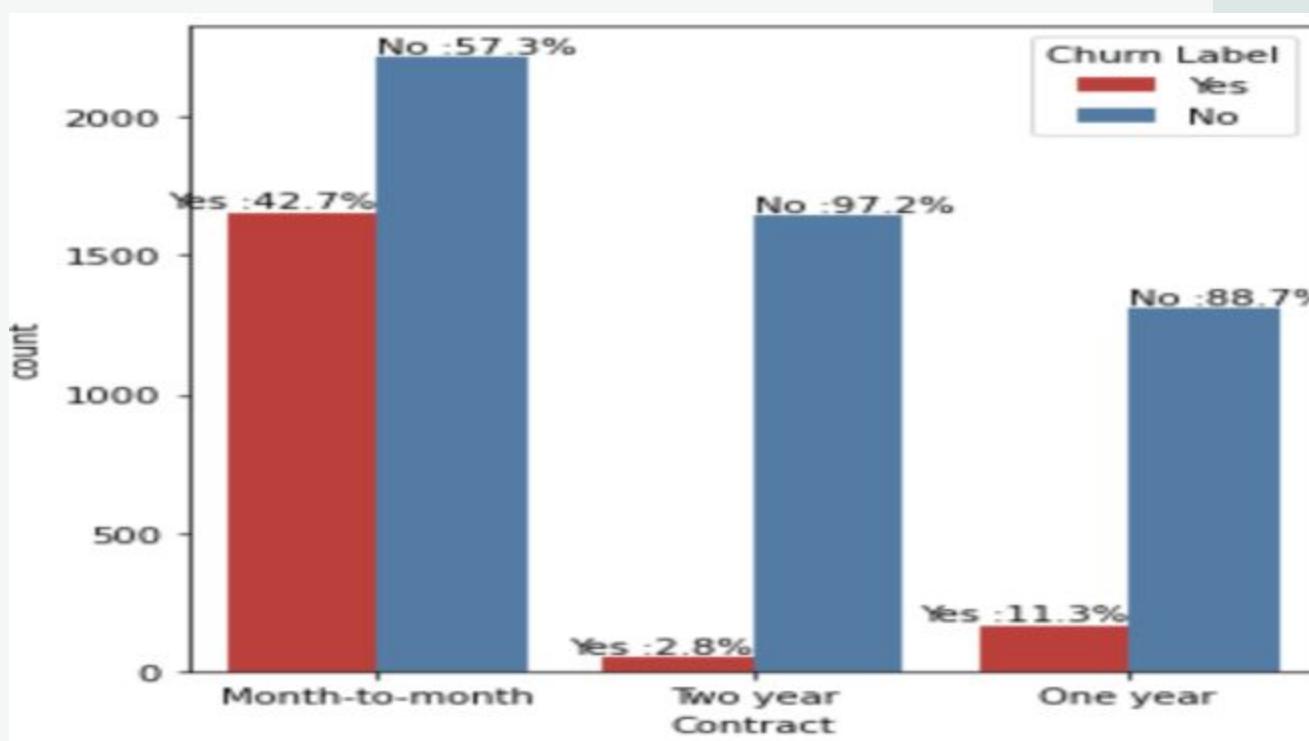
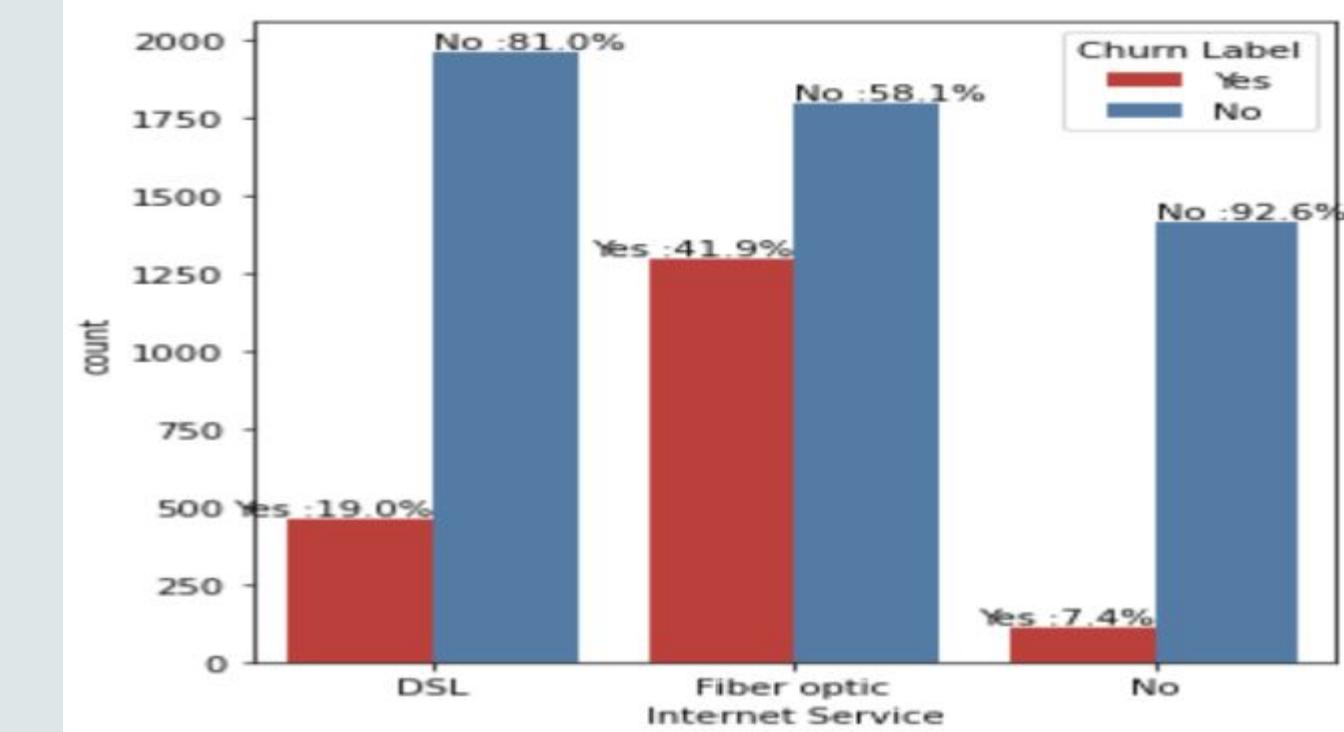
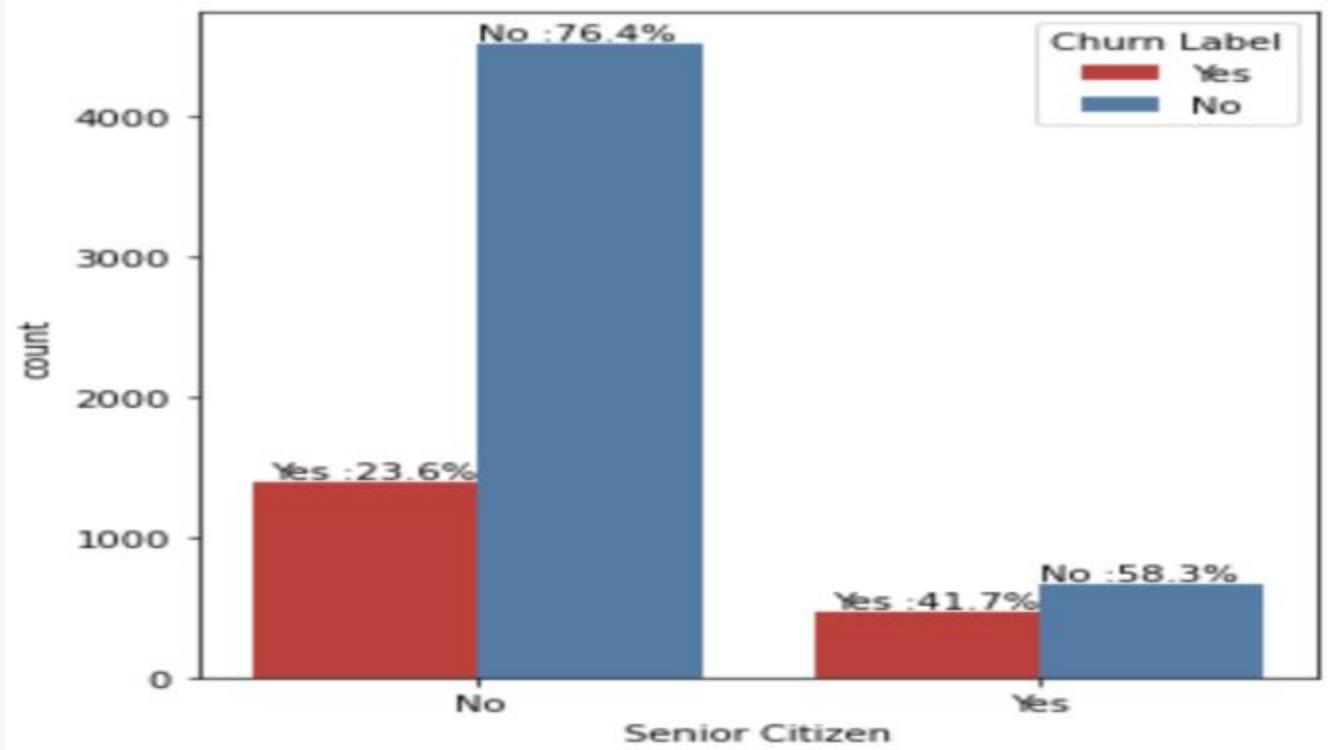
| | Column | d_type | unique_sample | n_uniques |
|----|-------------------|---------|--|-----------|
| 0 | CustomerID | object | [3668-QPYBK, 9237-HQITU, 9305-CDSKC, 7892-POOK...] | 7043 |
| 1 | Count | int64 | [1] | 1 |
| 2 | Country | object | [United States] | 1 |
| 3 | State | object | [California] | 1 |
| 4 | City | object | [Los Angeles, Beverly Hills, Huntington Park, ...] | 1129 |
| 5 | Zip Code | int64 | [90003, 90005, 90006, 90010, 90015] | 1652 |
| 6 | Lat Long | object | [33.964131, -118.272783, 34.059281, -118.30742...] | 1652 |
| 7 | Latitude | float64 | [33.964131, 34.059281, 34.048013, 34.062125, 3... | 1652 |
| 8 | Longitude | float64 | [-118.272783, -118.30742, -118.293953, -118.31... | 1651 |
| 9 | Gender | object | [Male, Female] | 2 |
| 10 | Senior Citizen | object | [No, Yes] | 2 |
| 11 | Partner | object | [No, Yes] | 2 |
| 12 | Dependents | object | [No, Yes] | 2 |
| 13 | Tenure Months | int64 | [2, 8, 28, 49, 10] | 73 |
| 14 | Phone Service | object | [Yes, No] | 2 |
| 15 | Multiple Lines | object | [No, Yes, No phone service] | 3 |
| 16 | Internet Service | object | [DSL, Fiber optic, No] | 3 |
| 17 | Online Security | object | [Yes, No, No internet service] | 3 |
| 18 | Online Backup | object | [Yes, No, No internet service] | 3 |
| 19 | Device Protection | object | [No, Yes, No internet service] | 3 |
| 20 | Tech Support | object | [No, Yes, No internet service] | 3 |
| 21 | Streaming TV | object | [No, Yes, No internet service] | 3 |
| 22 | Streaming Movies | object | [No, Yes, No internet service] | 3 |
| 23 | Contract | object | [Month-to-month, Two year, One year] | 3 |
| 24 | Paperless Billing | object | [Yes, No] | 2 |
| 25 | Payment Method | object | [Mailed check, Electronic check, Bank transfer...] | 4 |
| 26 | Monthly Charges | float64 | [53.85, 70.7, 99.65, 104.8, 103.7] | 1585 |
| 27 | Total Charges | object | [108.15, 151.65, 820.5, 3046.05, 5036.3] | 6531 |
| 28 | Churn Label | object | [Yes, No] | 2 |
| 29 | Churn Value | int64 | [1, 0] | 2 |
| 30 | Churn Score | int64 | [86, 67, 84, 89, 78] | 85 |
| 31 | CLTV | int64 | [3239, 2701, 5372, 5003, 5340] | 3438 |
| 32 | Churn Reason | object | [Competitor made better offer, Moved, Competit... | 20 |

데이터 EDA - 변수에 따른 이탈률(범주형)



데이터 EDA - 변수에 따른 이탈률(범주형)

영향력이 클 것으로 예상되는 변수 4개 확인



데이터 EDA - 결측치 확인

Numerical 변수가 Categorical로 분류 됨 -> float로 변환

그 중 11개 값은 결측

| | |
|------|---------|
| 0 | 108.15 |
| 1 | 151.65 |
| 2 | 820.50 |
| 3 | 3046.05 |
| 4 | 5036.30 |
| | ... |
| 7038 | 1419.40 |
| 7039 | 1990.50 |
| 7040 | 7362.90 |
| 7041 | 346.45 |
| 7042 | 6844.50 |

Name: Total Charges, Length: 7043, dtype: float64

| | |
|------|-----|
| 2234 | 0.0 |
| 2438 | 0.0 |
| 2568 | 0.0 |
| 2667 | 0.0 |
| 2856 | 0.0 |
| 4331 | 0.0 |
| 4687 | 0.0 |
| 5104 | 0.0 |
| 5719 | 0.0 |
| 6772 | 0.0 |
| 6840 | 0.0 |

Name: Total Charges, dtype: float64

데이터 EDA - 결측치 확인

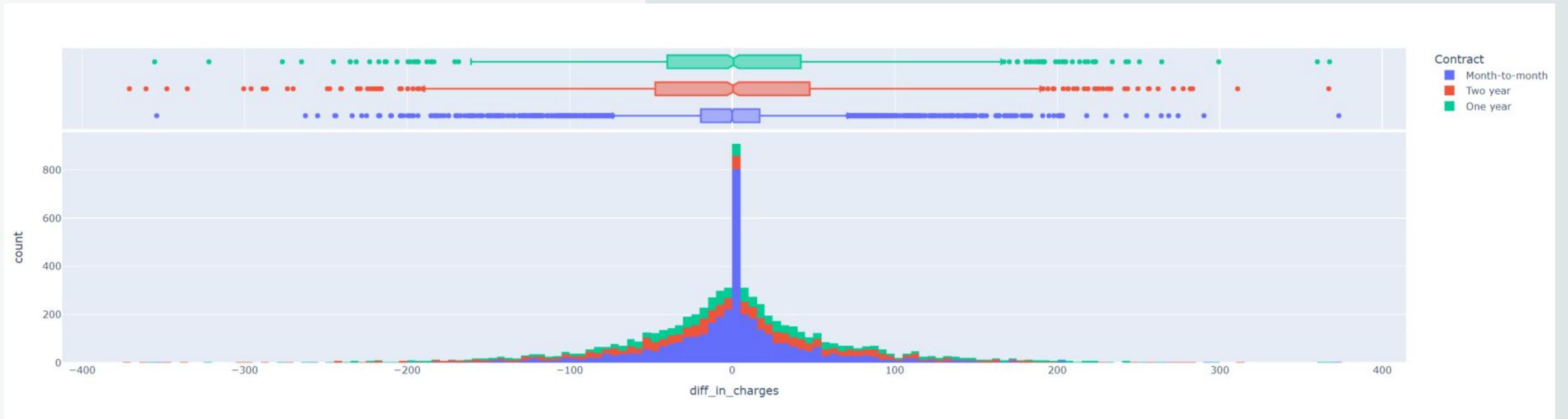
결측치 확인

결측치 확인

총 요금이 (이용 기간 * 달 요금) 값과 편차를 계산. → 차이를 고려하여 채우기
df[['Monthly Charges', 'Tenure Months', 'Total Charges']]

| | Monthly Charges | Tenure Months | Total Charges |
|------|-----------------|---------------|---------------|
| 0 | 53.85 | 2 | 108.15 |
| 1 | 70.70 | 2 | 151.65 |
| 2 | 99.65 | 8 | 820.5 |
| 3 | 104.80 | 28 | 3046.05 |
| 4 | 103.70 | 49 | 5036.3 |
| ... | ... | ... | ... |
| 7038 | 21.15 | 72 | 1419.4 |
| 7039 | 84.80 | 24 | 1990.5 |
| 7040 | 103.20 | 72 | 7362.9 |
| 7041 | 29.60 | 11 | 346.45 |
| 7042 | 105.65 | 66 | 6844.5 |

