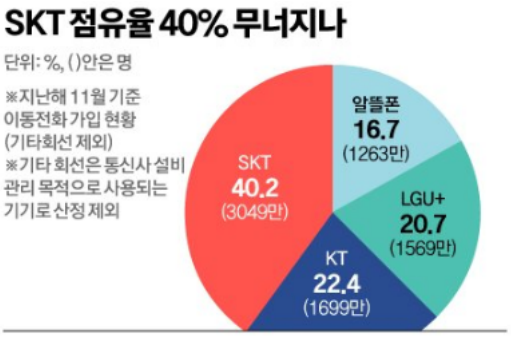
[SKN_2팀] 2차 프로젝트 결과 산출물 최종 보고서

주제: 이동 통신사 고객 이탈 예측 모델 구축 및 분석

이름	역할
김영현	EDA, Modeling, 전처리 전후 데이터의 후보 예측 모델의 성능 비교, 결과 산출물 보고서 정리
서종호(조장)	후보 예측 모델들의 하이퍼파라미터 튜닝 및 최종 모델 성능 개선, 결과 산출물 보고서 정리
장정원	EDA, 결과 산출물 보고서 정리, 발표 자료(PPT)
황양오	Data 설명, 후보 예측 모델의 장단점 분석 및 정리, 발표 자료(PPT)

1. 프로젝트 선정 배경 및 목표



2023년 1월 과기부 발표 [무선통신 서비스 가입 현황 통계] 참조

출처: https://www.joongang.co.kr/article/25132261

통신 3사의 기존 고객이 알뜰폰으로 갈아타는 '환승족'이 늘어나는 추세입니다. SKT 기준 41.3%(21년 11월) → 40.2%(22년 11월)으로 점유율 감소하였고, 알뜰폰 기준 14.1%(21년 11월) → 16.7%(22년 11월)로 급성장하였습니다. 가입자 확보 중심의 성장 전략은 포화상태로 한계가 있어 기존 가입자를 바탕으로 통신에 멤버쉽, 콘텐츠 경험, 서비스 등의 차별화된 편익에 초점을 맞춰 알뜰폰에 대응하는 분위기입니다.

이에 따른 통신사 고객의 이탈 관리 방안에 대한 분석 필요성을 느꼈습니다. 하지만 국내 통신사의 고객 데이터를 확보하는데 어려움이 있고, 미국 통신업계의 포화상태 및 AT&T, Verizon, T-Mobile 3사의 시장 점유, 고객 이탈 등을 고려했을 때, 한국의 상황과 유사하다고 판

단하여 Kaggle의 Telco Customer Churn Dataset을 분석하여 insight를 확보하고자 하였습니다.

2. 데이터 설명

- 데이터 설명 및 컬럼 정보

출처	Kaggle
Link	https://www.kaggle.com/datasets/blastchar/telco-customer-churn
Dataset Name	Telco Customer Churn
Total row	7043
Total column	21
Dataset 정보	- 캘리포니아에서 7,043명의 고객에게 집 전화 및 인터넷 서비스를 제공한 가상의 통신사에 대한 정보가 포함된 데이터셋 - 어떤 고객이 서비스를 떠났고, 남아있거나, 서비스에 가입했는지 확인
Column 정보	고객이 가입한 서비스 : 전화, 여러 회선, 인터넷, 온라인 보안, 온라인 백업, 기기 보호, 기술 지원, 스트리밍 TV 및 영화 등의 서비스 고객 계정 정보 : 고객 가입 기간, 계약, 결제 방법, 종이 없는 청구서, 월별 요금 및 총 요금 고객에 대한 인구통계학적 정보: 성별, 연령대, 파트너 및 부양가족이 있는지 여부

변수명	Type	결측치	Unique Value	설명	비고
customerID	index	X	aaaa-bbbbb	고객별 고유 ID	a : number, b : string
gender	Categorical	X	Male, Female	고객이 남성인지 여성인지 여부	
SeniorCitizen	Numerical	X	1, 0	고객이 65세 이상인지 여부	1 : senior
Partner	Categorical	X	Yes, No	고객의 배우자 유무	
Dependents	Categorical	X	Yes, No	고객의 부양가족(자녀, 부모, 조부모) 유무	
tenure	Numerical	X	Numerical	고객의 서비스 사용 개월 수	
PhoneService	Categorical	X	Yes, No	고객의 전화 서비스를 사용 여부	
MultipleLines	Categorical	X	Yes, No, No Phone service	고객의 다수의 전화 회선 사용 여부	Phone service 하위 option
InternetService	Categorical	X	DSL, Fiber optic, No	고객이 사용하는 인터넷 서비스	
OnlineSecurity	Categorical	X	Yes, No, No internet service	고객이 인터넷 보안 서비스 가입 여부	internet service 하위option
OnlineBackup	Categorical	X	Yes, No, No internet service	고객의 인터넷 백업 서비스 가입 여부	internet service 하위option
DeviceProtection	Categorical	X	Yes, No, No internet service	고객의 인터넷 장치 보호 서비스 가입 여부	internet service 하위option
TechSupport	Categorical	X	Yes, No, No internet service	고객의 인터넷 기술 지원 서비스 가입 여부	internet service 하위option
StreamingTV	Categorical	X	Yes, No, No internet service	고객의 TV 스트리밍 서비스 가입 여부	internet service 하위option
StreamingMovies	Categorical	X	Yes, No, No internet service	고객의 영화 스트리밍 서비스 가입 여부	internet service 하위option
Contract	Categorical	X	Month-to-month, One year, Two year	고객이 계약한 서비스 기간	
PaperlessBilling	Categorical	X	Yes, No	고객의 전자청구서 선택 여부	
PaymentMethod	Categorical	X	Electronic check, Mailed check, Bank transfer(automatic), Credit card(automatic)	고객의 결제수단	
MonthlyCharges	Numerical	X	Numerical	고객이 매월 지불하는 금액	
TotalCharges	Numerical	O	Numerical	고객이 서비스 가입 후 지불한 총 금액	
Churn	Categorical	X	Yes, No	고객 이탈 여부	