7043 rows × 3 columns



데이터 EDA - 결측치 대체

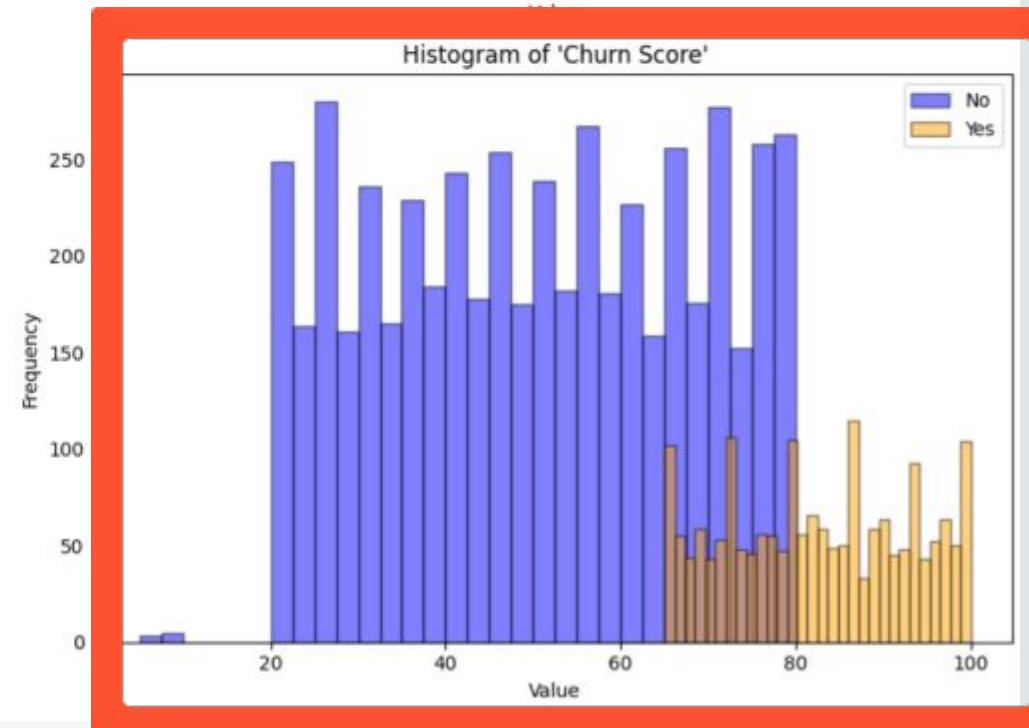
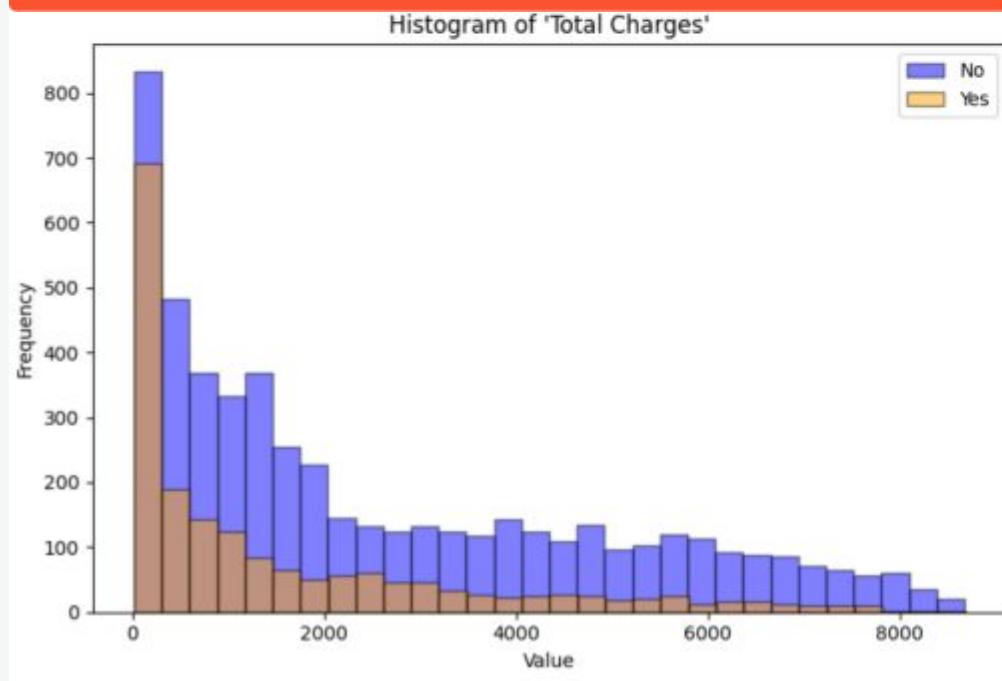
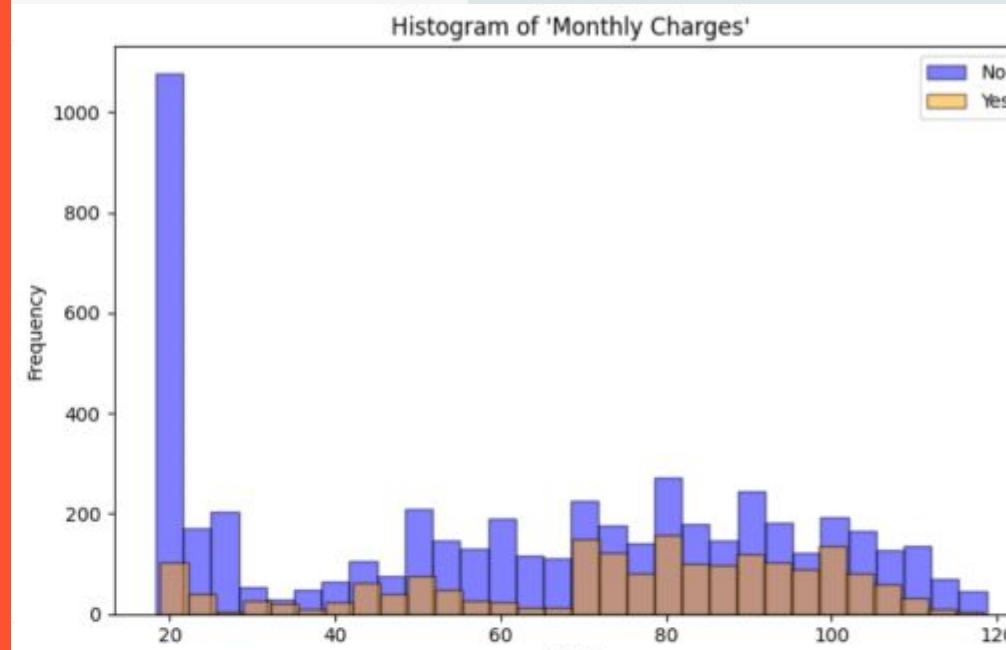
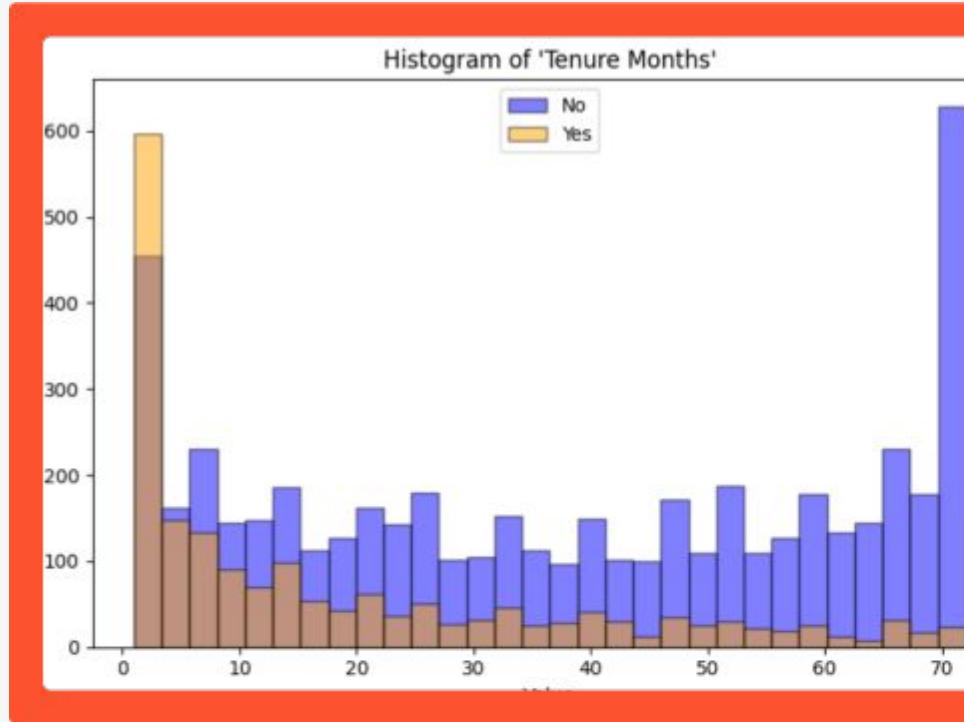
| | Total | Charges | diff_in_charges |
|----------------|-------|-----------|-----------------|
| Contract | | | |
| Month-to-month | 0.50 | 679.5500 | 0.0000 |
| | 0.80 | 2485.7300 | 24.8100 |
| | 0.90 | 3844.0600 | 54.0200 |
| | 0.95 | 4966.9200 | 85.3300 |
| One year | 0.50 | 2657.5500 | 0.7750 |
| | 0.80 | 5286.4600 | 55.0500 |
| | 0.90 | 6341.2500 | 92.2000 |
| | 0.95 | 7072.4725 | 133.3375 |
| Two year | 0.50 | 3623.9500 | 0.5000 |
| | 0.80 | 6399.2400 | 61.5300 |
| | 0.90 | 7457.6100 | 97.5700 |
| | 0.95 | 7922.3400 | 139.1800 |

약 20%의 고객이 20 달러 이상의 편차를 가짐

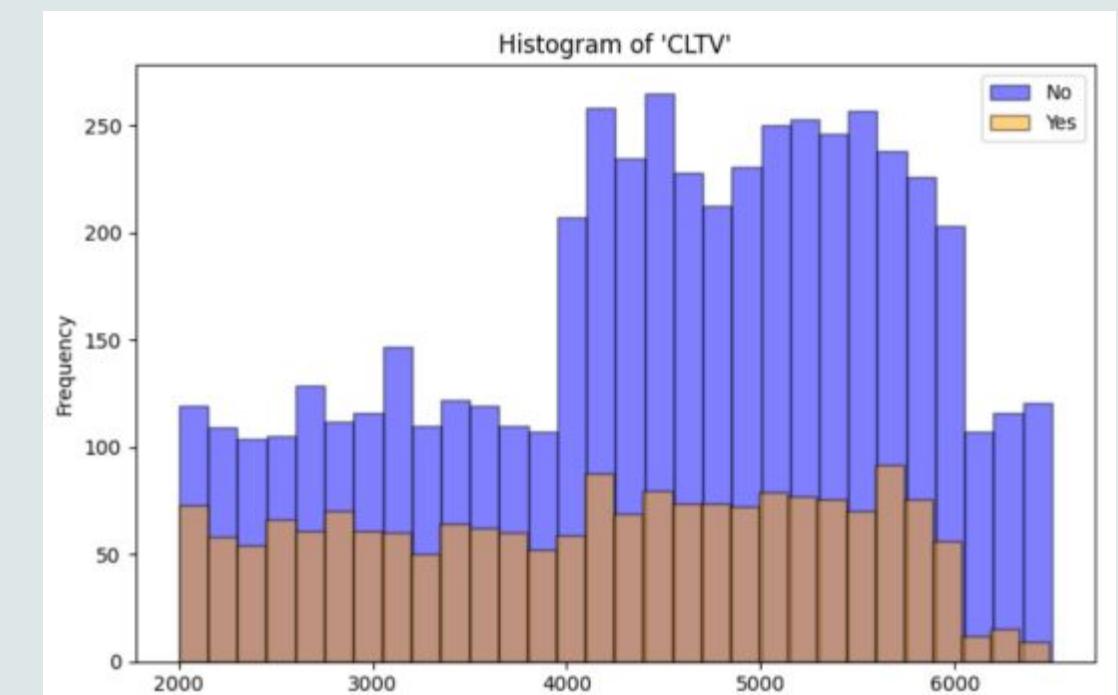


Monthly Charges x Tenure Months로 대체

데이터 EDA - 주요 변수(수치형)



Tenure Months, Churn Score
를 주요 변수로 예상



데이터 전처리

#필요 없는 열 지우기

```
df = df.drop(['Customer ID', 'City', 'Zip Code', 'Count', 'Country', 'State', 'Lat Long', 'Latitude', 'Longitude',  
'Churn Reason', 'Churn Value'], axis=1)
```

#Churn 값이 1이면 Yes, 0이면 No거나 다른 열의 값들도 다 바꿔주기

```
df = df.replace({'Yes': 1, 'No': 0, 'No phone service': 0, 'No internet service': 0})
```

#Gender의 값들도 바꿔주기

```
df = df.replace({'Male': 1, 'Female': 0})
```

#unique 값이 적은 컬럼 원핫 인코딩

```
df = pd.get_dummies(df, columns=['Contract', 'Payment Method', 'Internet Service'])  
df
```