3. 데이터 로드 및 확인

데이터 확인

```
# 데이터 및 컬럼 개수 확인
print(df.shape)
display(df.head())
```

✓ 0.0s Python

```
(7043, 21)
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

5 rows × 21 columns

```
df.info()
```

```
<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
```

- 총 7,043개의 데이터와 21개의 컬럼이 존재하며, 데이터 형태와 맞지 않는 데이터 타입으로 되어 있어 전처리가 필요함을 알 수 있습니다
    - ‘SeniorCitizen’의 object형 변환, ‘TotalCharges’의 float형 변환

데이터 타입 변경 및 결측치 처리

```
df['SeniorCitizen'] = df['SeniorCitizen'].astype('object')
df['TotalCharges'] = pd.to_numeric(df.TotalCharges, errors='coerce')
df.isnull().sum() # 결측치 11개 존재
```

```
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
```

```
df[df['TotalCharges'].isnull()]['tenure'].value_counts()
```

✓ 0.0s

```
0    11
Name: tenure, dtype: int64
```

```
# 결측치 처리
df['TotalCharges'] = df['TotalCharges'].fillna(0)
df.isnull().sum()
```

✓ 0.0s

```
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        0
Churn               0
dtype: int64
```

- 범주형인 'SeniorCitizen' 변수와 연속형인 'TotalCharges' 변수의 타입 변경 후 결측치 확인
  - 'TotalCharges' 변수는 11개의 결측치가 존재하며 이 결측치는 이용 개월 수가 0인 고객들에 해당됩니다. 따라서 결측치를 0으로 처리하였습니다.

연속형 데이터 요약 통계량

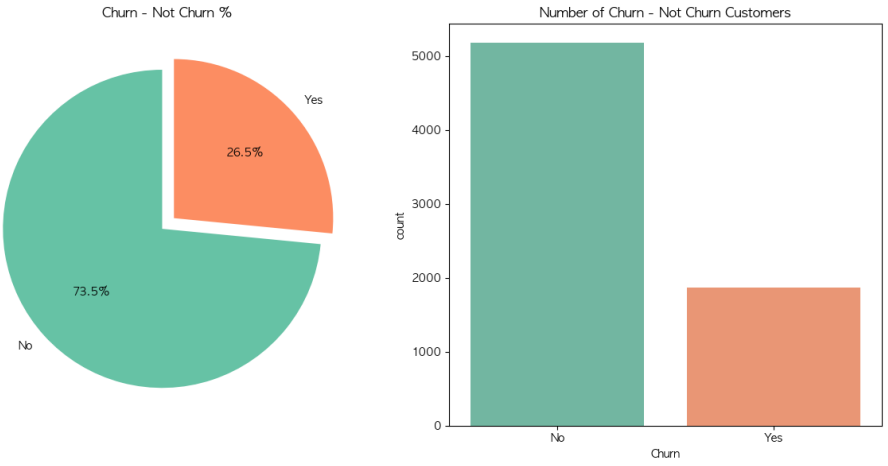
```
# 요약 통계량
df.describe()
```

	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000
mean	32.371149	64.761692	2279.734304
std	24.559481	30.090047	2266.794470
min	0.000000	18.250000	0.000000
25%	9.000000	35.500000	398.550000
50%	29.000000	70.350000	1394.550000
75%	55.000000	89.850000	3786.600000
max	72.000000	118.750000	8684.800000

- 이용 개월 수를 나타내는 'tenure', 월별 요금을 나타내는 'MonthlyCharges', 가입 후 총 요금을 나타내는 'TotalCharges'에 대한 요약 통계량
  - 이용 개월 수가 0인 고객들은 'TotalCharges'의 최소 값인 0으로 채워짐을 확인했습니다.
  - 월별 요금의 범위는 최소 18.25 달러에서 최대 118.75 달러로 큰 차이가 있음을 확인할 수 있습니다.
  - 가입 후 총 요금 역시 이용 개월 수와 월별 요금에 따라 고객들 간의 차이가 크다는 것을 알 수 있습니다.
- 고객의 이용 개월 수와 요금 지불 패턴이 이탈 예측에 큰 영향을 끼치는 변수인지 추가 확인이 필요하다고 판단했습니다.

4. EDA 및 결과 분석

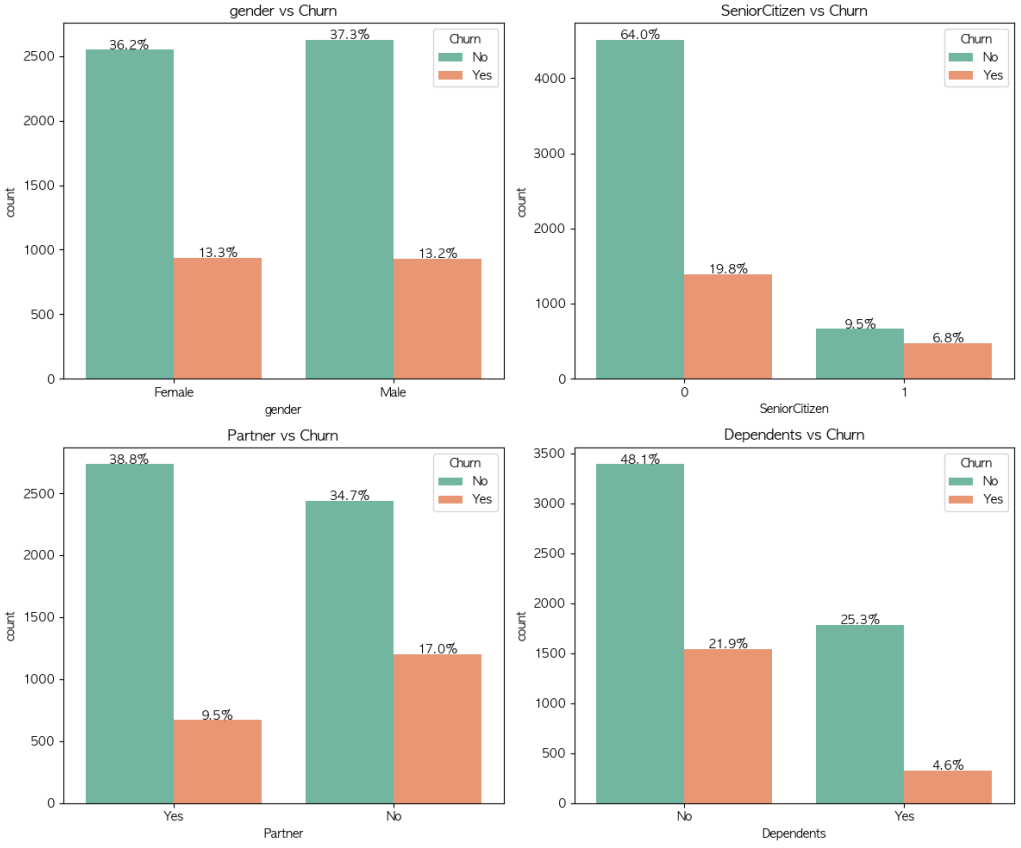
4-1. Target 변수('Churn') 시각화



- 파이 그래프와 막대 그래프를 통한 고객 이탈 여부(Yes/No) 시각화
- 이탈하지 않은 고객과 이탈한 고객의 수가 약 3:1의 비율로, **클래스 불균형(class imbalance) 상태**임을 알 수 있습니다.

4-2. Categorical 변수 시각화

> **Group 1: 고객 정보**
> **Gender | SeniorCitizen | Partner | Dependents**



- **고객 정보 관련 변수들과 'Churn' 변수의 관계 시각화**
- **가설**
  - 성별은 이탈 여부에 큰 영향이 없을 것

- 고령 고객일수록 이탈하지 않을 확률이 높을 것
- 배우자가 있으면 이탈할 확률이 높을 것
- 부양 가족이 있으면 이탈할 확률이 높을 것

- 시각화 결과

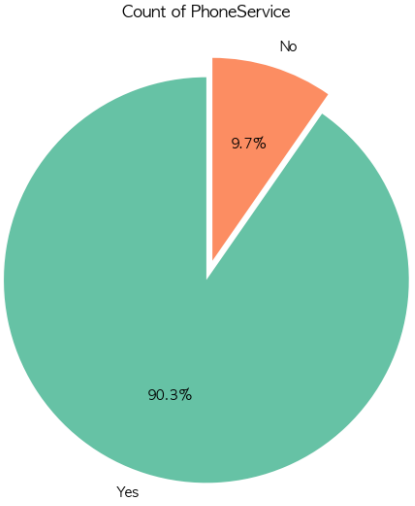
  1) 성별은 이탈을 예측하는 지표가 아님이 확인되었습니다.

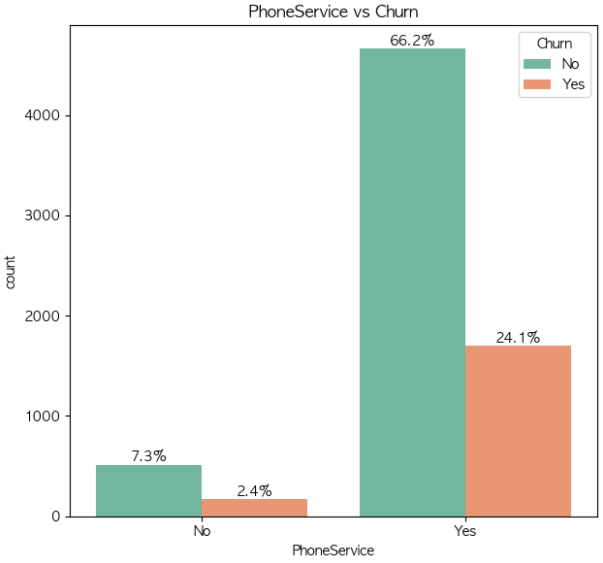
  2) 고령 고객은 전체 고객의 16%로 소수에 해당되지만, 비고령 고객과 비교했을 때 고령 고객은 훨씬 높은 이탈율을 보입니다.

  3) 배우자가 없는 경우 이탈 가능성이 더 높습니다.

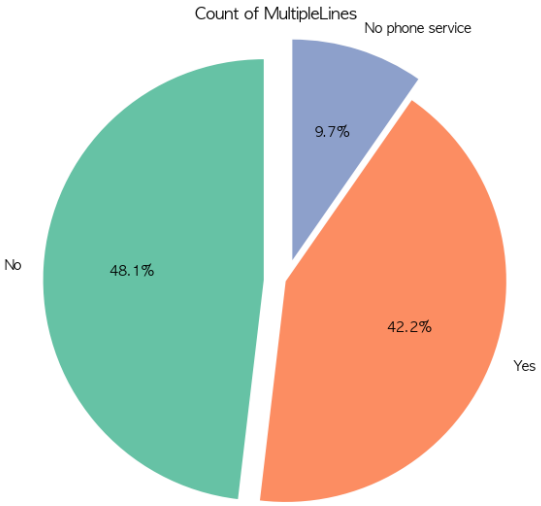