데이터 EDA - Correlation

상관계수 Abs(0.6) 이상 파악

Tenure Months - Contract(Month-to-Month) **-0.65**

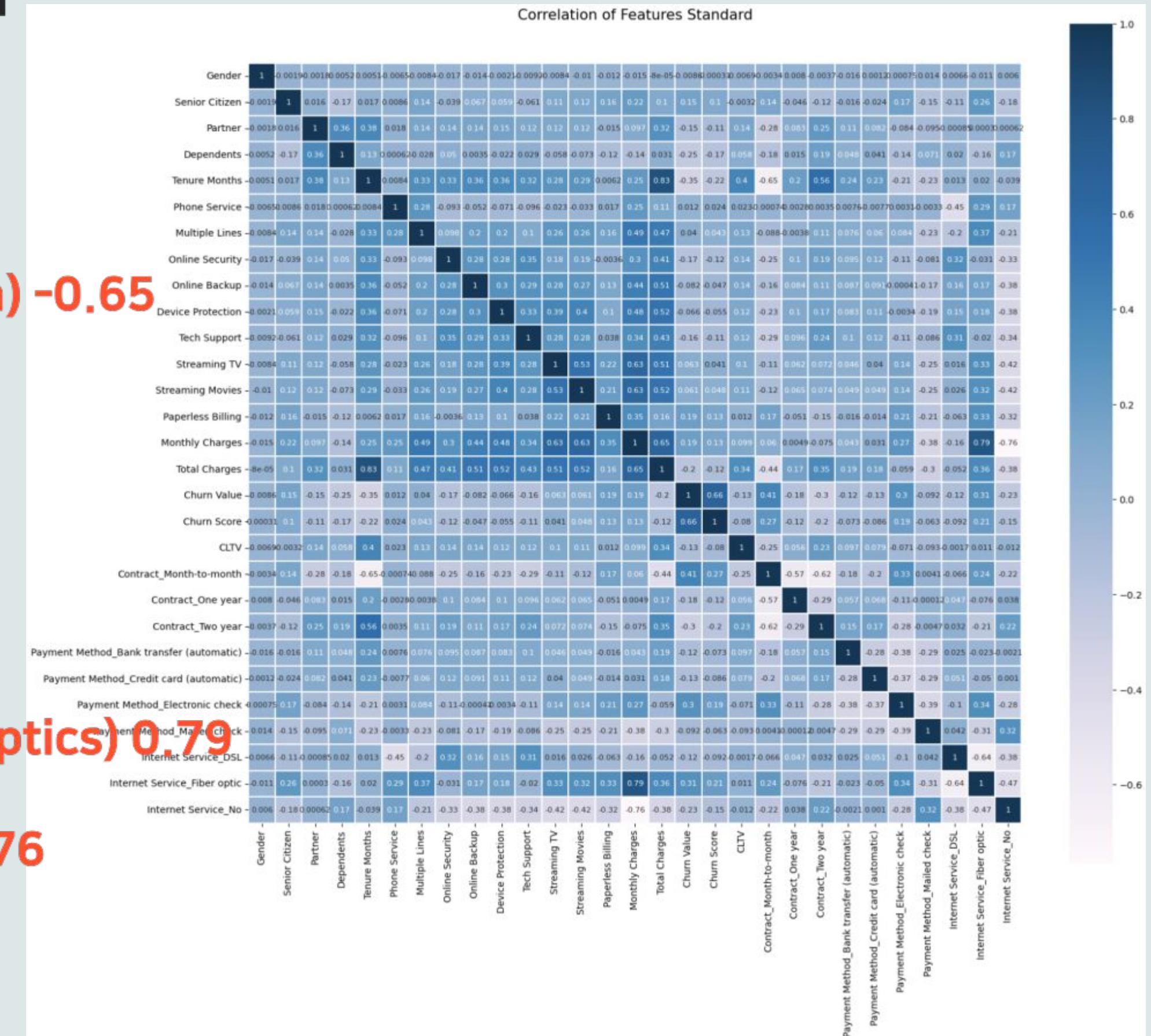
Streaming TV - Monthly Charges **0.63**

Streaming Movies - Monthly Charges **0.63**

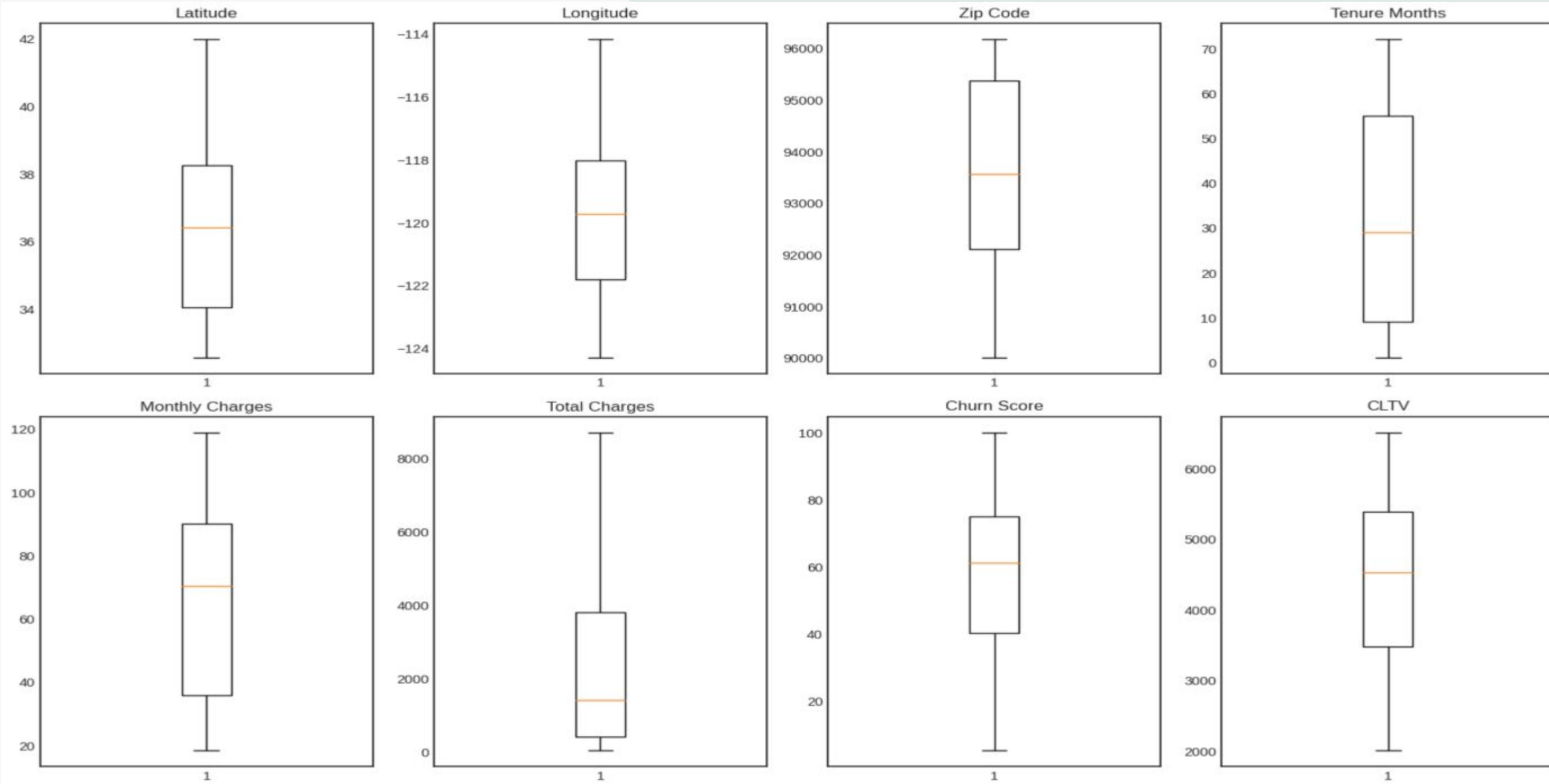
Monthly Charges - Total Charges **0.65**

Monthly Charges - Internet Service(Fiber Optics) **0.79**

Monthly Charges - Internet Service(No) **-0.76**



데이터 EDA - Outlier 확인



데이터 전처리 - Scaling

연속형 변수 전처리 MinMaxScaler, StandardScaler 사용

```
# 스케일링할 컬럼 선택 -> 수치형 변수  
cols_to_scale = ['Tenure Months', 'Monthly Charges', 'Churn Score', 'Total Charges', 'CLTV']  
  
# StandardScaler 객체 생성 후 fit & transform 수행  
sdscaler = StandardScaler()  
df[cols_to_scale] = sdscaler.fit_transform(df[cols_to_scale])  
  
# 데이터 프레임으로 저장  
stdscaled_df = df  
  
# MinMaxScaler 객체 생성 후 fit & transform 수행  
mMscaler = MinMaxScaler()  
df2[cols_to_scale] = mMscaler.fit_transform(df2[cols_to_scale])  
  
# 데이터 프레임으로 저장  
mMscaled_df = df2
```

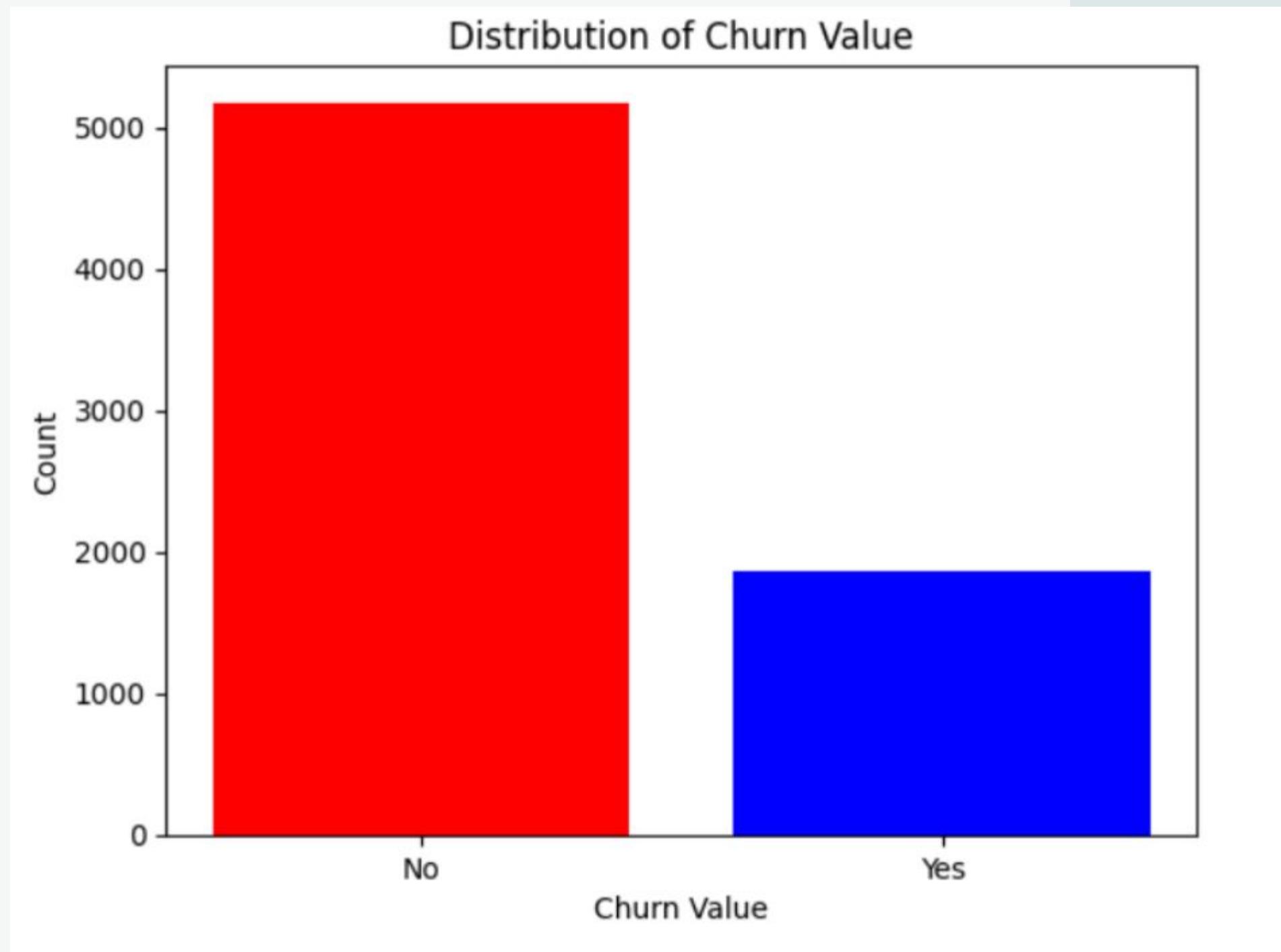
Gender를 제외하고 변수간 상관성이 존재함. -> Gender drop
stdscaled_df = stdscaled_df.drop('Gender', axis=1)
mMscaled_df = mMscaled_df.drop('Gender', axis=1)

**기본적인 Scaler
StandardScaler**

**추가적으로
MinMaxScaler 사용**

**+ 상관관계 낮은
Gender Drop**

데이터 전처리 - OverSampling



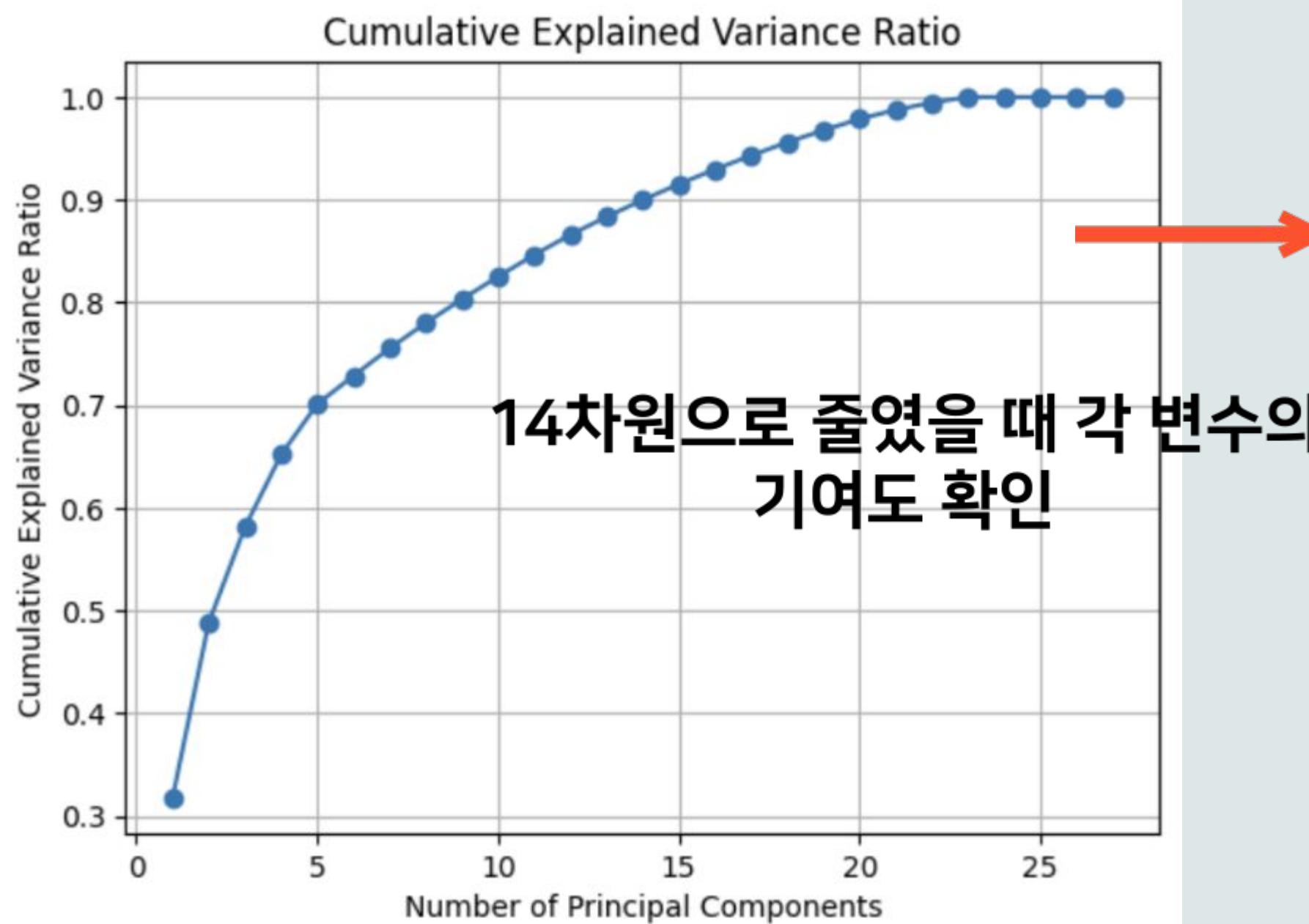
종속 변수 Churn Value 불균형

7043개 중
No : 5174, Yes : 1869
SMOTE > OverSampling 수행

```
minMax Churn Value
0    5174
1    5174
Name: Churn Value, dtype: int64
standard Churn Value
0    5174
1    5174
Name: Churn Value, dtype: int64
```

데이터 전처리 - PCA

StandardScaler - 14차원 축소



Explained Variance up to Dimension 14: 0.90

Tech Support Contribution: 0.3834
Online Security Contribution: 0.3710
Senior Citizen Contribution: 0.3023
Contract_Month-to-month Contribution: 0.2710
Contract_Two year Contribution: 0.2580
Multiple Lines Contribution: 0.2552
Paperless Billing Contribution: 0.2497
Payment Method_Electronic check Contribution: 0.2304
Online Backup Contribution: 0.2194
Device Protection Contribution: 0.2128
Streaming Movies Contribution: 0.1934
Payment Method_Mailed check Contribution: 0.1770
Internet Service_DSL Contribution: 0.1601
Internet Service_Fiber optic Contribution: 0.1582
Partner Contribution: 0.1509
Streaming TV Contribution: 0.1487
Dependents Contribution: 0.1285
Phone Service Contribution: 0.1009
Monthly Charges Contribution: 0.0954
Total Charges Contribution: 0.0753
Payment Method_Credit card (automatic) Contribution: 0.0557
Tenure Months Contribution: 0.0221
Churn Score Contribution: 0.0200
Contract_One year Contribution: 0.0130
CLTV Contribution: 0.0100
Payment Method_Bank transfer (automatic) Contribution: 0.0023
Internet Service_No Contribution: 0.0019

데이터 전처리 - PCA

StandardScaler - 14차원 축소

```
Cross-validation scores: [0.79408213 0.81461353 0.81038647 0.80483384 0.81510574]
Mean CV score: 0.8078043405286278
Model score: 0.8144927536231884
Confusion matrix:
[[765 236]
 [148 921]]
F1 score: 0.8274932614555256
Recall score: 0.8615528531337698
```

Gradient Boost

```
Cross-validation scores: [0.81763285 0.8423913 0.83756039 0.82296073 0.84108761]
Mean CV score: 0.8323265758862763
Model score: 0.8439613526570048
Confusion matrix:
[[812 189]
 [134 935]]
F1 score: 0.8527131782945737
Recall score: 0.8746492048643593
```

XGBoost

```
Cross-validation scores: [0.83160801 0.82677709 0.83643892 0.81837017 0.84530387]
Mean CV score: 0.8316996099424637
Model score: 0.8360708534621578
Confusion matrix:
[[1213 301]
 [ 208 1383]]
F1 score: 0.8445801526717557
Recall score: 0.86926461345066
```

LightGBM

```
Cross-validation scores: [0.83712905 0.84817115 0.83574879 0.81906077 0.84461326]
Mean CV score: 0.836944606491808
Model score: 0.8553945249597423
Confusion matrix:
[[1270 244]
 [ 205 1386]]
F1 score: 0.8606022974231604
Recall score: 0.8711502199874293
```