  4) 부양 가족이 없는 경우 이탈 가능성이 더 높습니다.

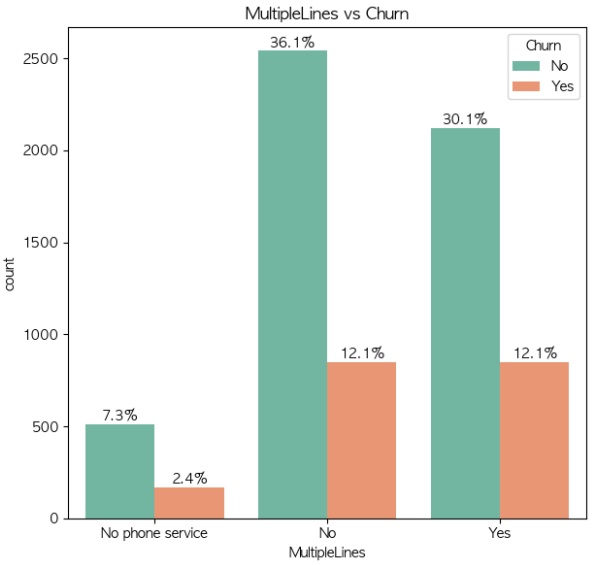
Group 2: 전화 및 인터넷 서비스 정보
PhoneService | MultipleLines | InternetService



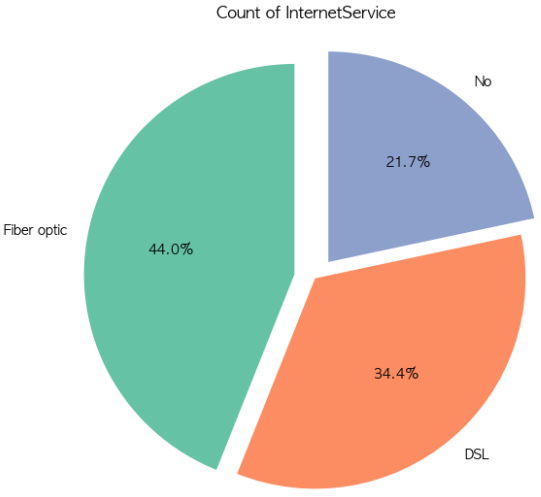


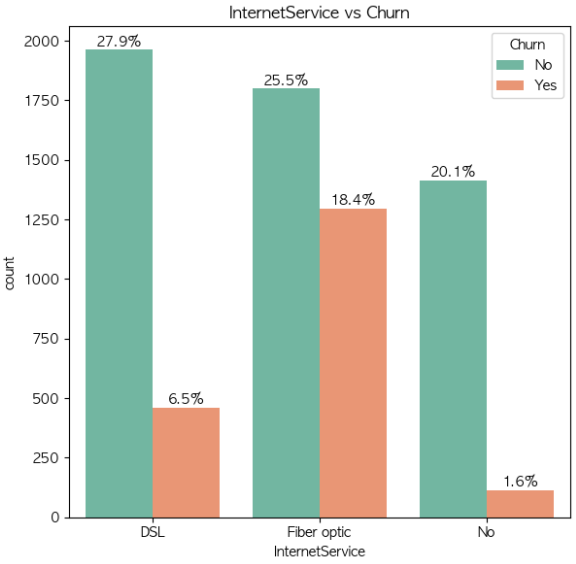
- **통신 서비스 사용 유무 (Yes/No) 와 'Churn'과의 관계 시각화**
- **가설**
  - 통신 서비스 사용 유무는 이탈 여부에 큰 영향이 없을 것
- **시각화 결과**
  - 통신 서비스를 이용하는 고객의 비율이 매우 높음을 확인했습니다.
  - 통신 서비스를 이용하지 않은 고객의 이탈 비율이 통신 서비스를 이용하는 고객보다 상대적으로 더 낮지만, 통신 서비스를 이용하는 고객의 이탈하지 않은 비율가 압도적으로 높아 큰 상관관계는 없을 것 이라고 판단됩니다.





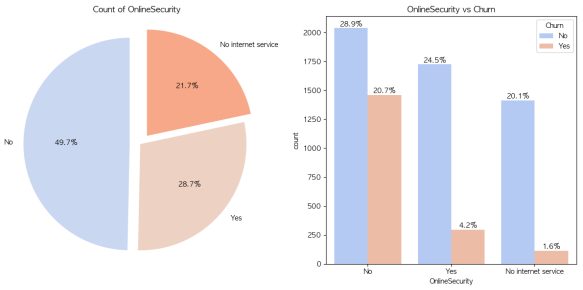
- **다수의 전화 회선 사용 유무 (Yes/No/No phone service) 와 'Churn'과의 관계 시각화**
- 가설
  - 다수의 전화 회선을 사용할 경우, 이탈이 더 번거로울 것으로 생각하여 이탈 확률이 더 적을 것
- 시각화 결과
  - 다수의 전화 회선을 사용하지 않는 고객들의 이탈하지 않는 비율이 조금 더 높으나, 큰 상관관계는 없는 것으로 판단하였습니다.



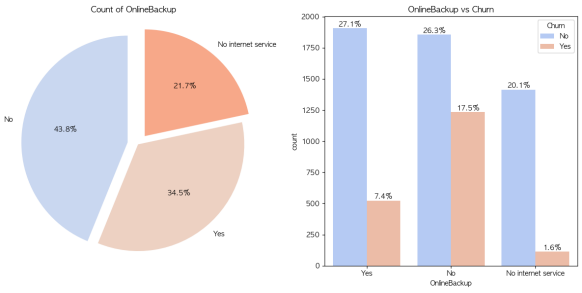


- 고객이 사용하는 인터넷 서비스와 'Churn'과의 관계 시각화
- 시각화 결과
  - Fiber optic에 비해 더 느리고 비용이 더 높은 DSL을 사용하는 고객이 이탈 가능성이 더 낮은 흥미로운 결과를 보았습니다.

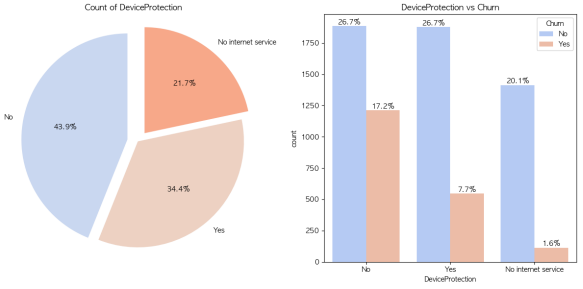
> **Group 3: 고객이 가입한 서비스 정보**
> **Online Security | Online Backup | DeviceProtection | TechSupport | StreamingTV | StreamingMovies**



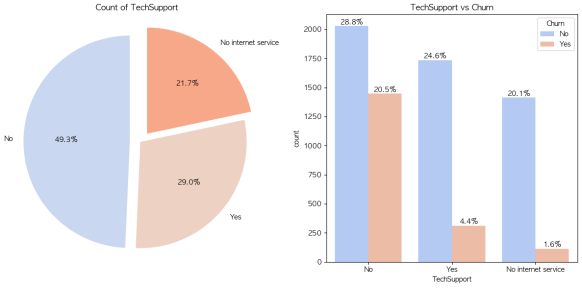
- **인터넷 보안 서비스 유무 (Yes/No/No phone service)와 'Churn'과의 관계 시각화**
- 가설
  - 부가 서비스를 많이 가입한 고객일수록 이탈 확률이 낮을 것
- 시각화 결과
  - 인터넷 보안 서비스를 이용하는 고객의 이탈 비율이 더 낮습니다.



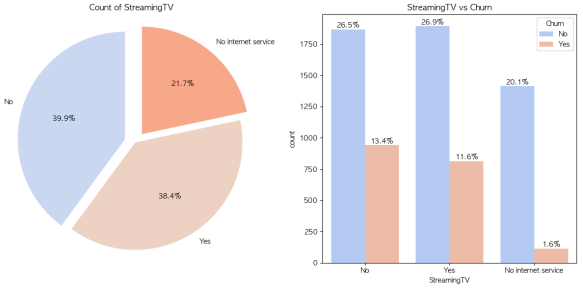
- **인터넷 백업 서비스 유무 (Yes/No/No phone service)와 'Churn'과의 관계 시각화**
- 가설
  - 부가 서비스를 많이 가입한 고객일수록 이탈 확률이 낮을 것
- 시각화 결과
  - 인터넷 백업 서비스를 이용하는 고객의 이탈 비율이 더 낮습니다.



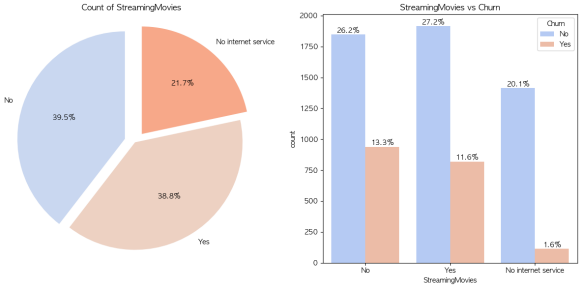
- **디바이스 보호 서비스 유무 (Yes/No/No phone service)와 'Churn'과의 관계 시각화**
- 가설
  - 부가 서비스를 많이 가입한 고객일수록 이탈 확률이 낮을 것
- 시각화 결과
  - 디바이스 보호 서비스를 이용하는 고객의 이탈 비율이 더 낮습니다.



- **기술 지원 서비스 유무 (Yes/No/No phone service)와 'Churn'과의 관계 시각화**
- 가설