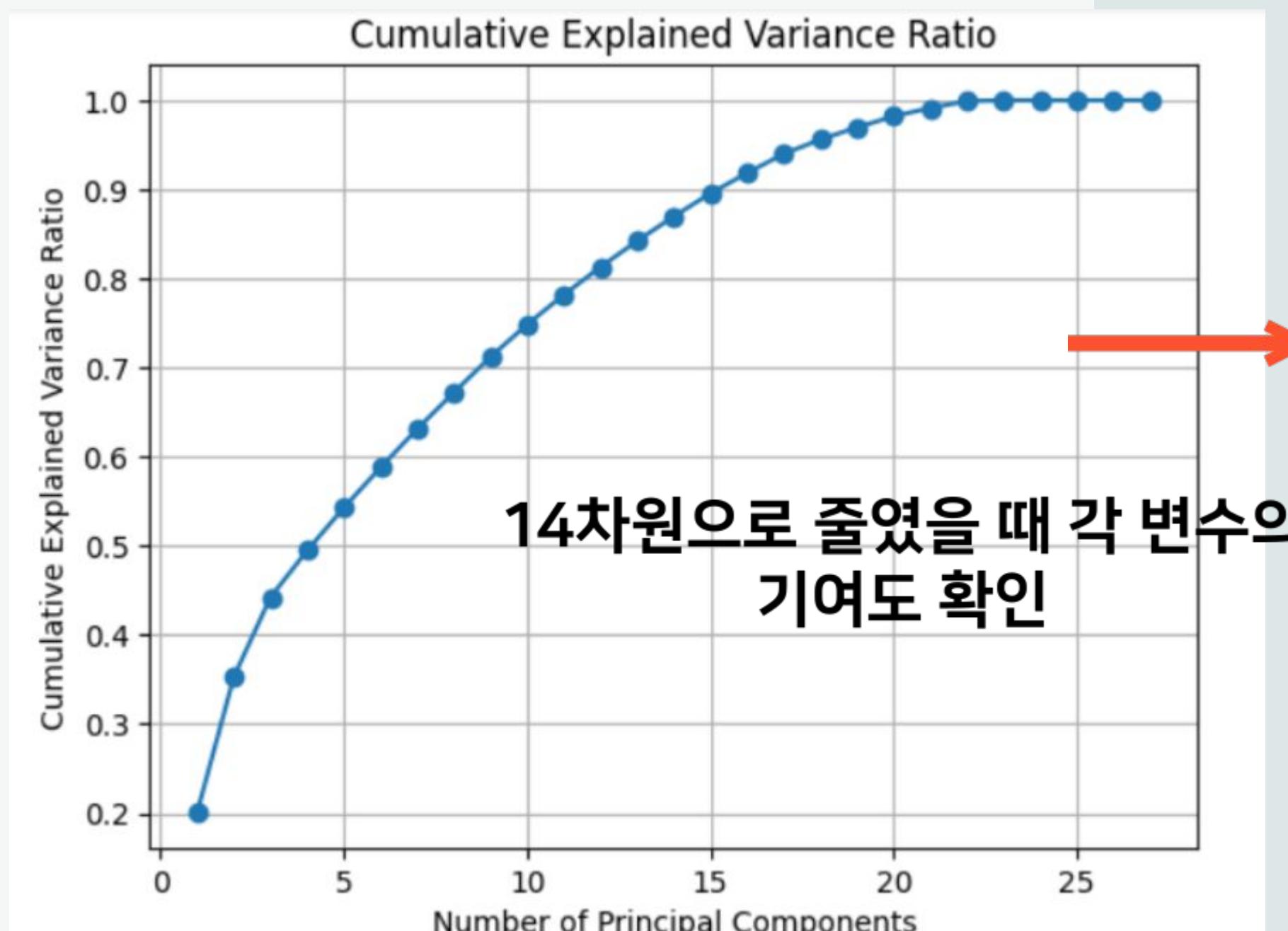
RF



Random Forest
0.86 > MAX F1 Score

데이터 전처리 - PCA

MinMaxScaler - 14차원 축소



Explained Variance up to Dimension 14: 0.85
Device Protection Contribution: 0.6059
Senior Citizen Contribution: 0.4146
Dependents Contribution: 0.3354
Streaming TV Contribution: 0.2955
Partner Contribution: 0.2283
Online Backup Contribution: 0.2015
Tech Support Contribution: 0.1873
Streaming Movies Contribution: 0.1829
Online Security Contribution: 0.1571
Payment Method_Mailed check Contribution: 0.1515
Multiple Lines Contribution: 0.1245
Payment Method_Credit card (automatic) Contribution: 0.0922
Paperless Billing Contribution: 0.0804
Contract_Two year Contribution: 0.0703
Payment Method_Bank transfer (automatic) Contribution: 0.0570
Tenure Months Contribution: 0.0534
Phone Service Contribution: 0.0530
Internet Service_Fiber optic Contribution: 0.0517
Contract_One year Contribution: 0.0384
Contract_Month-to-month Contribution: 0.0319
Internet Service_DSL Contribution: 0.0305
Monthly Charges Contribution: 0.0277
Internet Service_No Contribution: 0.0213
Total Charges Contribution: 0.0190
Gender Contribution: 0.0098
Churn Score Contribution: 0.0079
CLTV Contribution: 0.0078
Payment Method_Electronic check Contribution: 0.0022

데이터 전처리 - PCA

StandardScaler - 14차원 축소

```
Cross-validation scores: [0.9263285  0.92572464  0.9384058   0.92326284  0.93353474]
Mean CV score: 0.9294513040559277
Model score: 0.9314009661835749
Confusion matrix:
[[ 917   84]
 [ 58 1011]]
F1 score: 0.934380776340111
Recall score: 0.9457436856875585
```

Gradient Boost

```
Cross-validation scores: [0.93297101 0.92391304 0.93719807 0.93353474 0.94441088]
Mean CV score: 0.9344055489878424
Model score: 0.9439613526570049
Confusion matrix:
[[ 925   76]
 [ 40 1029]]
F1 score: 0.9466421343146273
Recall score: 0.9625818521983162
```

XGBoost

```
Cross-validation scores: [0.92477571 0.93995859 0.93029676 0.92472376 0.95372928]
Mean CV score: 0.9346968189149308
Model score: 0.9388083735909822
Confusion matrix:
[[1394  120]
 [ 70 1521]]
F1 score: 0.9412128712871287
Recall score: 0.9560025141420491
```

LightGBM

```
Cross-validation scores: [0.92477571 0.93581781 0.92615597 0.92748619 0.94475138]
Mean CV score: 0.931797410292486
Model score: 0.9391304347826087
Confusion matrix:
[[1402  112]
 [ 77 1514]]
F1 score: 0.9412496114392293
Recall score: 0.9516027655562539
```

RF



XGBoost
0.94 > MAX F1 Score

Modeling - 기본 값

StratifiedKFold 1,2,3,4,5 평균

StandardScaler

| 모델 | 정확도 | 재현율 | 정밀도 | F1 스코어 |
|--------------|------|------|------|--------|
| GBM | 0.94 | 0.96 | 0.92 | 0.94 |
| XGBoost | 0.95 | 0.96 | 0.93 | 0.95 |
| LightGBM | 0.95 | 0.96 | 0.93 | 0.95 |
| RandomForest | 0.94 | 0.97 | 0.92 | 0.95 |

MinMaxScaler

| 모델 | 정확도 | 재현율 | 정밀도 | F1 스코어 |
|--------------|------|------|------|--------|
| GBM | 0.93 | 0.96 | 0.92 | 0.94 |
| XGBoost | 0.95 | 0.96 | 0.93 | 0.95 |
| LightGBM | 0.95 | 0.97 | 0.93 | 0.95 |
| RandomForest | 0.94 | 0.97 | 0.92 | 0.94 |

| Model | Accuracy | Recall | Precision | F1-Score |
|----------|----------|--------|-----------|----------|
| CatBoost | 0.93 | 0.86 | 0.86 | 0.86 |

Optimization - 최적화

GridSearch 이용 - 최적화 StandardScaler, MinMaxScaler 동일범위

| 파라미터 | GBM | XGB | LGBM | RF | CatBoost |
|-------------------|-------------------|-------------------|-------------------|-------------------|------------------|
| 양상블에 포함될 기본 모델 개수 | | | | | |
| n_estimators | 50, 100, 200, 500 | 50, 100, 200, 500 | 50, 100, 200, 500 | 50, 100, 200, 500 | 300, 500, 600 |
| 학습률 | | | | | |
| learning_rate | 0.1, 0.01, 0.001 | 0.1, 0.01, 0.001 | 0.1, 0.01, 0.001 | N/A | 0.009, 0.01, 0.1 |
| 최대 깊이 | | | | | |
| max_depth | 4, 6, 8 | 4, 6, 8 | 4, 6, 8 | 4, 6, 8 | 4, 6, 8 |

Optimization - 최적화

StandardScaler 최적값

| 모델 | n_estimators | learning_rate | max_depth |
|------------------|--------------|---------------|-----------|
| XGB | 500 | 0.1 | 4 |
| GradientBoosting | 200 | 0.1 | 4 |
| LGBM | 500 | 0.1 | 6 |
| RandomForest | 500 | N/A | 8 |

MinMaxScaler 최적값

| 모델 | n_estimators | learning_rate | max_depth |
|------------------|--------------|---------------|-----------|
| XGB | 500 | 0.1 | 4 |
| GradientBoosting | 200 | 0.1 | 4 |
| LGBM | 500 | 0.1 | 6 |
| RandomForest | 500 | N/A | 8 |

| 모델 | depth | iterations | learning_rate |
|----------|-------|------------|---------------|
| CatBoost | 6 | 300 | 0.009 |

Modeling - 최적화 된 모델

StratifiedKFold 1,2,3,4,5 평균

StandardScaler

| Model | Accuracy | Recall | Precision | F1-Score |
|--------------|----------|--------|-----------|----------|
| GBM | 0.94 | 0.96 | 0.93 | 0.95 |
| XGBoost | 0.94 | 0.96 | 0.93 | 0.95 |
| LightGBM | 0.94 | 0.96 | 0.93 | 0.95 |
| RandomForest | 0.94 | 0.97 | 0.92 | 0.95 |

MinMaxScaler

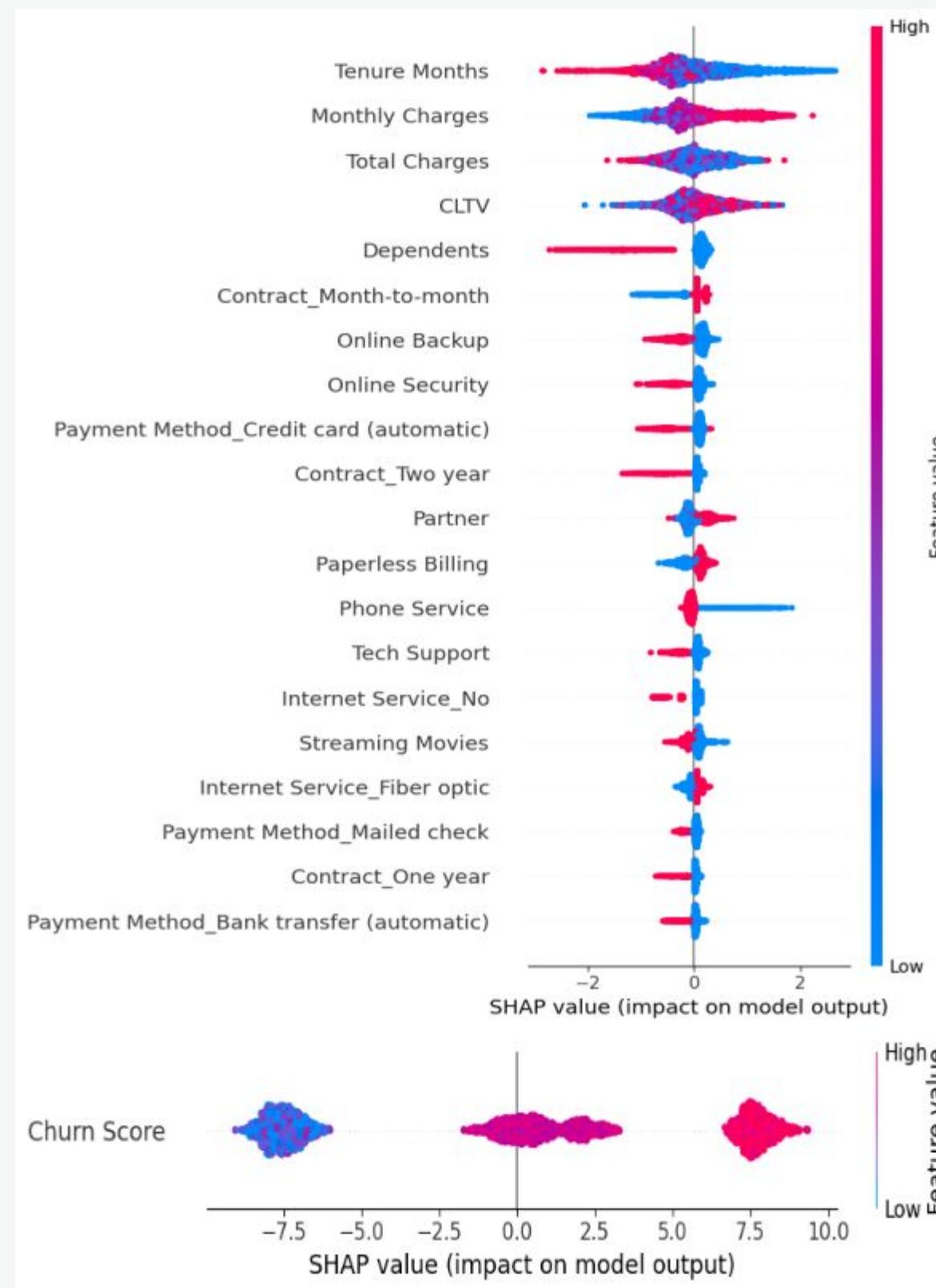
| Model | Accuracy | Recall | Precision | F1-Score |
|--------------|----------|--------|-----------|----------|
| GBM | 0.94 | 0.97 | 0.93 | 0.95 |
| XGBoost | 0.94 | 0.96 | 0.93 | 0.95 |
| LightGBM | 0.94 | 0.96 | 0.93 | 0.95 |
| RandomForest | 0.94 | 0.97 | 0.92 | 0.94 |

| Model | Accuracy | Recall | Precision | F1-Score |
|----------|----------|--------|-----------|----------|
| CatBoost | 0.93 | 0.86 | 0.86 | 0.86 |