  - 부가 서비스를 많이 가입한 고객일수록 이탈 확률이 낮을 것
- 시각화 결과
  - 기술 지원 서비스를 이용하는 고객의 이탈 비율이 더 낮습니다.



- TV 스트리밍 서비스 유무 (Yes/No/No phone service) 와 Churn과의 관계 시각화
- 가설



- 영화 스트리밍 서비스 유무 (Yes/No/No phone service) 와 Churn과의 관계 시각화
- 가설

- 부가 서비스를 많이 가입한 고객일수록 이탈 확률이 낮을 것

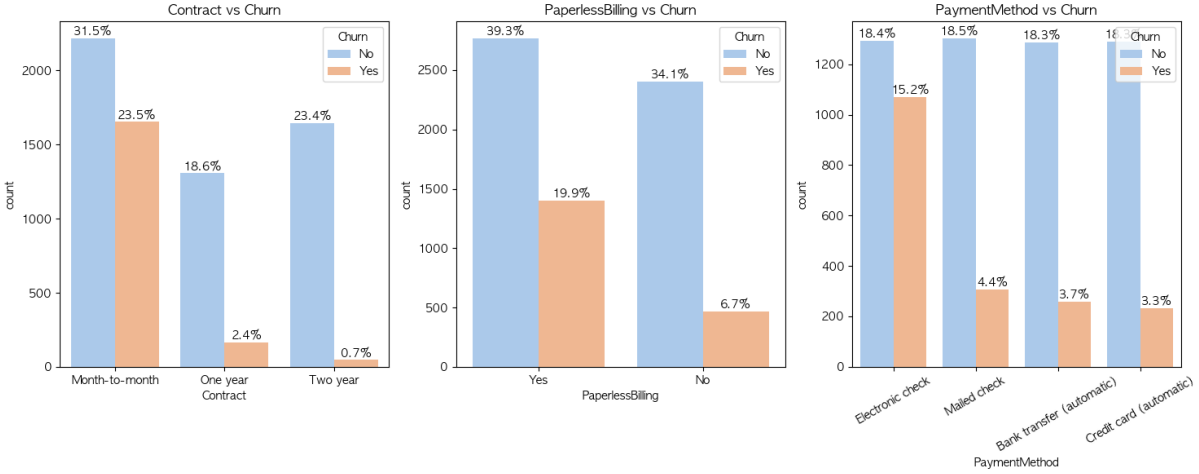
- 시각화 결과
  - TV 스트리밍 서비스 유무와 이탈률은 상관 관계가 없었습니다.

- 부가 서비스를 많이 가입한 고객일수록 이탈 확률이 낮을 것

- 시각화 결과
  - 영화 스트리밍 서비스 유무와 이탈률은 상관 관계가 없었습니다.

위 6개 서비스 칼럼에 대한 시각화를 바탕으로, 서비스를 이용하지 않는 고객들이 이용하는 고객보다 이탈율이 더 높은 것을 바탕으로 해당 통신사 기업에서는 더 많은 고객들이 서비스를 이용하도록 홍보 및 이벤트를 통해 장려할 필요가 있음을 알 수 있습니다.

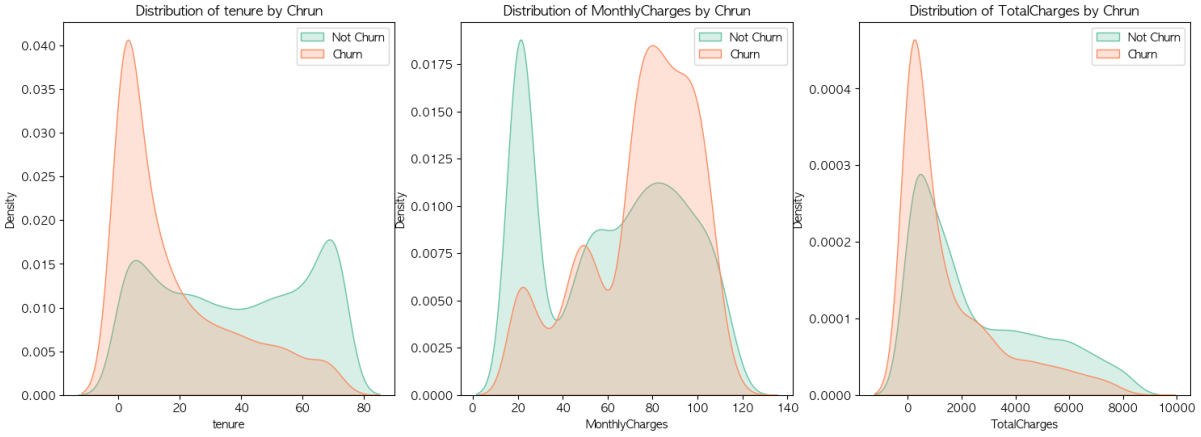
Group 4: 계약 기간 및 결제 정보
Contract | PaperlessBilling | PaymentMethod



- **계약 기간, 전자 청구서 여부, 결제 수단과 'Churn'의 관계 시각화**
- 가설
  - 계약 기간이 길수록 이탈 확률이 적을 것
  - 자동 결제가 아닌 수단으로 결제하는 고객의 이탈 확률이 더 높을 것
- 시각화 결과
  - 계약 기간이 짧을수록 이탈 확률이 높았습니다.
  - 전자 청구서를 받는 고객들의 이탈 확률이 더 높았습니다.