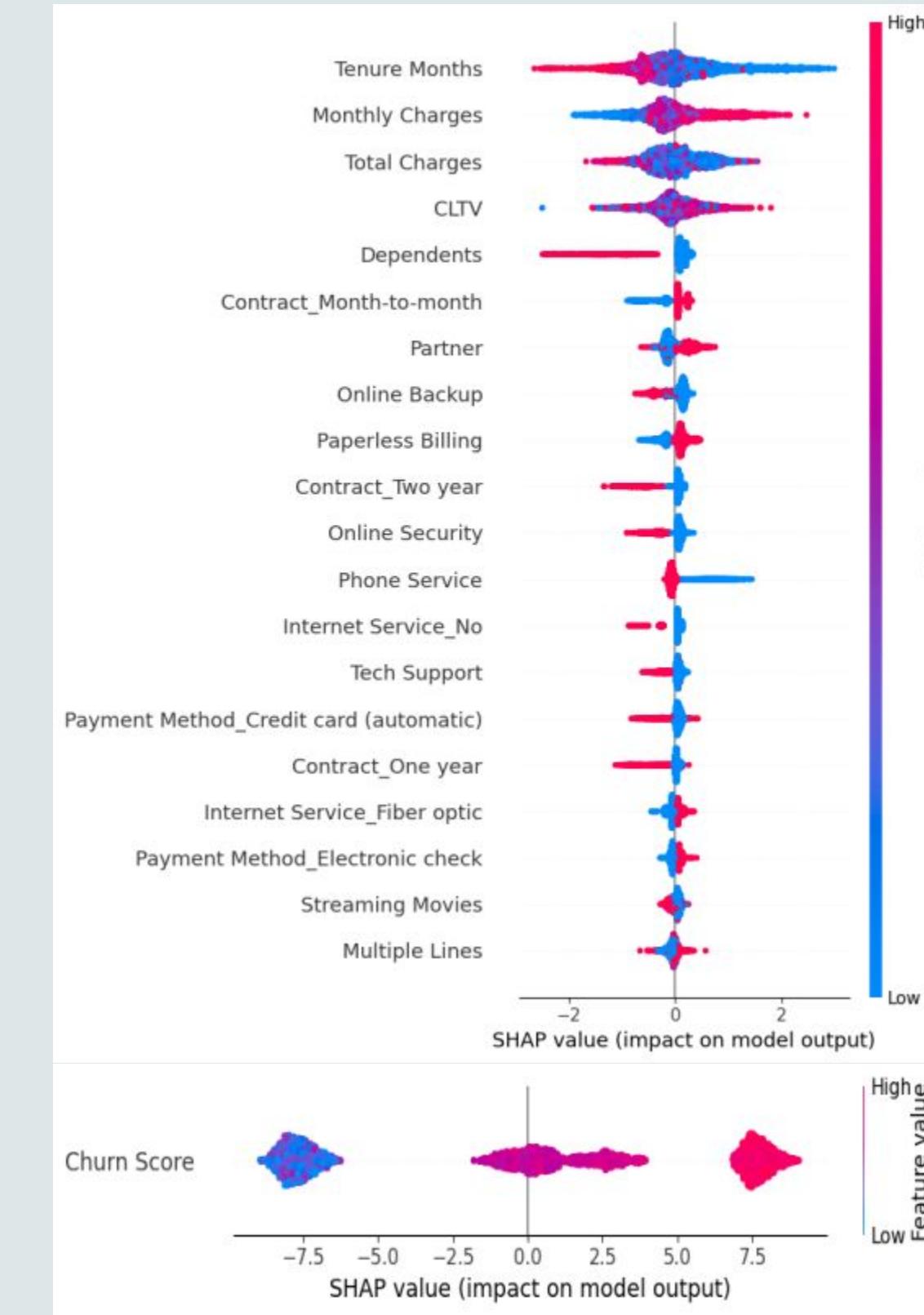
XAI - SHAP

XGBoost

StandardScaler



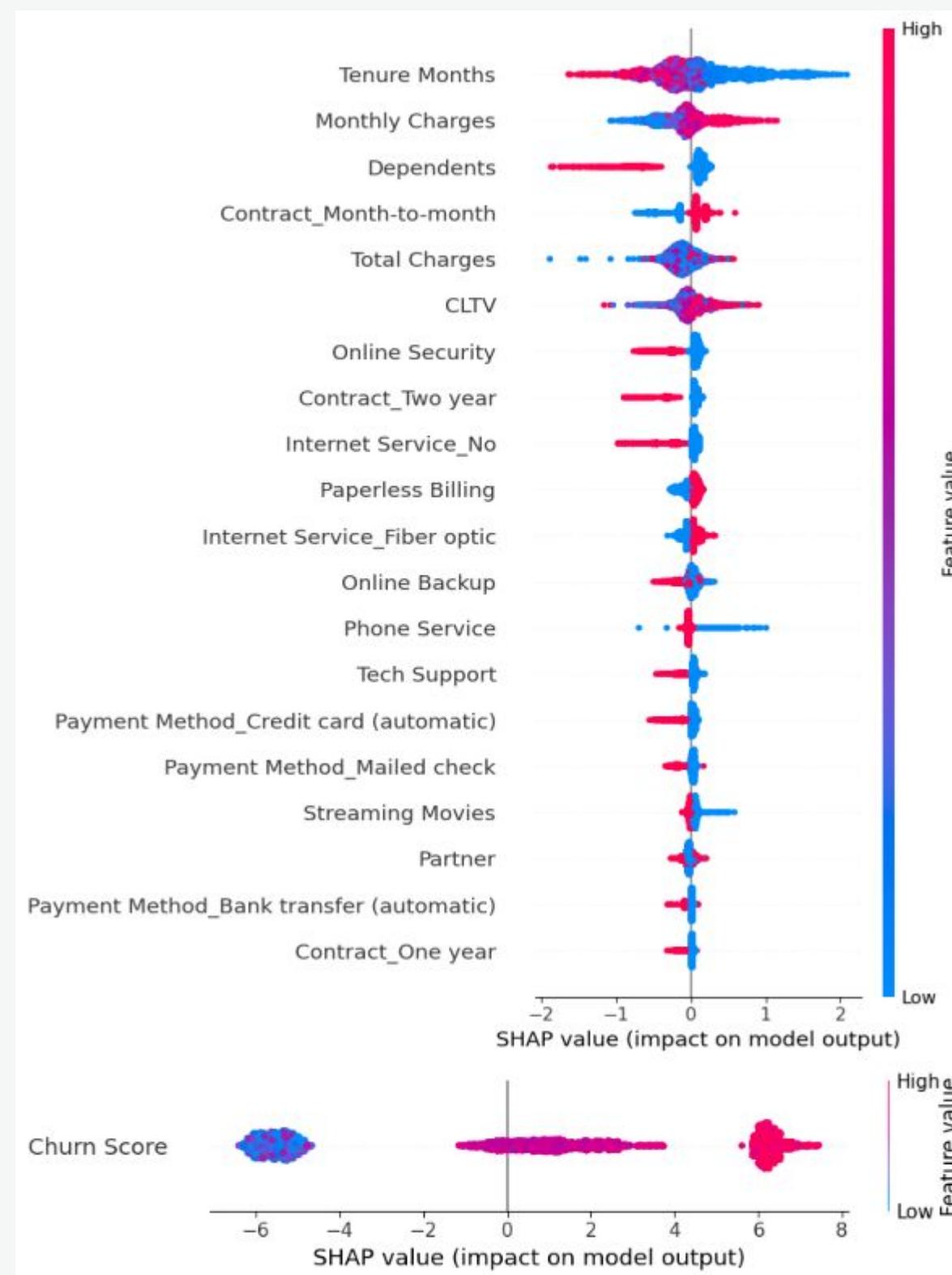
MinMaxScaler



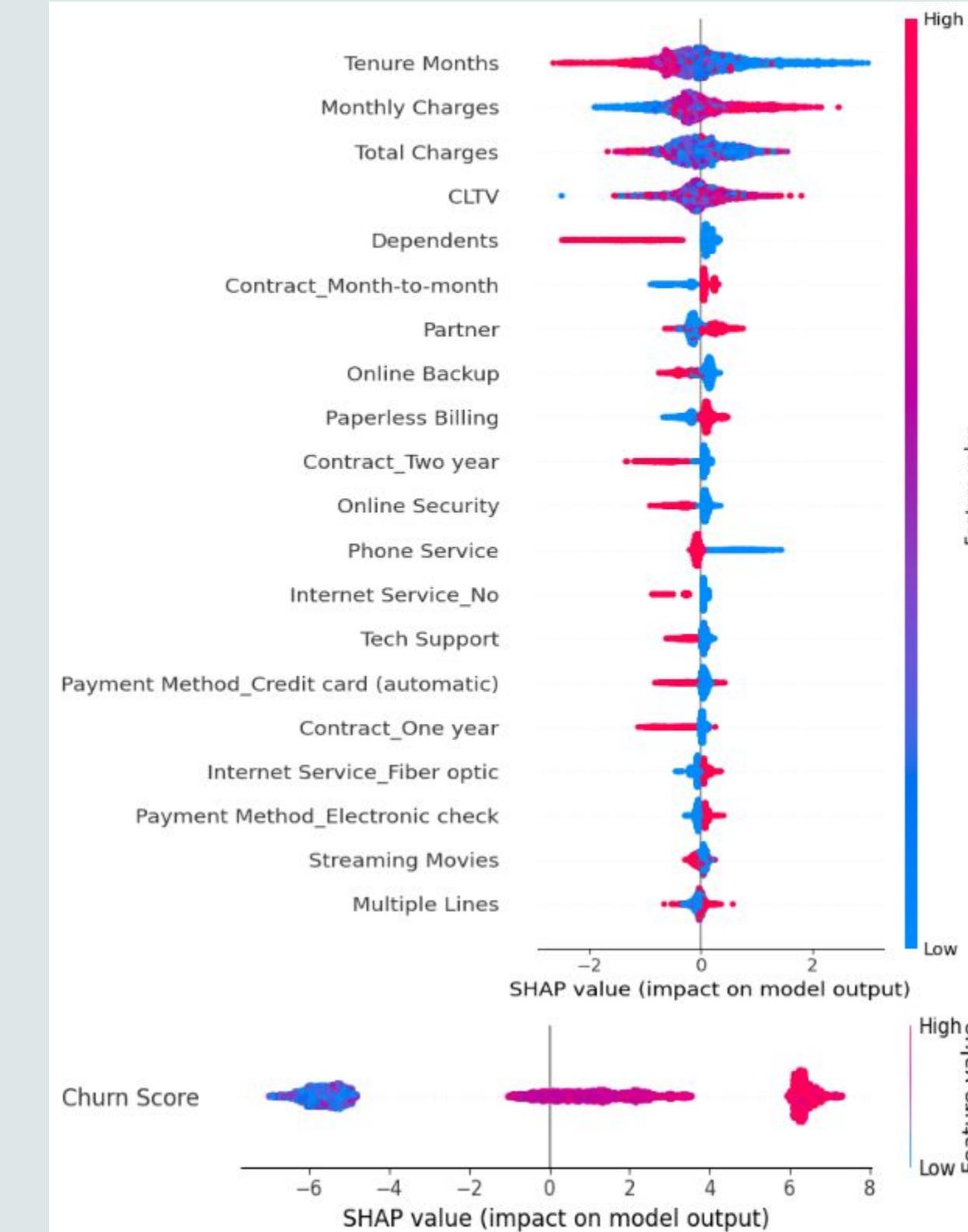
XAI - SHAP

Gradient Boosting

StandardScaler



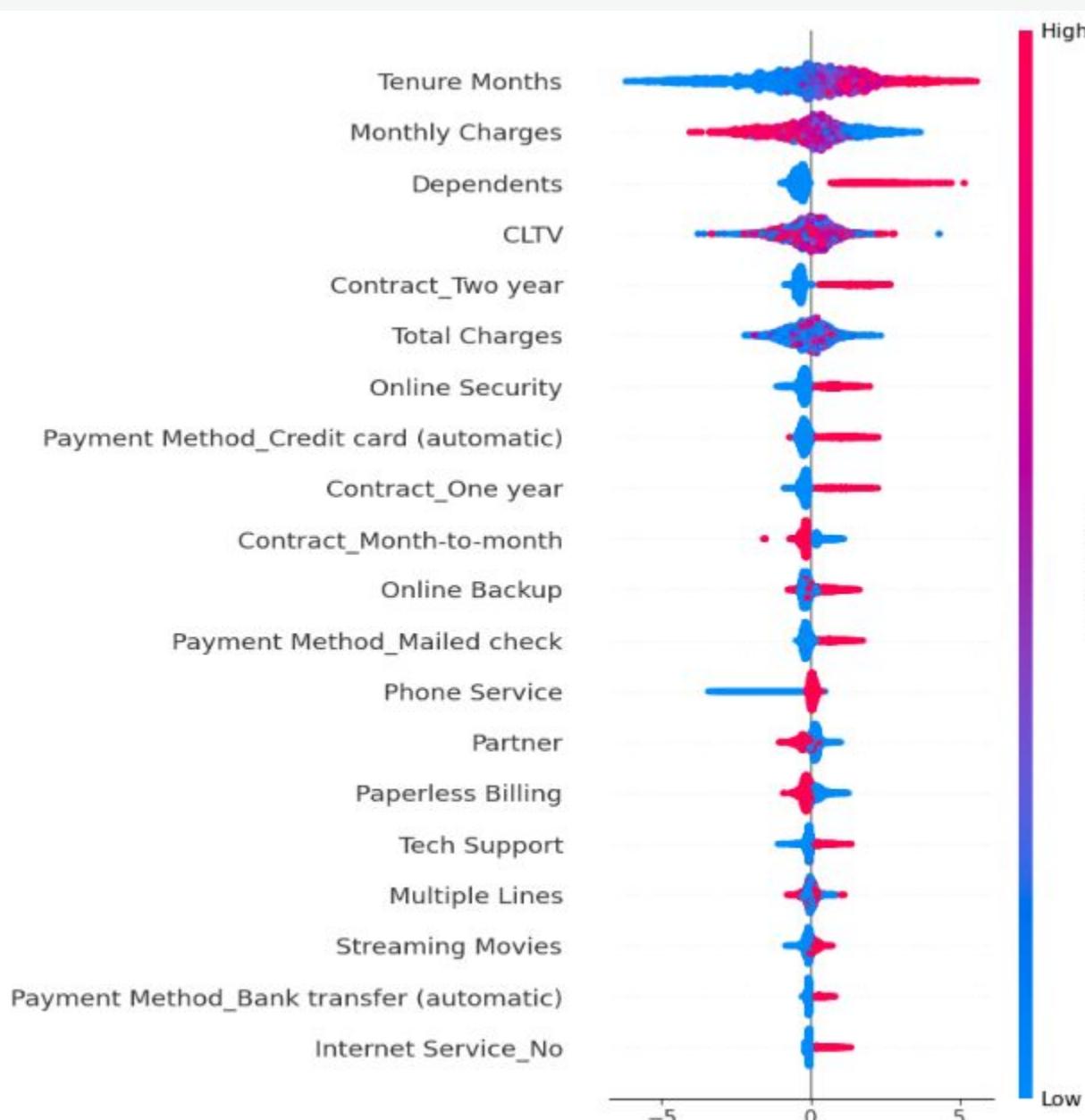
MinMaxScaler



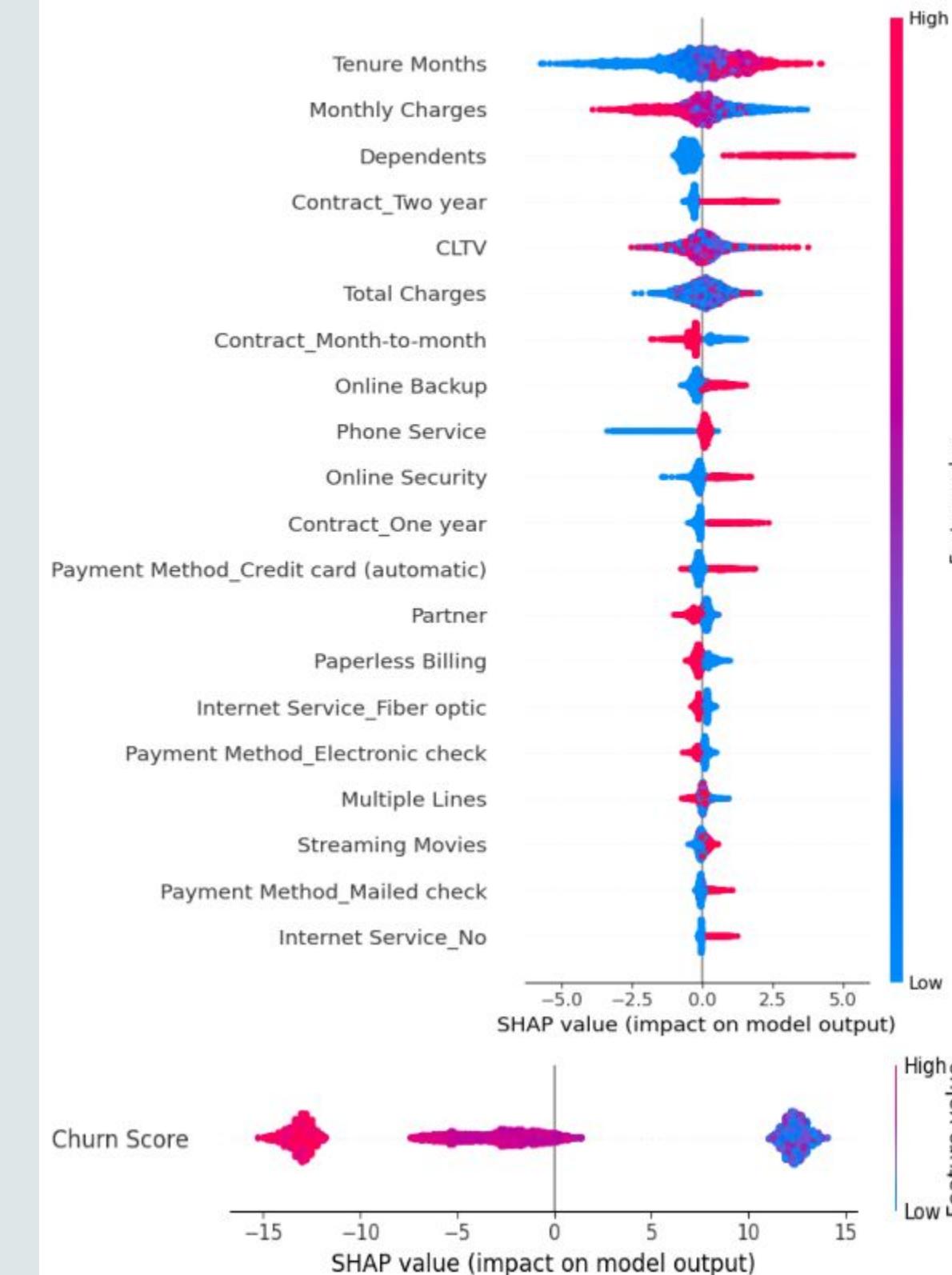
XAI - SHAP

LightGBM

StandardScaler



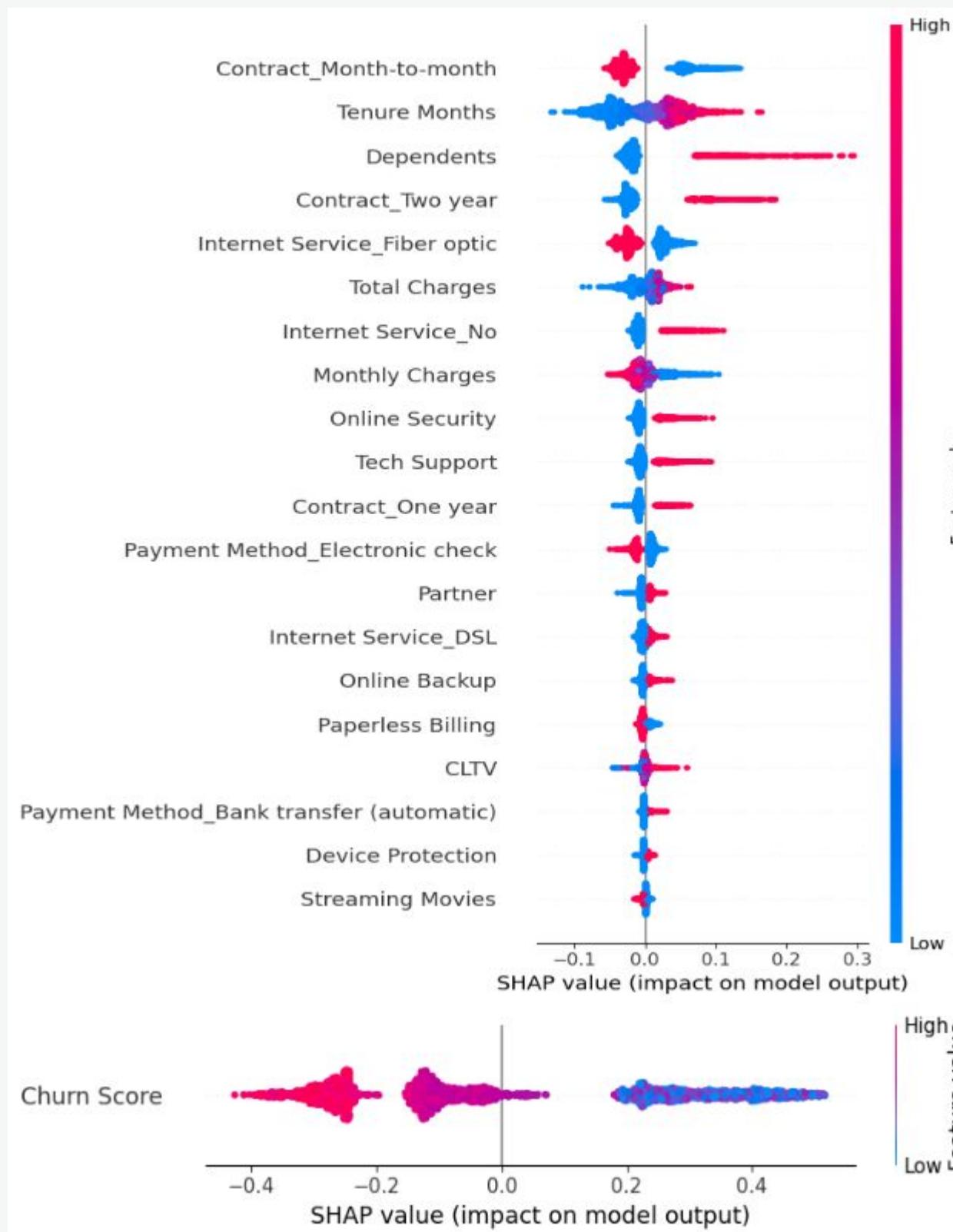
MinMaxScaler



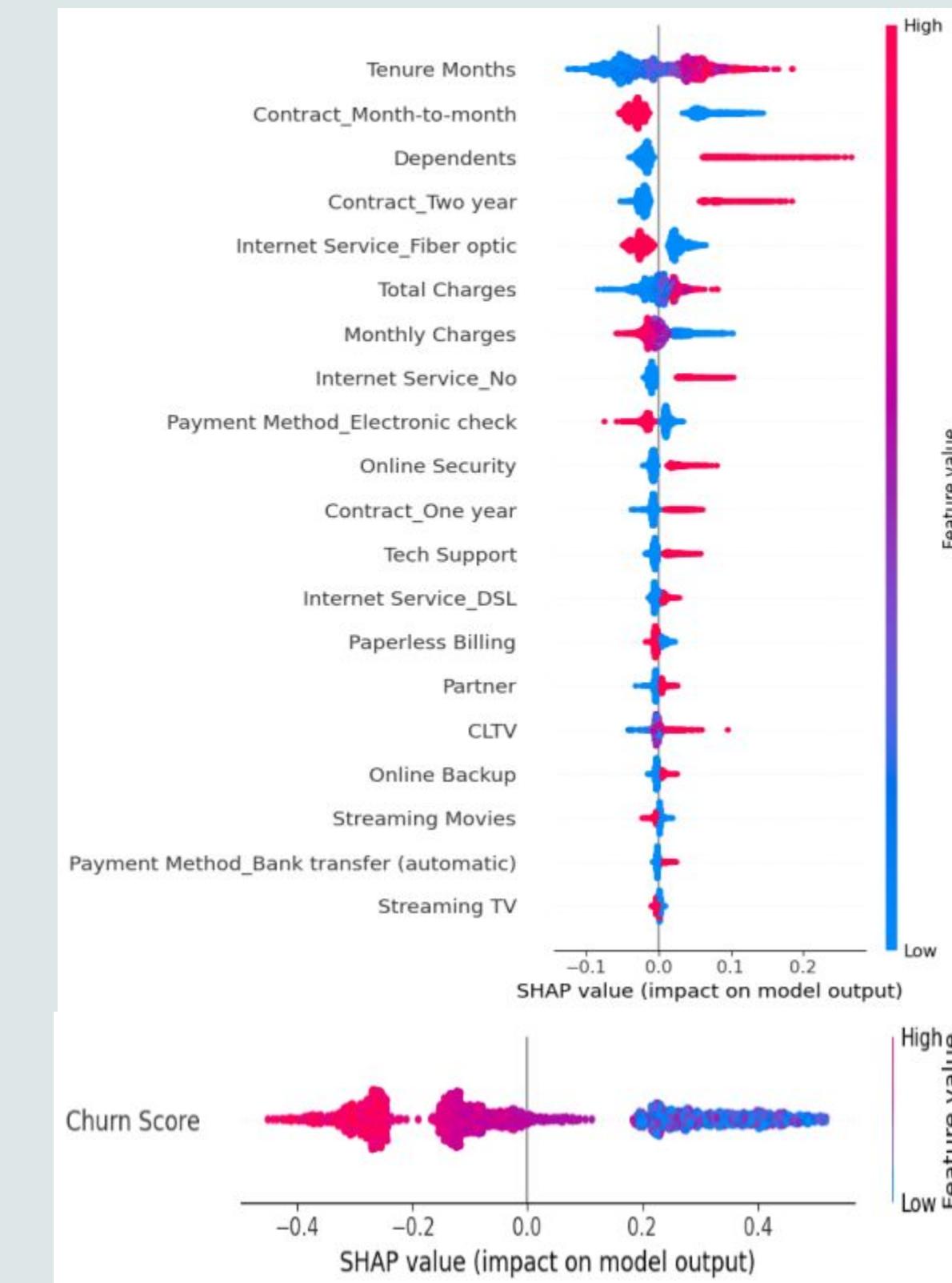
XAI - SHAP

Random Forest

StandardScaler

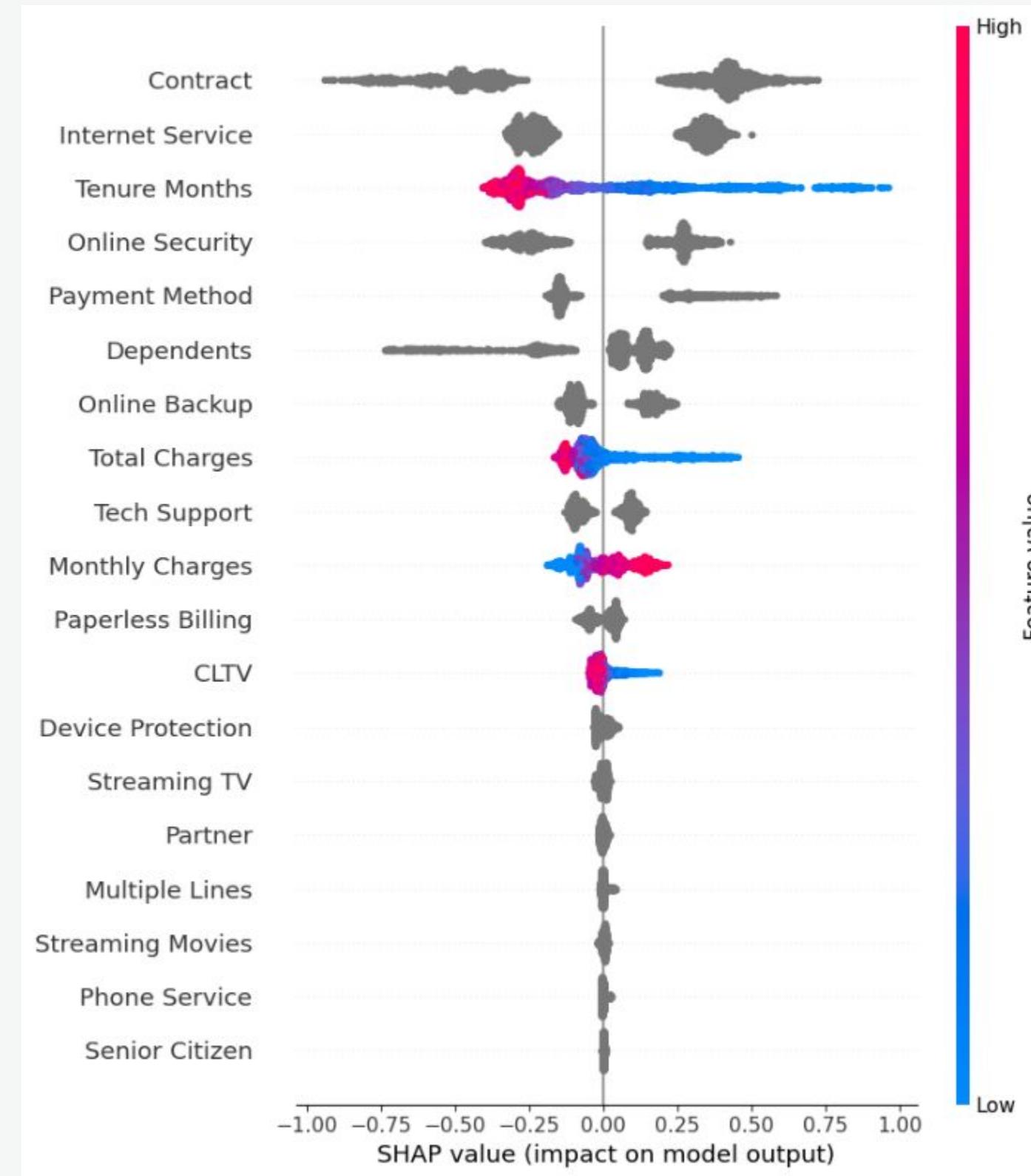


MinMaxScaler

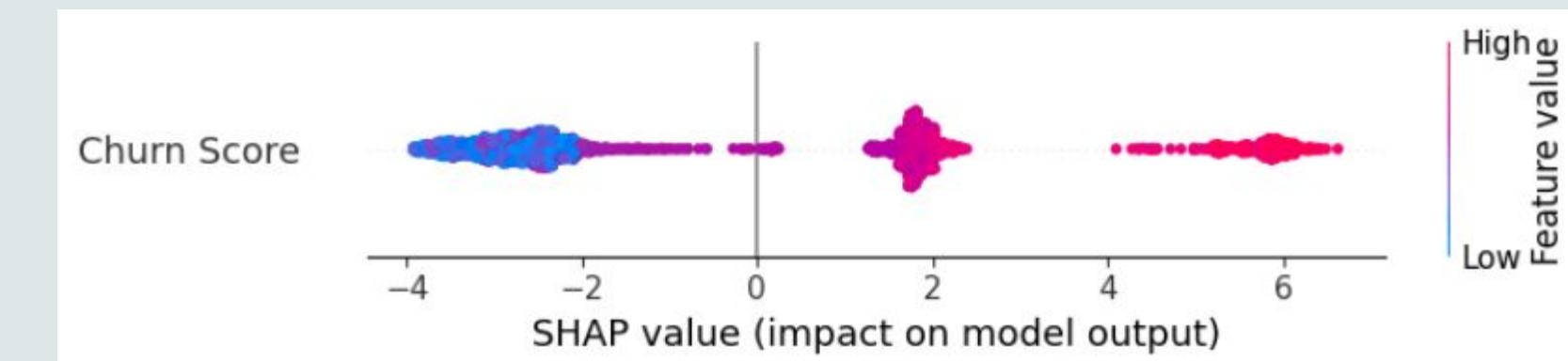


XAI - SHAP

CatBoost



각 모델, Scaler마다
변수의 중요도,
종속변수 상관관계 파악



최종 결론

- 높은 예측력으로 고객 이탈 사전 방지 가능
- XAI 이용, 이탈 사유 분석 가능
- 데이터 분석으로 고객의 Needs를 파악, 서비스 향상 기여
- 고객은 양질의 서비스를, 통신사는 고객 이탈 방지로 상호이익
- 해외 데이터이기 때문에, 국내 적용 가능 미지수

뉴스 기사

<https://www.edaily.co.kr/news/read?newsId=02994646635609248&mediaCodeNo=257&OutLnkChk=Y>

https://biz.chosun.com/it-science/ict/2022/04/20/2BIRWZY2PNBZJGOAA3FT7E7M2M/?utm_source=naver&utm_medium=original&utm_campaign=biz

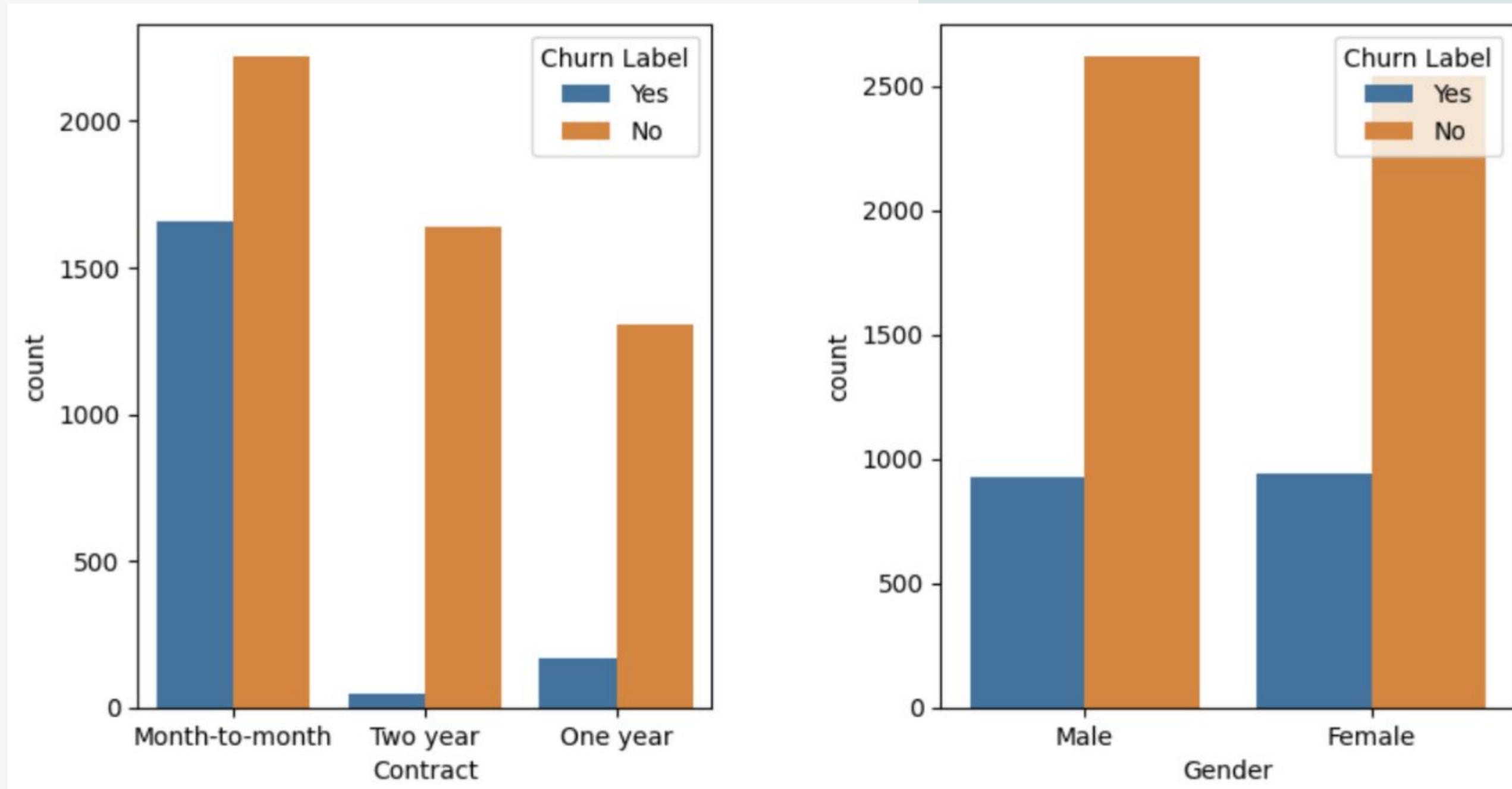
<https://www.yna.co.kr/view/AKR20220909033100017?input=1195m>

감사합니다

별첨

별첨 - EDA

Contract, Gender에 따른 이탈율

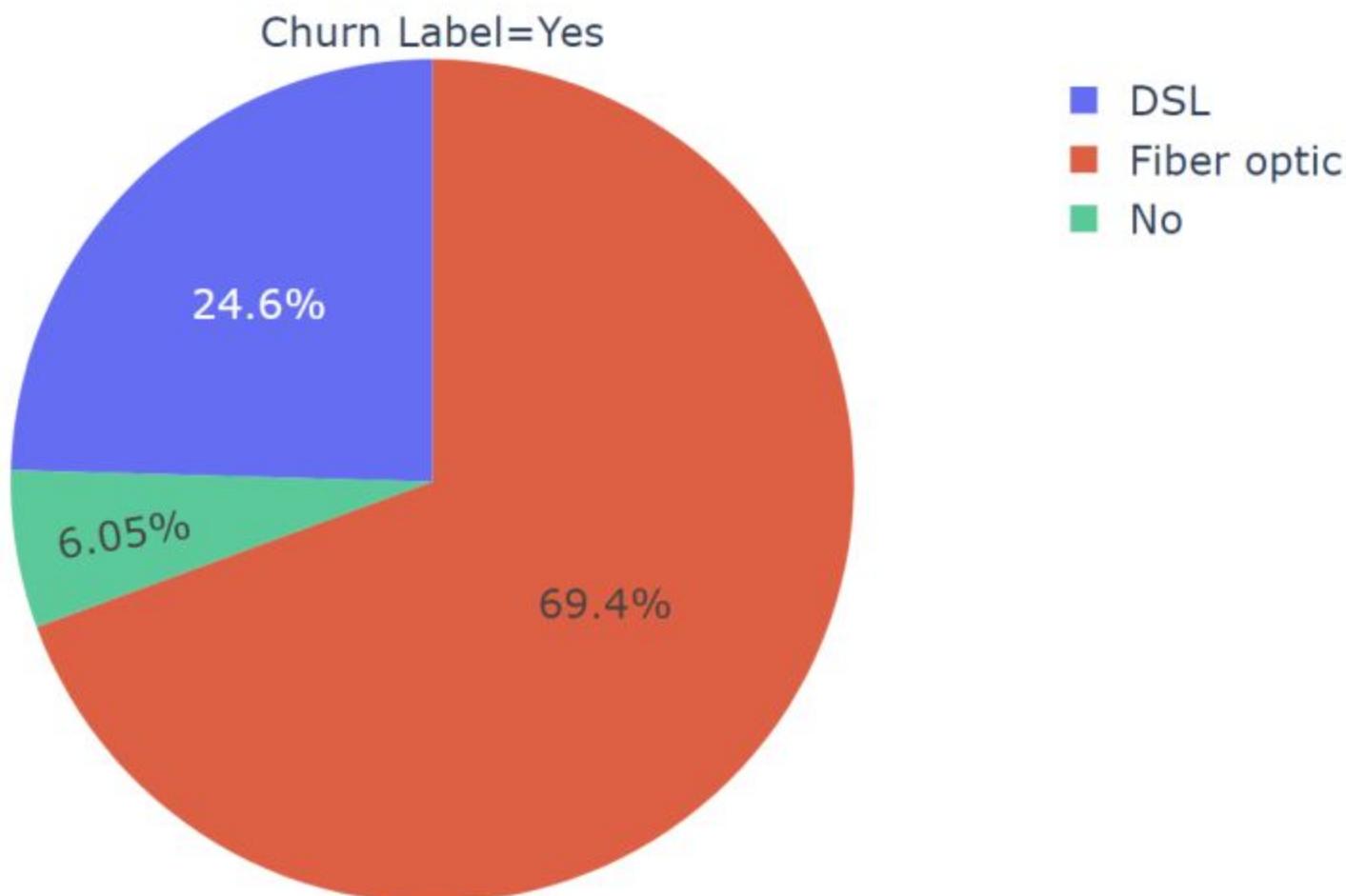
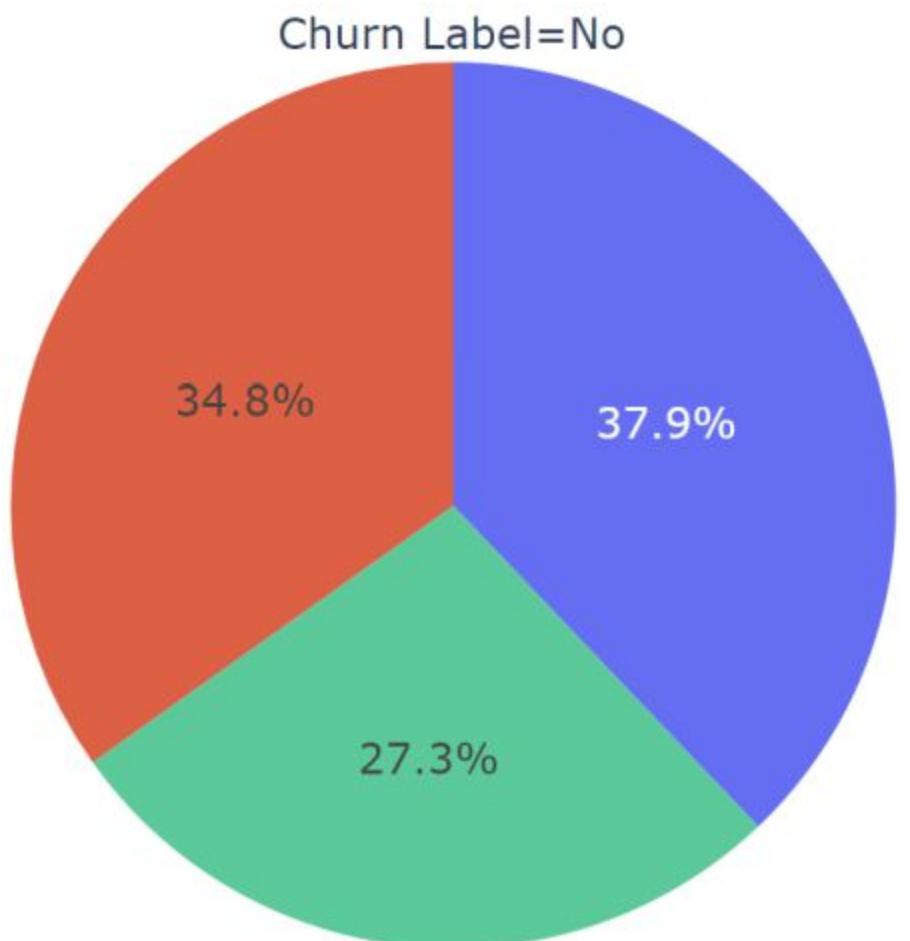


- Gender에 따른 이탈 변화 없음
- Contract에 따른 이탈 변화가 큼

별첨 - EDA

Internet Service에 따른 이탈율

Churn rate by Internet Service

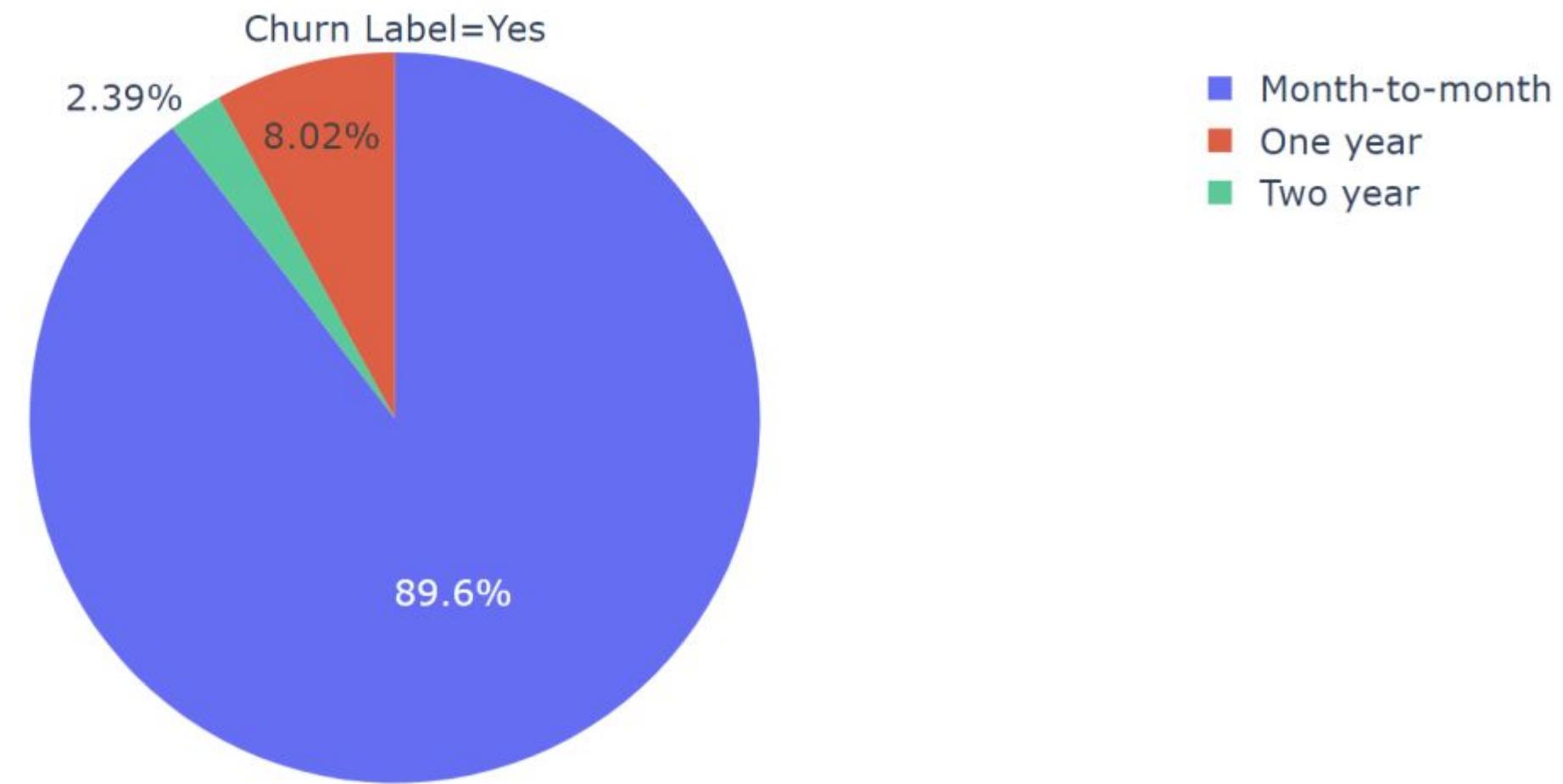


이탈 고객중
Fiber Optic 다수

별첨 - EDA

Fiber Optic 과 Contract 분석

fiber user by Contract



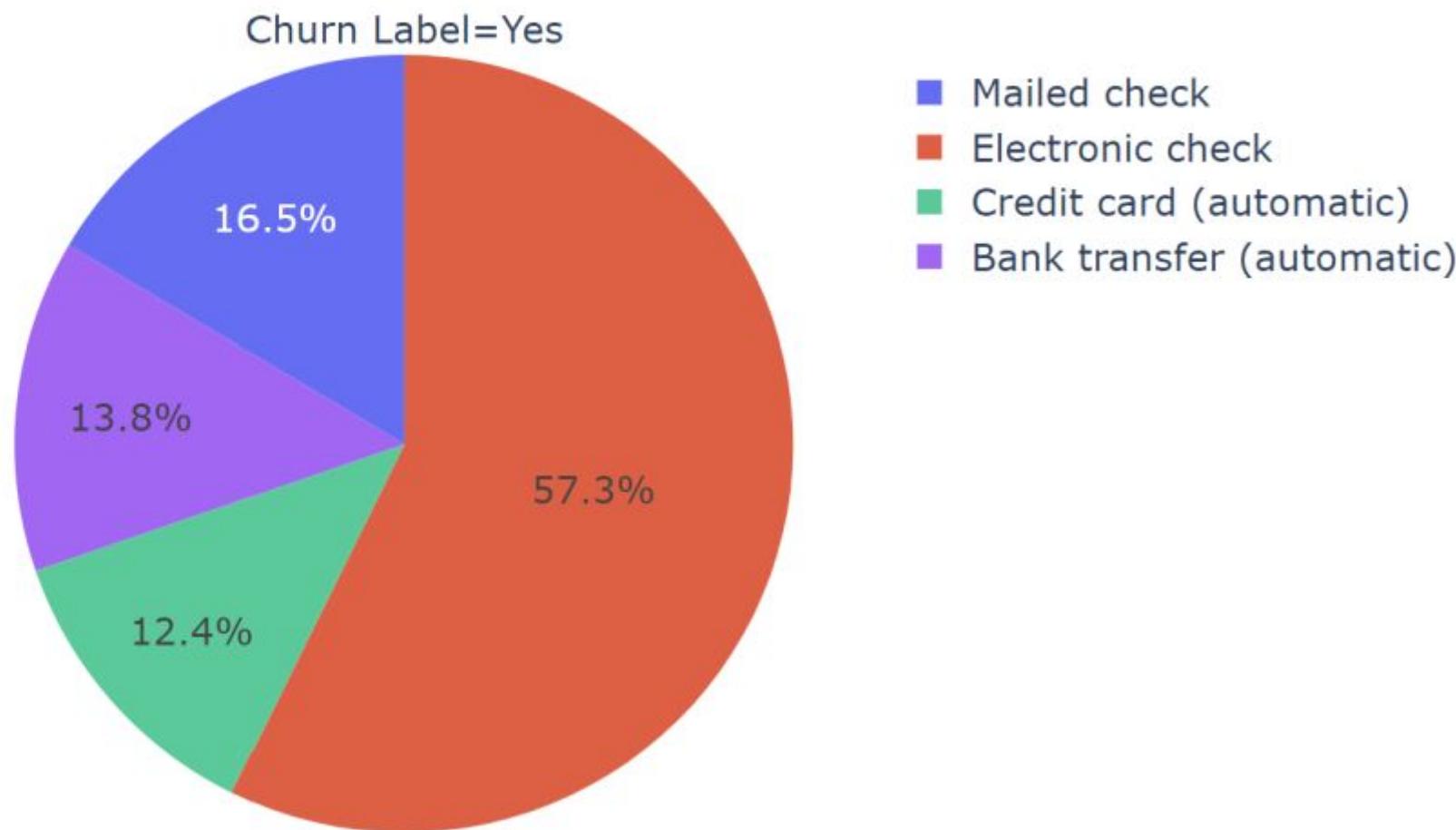
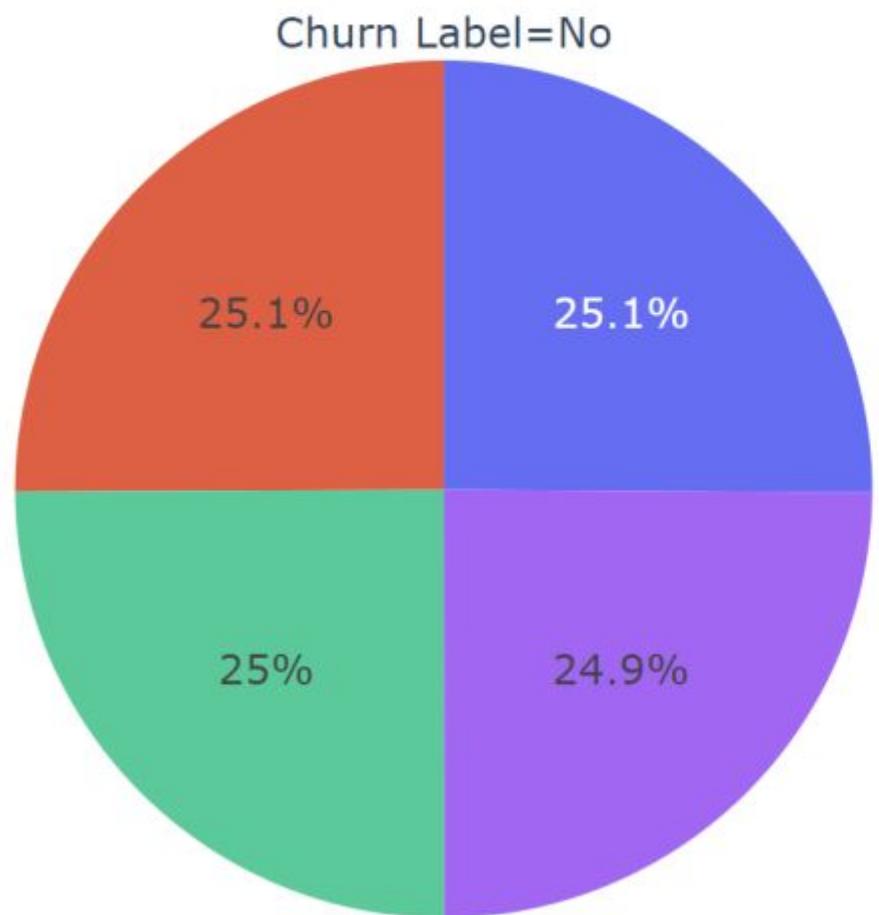
이탈 고객중
값에 따라 이탈율 차이가 컸던
Internet Service
(Fiber Optic),
Contract 분석

Fiber Optic
사용자중
Month to Month이
압도적으로 높음

별첨 - EDA

Fiber Optic 과 Payment Method 분석

fiber user by Payment Method



주요 변수로 파악되는
Fiber Optic과
Payment Method도 비교

이탈 고객 중
Electronic check이
앞도적으로 높음

별첨 - PCA

개인별로 차원을 정해 PCA 차원 축소 시도

- 차원을 축소하여 시간 단축 및 과적합 방지
- 주성분요소로 분석을 하여, 더 높은 정확도를 만들고자 시도
- 개인별 모델 학습 후, 겹치는 차원 제외
(개인별 의견, 기준 반영)

별첨 - PCA

PCA 축소 기여 변수 StandardScaler

```
Explained Variance up to Dimension 15: 0.92
Online Backup Contribution: 0.6301
Online Security Contribution: 0.3672
Device Protection Contribution: 0.3461
Senior Citizen Contribution: 0.2663
Dependents Contribution: 0.2060
Streaming TV Contribution: 0.2028
Partner Contribution: 0.1890
Multiple Lines Contribution: 0.1621
Payment Method_Mailed check Contribution: 0.1560
Contract_Month-to-month Contribution: 0.1491
Contract_Two year Contribution: 0.1339
Tech Support Contribution: 0.1225
Tenure Months Contribution: 0.0941
Streaming Movies Contribution: 0.0911
Internet Service_Fiber optic Contribution: 0.0889
Total Charges Contribution: 0.0884
Internet Service_No Contribution: 0.0876
Payment Method_Electronic check Contribution: 0.0675
Payment Method_Credit card (automatic) Contribution: 0.0450
Payment Method_Bank transfer (automatic) Contribution: 0.0435
Phone Service Contribution: 0.0421
Paperless Billing Contribution: 0.0339
Contract_One year Contribution: 0.0152
Monthly Charges Contribution: 0.0121
CLTV Contribution: 0.0056
Churn Score Contribution: 0.0047
Internet Service_DSL Contribution: 0.0013
```

개인별 PCA 분석 취합

노인, 인터넷 서비스, 계약, 지불방법에 대한 내용을 고루 포함

초반에 Churn reason에 포함된 서비스 불만족을
유추할 수 있는 Tech Support가 높게 반영

StandardScaler 15차원 축소

별첨 - PCA

개인별 PCA 분석 취합

StandardScaler - 15차원 결과

```
Final Test Results for GBM  
Accuracy: 0.95  
Recall: 0.89  
Precision: 0.90  
F1-Score: 0.90  
Confusion Matrix:  
[[1515  56]  
 [ 57 485]]
```

```
Final Test Results for XGBoost  
Accuracy: 0.94  
Recall: 0.89  
Precision: 0.90  
F1-Score: 0.89  
Confusion Matrix:  
[[1515  56]  
 [ 62 480]]
```

```
Final Test Results for LightGBM  
Accuracy: 0.95  
Recall: 0.89  
Precision: 0.90  
F1-Score: 0.89  
Confusion Matrix:  
[[1515  56]  
 [ 58 484]]
```

```
Final Test Results for RandomForest  
Accuracy: 0.94  
Recall: 0.86  
Precision: 0.90  
F1-Score: 0.88  
Confusion Matrix:  
[[1520  51]  
 [ 76 486]]
```

```
Final Test Results for CatBoost  
Accuracy: 0.93  
Recall: 0.86  
Precision: 0.86  
F1-Score: 0.86  
Confusion Matrix:  
[[1496  75]  
 [ 76 466]]
```

각 기본 파라미터로
GBM, XGBoost, LightGBM,
RandomForest, CatBoost
모델 학습
성능이 좋지 않음

별첨 - PCA

PCA 축소 기여 변수 MinMaxScaler

```
Explained Variance up to Dimension 19: 0.96
Internet Service_No Contribution: 0.5621
Internet Service_Fiber optic Contribution: 0.3681
Phone Service Contribution: 0.3215
Online Security Contribution: 0.2751
Contract_Month-to-month Contribution: 0.2684
Tech Support Contribution: 0.2381
Contract_Two year Contribution: 0.2377
Internet Service_DSL Contribution: 0.2040
Device Protection Contribution: 0.1657
Online Backup Contribution: 0.1639
Payment Method_Mailed check Contribution: 0.1285
Streaming Movies Contribution: 0.1183
Senior Citizen Contribution: 0.1152
Streaming TV Contribution: 0.1140
Monthly Charges Contribution: 0.1001
Payment Method_Electronic check Contribution: 0.0888
Paperless Billing Contribution: 0.0760
Churn Score Contribution: 0.0756
Dependents Contribution: 0.0531
CLTV Contribution: 0.0449
Tenure Months Contribution: 0.0433
Contract_One year Contribution: 0.0307
Payment Method_Bank transfer (automatic) Contribution: 0.0249
Partner Contribution: 0.0246
Payment Method_Credit card (automatic) Contribution: 0.0148
Gender Contribution: 0.0131
Total Charges Contribution: 0.0102
Multiple Lines Contribution: 0.0051
```

개인별 PCA 분석 취합

타 차원 대비
유의하게 생각한 변수를 가장 많이 포함

Internet Service, Contract, Payment Method

MinMaxScaler - 19차원 축소

별첨 - PCA

MinMaxScaler - 19차원 결과

개인별 PCA 분석 취합

```
Final Test Results for GBM  
Accuracy: 0.99  
Recall: 0.98  
Precision: 1.00  
F1-Score: 0.99  
Confusion Matrix:  
[[1570  1]  
 [ 12 530]]
```

```
Final Test Results for XGBoost  
Accuracy: 1.00  
Recall: 0.99  
Precision: 1.00  
F1-Score: 1.00  
Confusion Matrix:  
[[1571  0]  
 [  5 537]]
```

```
Final Test Results for LightGBM  
Accuracy: 1.00  
Recall: 0.99  
Precision: 1.00  
F1-Score: 0.99  
Confusion Matrix:  
[[1571  0]  
 [  6 536]]
```

```
Final Test Results for CatBoost  
Accuracy: 0.93  
Recall: 0.86  
Precision: 0.86  
F1-Score: 0.86  
Confusion Matrix:  
[[1496  75]  
 [ 76 466]]
```

```
Final Test Results for RandomForest  
Accuracy: 0.99  
Recall: 0.97  
Precision: 1.00  
F1-Score: 0.99  
Confusion Matrix:  
[[1571  0]  
 [ 14 528]]
```

지나치게 과적합이 됨

별첨 - PCA

PCA 축소 기여 변수 MinMaxScaler

Explained Variance up to Dimension 17: 0.94
Dependents Contribution: 0.8202
Partner Contribution: 0.3528
Senior Citizen Contribution: 0.3501
Churn Score Contribution: 0.1333
Tech Support Contribution: 0.1251
Contract_Two year Contribution: 0.1148
Contract_One year Contribution: 0.0753
Tenure Months Contribution: 0.0697
Payment Method_Mailed check Contribution: 0.0656
Device Protection Contribution: 0.0612
Contract_Month-to-month Contribution: 0.0546
Internet Service_No Contribution: 0.0484
Payment Method_Bank transfer (automatic) Contribution: 0.0461
Streaming Movies Contribution: 0.0386
Total Charges Contribution: 0.0379
Phone Service Contribution: 0.0317
Internet Service_DSL Contribution: 0.0304
CLTV Contribution: 0.0298
Streaming TV Contribution: 0.0277
Multiple Lines Contribution: 0.0181
Payment Method_Credit card (automatic) Contribution: 0.0161
Internet Service_Fiber optic Contribution: 0.0129
Paperless Billing Contribution: 0.0123
Payment Method_Electronic check Contribution: 0.0084
Online Security Contribution: 0.0083
Online Backup Contribution: 0.0050
Monthly Charges Contribution: 0.0014

개인별 PCA 분석 취합

**Senior Citizen, Contract 등
유의한 변수를 많이 포함하여 17차원으로 축소**

| | |
|--------------|--|
| GBM | 정확도: 0.8248, 정밀도: 0.8074, 재현율: 0.8529, F 스코어: 0.8289 |
| LightGBM | 정확도: 0.8653, 정밀도: 0.8501, 재현율: 0.8866, F 스코어: 0.8669 |
| XGBoost | 정확도: 0.8745, 정밀도: 0.8596, 재현율: 0.8953, F 스코어: 0.8757 |
| RandomForest | 정확도: 0.8669, 정밀도: 0.8598, 재현율: 0.8765, F 스코어: 0.8665 |

중요 성능 지표 - 원본 데이터보다 낮음

별첨 - PCA

PCA 축소 모델 성능 StandardScaler

standard 시 : 12차원

scalingName : Pca_smote_standard, modelName : Gradient Boost

- accuracy : 0.9584421910328708
- precision : 0.9503871620564851
- recall : 0.9665104446678043
- f1-score : 0.9583479758492667

scalingName : Pca_smote_standard, modelName : XGB

- accuracy : 0.9664499998093561
- precision : 0.9583626496100959
- recall : 0.9746047700303094
- f1-score : 0.9663782994495523

scalingName : Pca_smote_standard, modelName : LGBM

- accuracy : 0.9647935897875846
- precision : 0.9584813556450102
- recall : 0.9709777705055982
- f1-score : 0.9646423360030905

scalingName : Pca_smote_standard_randomforest, modelName : RandomForest

- accuracy : 0.9583048320617381
- precision : 0.9526053701440045
- recall : 0.9637202652267751
- f1-score : 0.9581081136683551

개인별 PCA 분석 취합

차원 선정 기준 - StandardScaler 12차원 축소

1. 선정한 중요 변수들에 대한 요소들이
상위권에 포진되어있는지에 따라서 지정

2. 적절한 설명력을 갖춘 차원에 대해서 선정

3. 모든 feature값들의 설명력이 높은건 불가능
어느 정도의 tradeOff를 하여 평균적으로 높은 차원을 지정

별첨 - PCA

PCA 축소 모델 성능

MinMaxScaler

minMax 시 : 13차원으로 선정

scalingName : Pca_minMax, modelName : Gradient Boost

- accuracy : 0.9431181344344928
- precision : 0.9305343205317126
- recall : 0.9564650974342193
- f1-score : 0.9432947956918774

scalingName : Pca_minMax, modelName : XGB

- accuracy : 0.9464320030198003
- precision : 0.9386068003807913
- recall : 0.9542359146973345
- f1-score : 0.9462834282338612

scalingName : Pca_minMax, modelName : LGBM

- accuracy : 0.9462935001849246
- precision : 0.937164598916767
- recall : 0.9556298356747155
- f1-score : 0.9462380388497476

scalingName : Pca_minMax_randomforest, modelName : RandomForest

- accuracy : 0.9465696479568686
- precision : 0.9329217816613282
- recall : 0.9612121424619964
- f1-score : 0.9467885624246513

개인별 PCA 분석 취합

MinMaxScaler 13차원 축소

StandardScaler과 MinMaxScaler 중

StandardScaler,

그중 LightGBM 에서 성능 우수

별첨 - PCA

PCA 축소 모델 성능

MinMaxScaler

```
Explained Variance up to Dimension 18: 0.96
Internet Service_No Contribution: 0.5359
Internet Service_Fiber optic Contribution: 0.3473
Phone Service Contribution: 0.3024
Contract_Month-to-month Contribution: 0.2940
Online Security Contribution: 0.2872
Contract_Two year Contribution: 0.2596
Tech Support Contribution: 0.2417
Internet Service_DSL Contribution: 0.1984
Dependents Contribution: 0.1853
Online Backup Contribution: 0.1454
Payment Method_Mailed check Contribution: 0.1444
Device Protection Contribution: 0.1192
Streaming Movies Contribution: 0.1166
Monthly Charges Contribution: 0.1061
CLTV Contribution: 0.0921
Streaming TV Contribution: 0.0880
Paperless Billing Contribution: 0.0874
Churn Score Contribution: 0.0794
Partner Contribution: 0.0740
Payment Method_Electronic check Contribution: 0.0700
Tenure Months Contribution: 0.0659
Payment Method_Bank transfer (automatic) Contribution: 0.0413
Senior Citizen Contribution: 0.0389
Payment Method_Credit card (automatic) Contribution: 0.0352
Total Charges Contribution: 0.0321
Contract_One year Contribution: 0.0221
Multiple Lines Contribution: 0.0175
```

개인별 PCA 분석 취합

MinMaxScaler 18차원 축소

대부분 변수 하나의 기여도가 지나치게 높은 것들이 많았는데, 18차원은 고르게 분포하여 선정

별첨 - PCA

PCA 축소 모델 성능
MinMaxScaler

개인별 PCA 분석 취합

MinMaxScaler 18차원

GBM

정확도: 0.8290, 정밀: 0.8118, 재현율: 0.8674, F1 스코어: 0.8387

XGBoost

정확도: 0.8709, 정밀: 0.8542, 재현율: 0.9019, F1 스코어: 0.8774

LightGBM

정확도: 0.8673, 정밀: 0.8528, 재현율: 0.8957, F1 스코어: 0.8737

RandomForest

정확도: 0.8641, 정밀: 0.8588, 재현율: 0.8793, F1 스코어: 0.8689

각 변수를 고르게 PCA 차원 축소를 하였지만,

성능은 좋지 않았음