  - 전자 수표로 결제하는 고객들의 이탈 확률이 매우 높았습니다.

4-3. Numerical 변수 시각화

- **연속형 변수 'tenure', 'MonthlyCharges', 'TotalCharges'의 시각화**
- 시각화 결과
  - 세 변수 모두 이상치가 없었습니다.
  - TotalCharges는 양의 왜도 분포를 보여주었습니다.

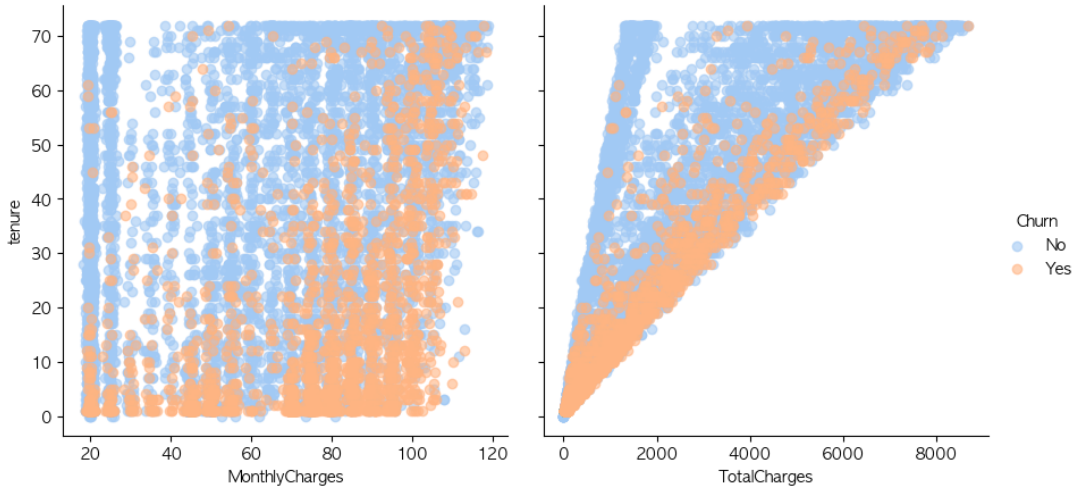


- **서비스 이용 기간, 월 비용, 누적 비용과 'Churn'의 관계 시각화**
- 가설
  - 장기 고객일수록 이탈률이 적을 것
  - 월 비용이 높을수록 이탈률이 높을 것
  - 누적 비용이 높을수록 이탈률이 적을 것 (장기 고객)
- 시각화 결과

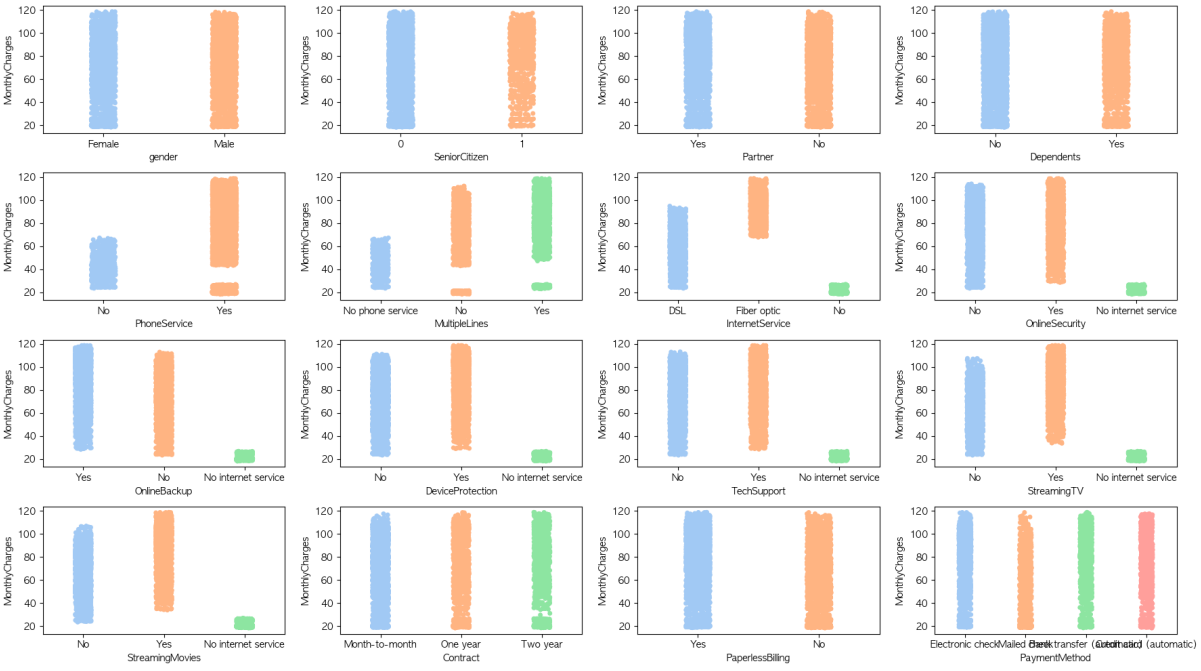
  1) 장기 고객일수록 이탈률이 적고, 신규 고객의 이탈률이 압도적으로 높았습니다.

  2) 월 비용이 낮을수록 이탈률이 낮고, 높을수록 이탈률이 높았습니다.

  3) 누적 비용이 낮을때 이탈률이 비이탈률 보다 상대적으로 더 높았습니다.

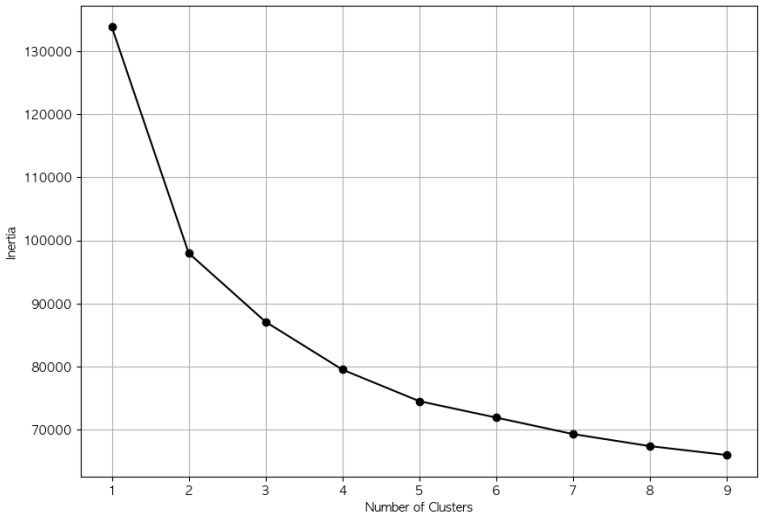


- 월 비용, 누적 비용과 서비스 이용 기간의 관계 시각화
- 월 비용과 서비스 이용 기간은 낮은 상관관계를 보였습니다.
- 총 비용과 서비스 이용 기간은 정비례했습니다. 이탈한 고객들은 이탈하지 않은 고객들에 비해 거의 선형관계를 보입니다.

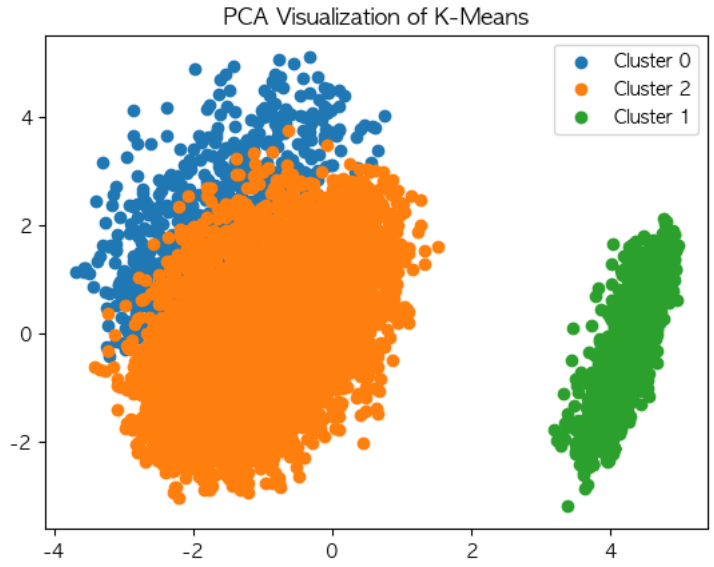


- PhoneService와 InternetService 변수들이 월별 금액에 가장 큰 영향을 끼침을 확인 가능하며, InternetService같은 경우는 광통신을 사용하는 고객의 평균 월별 요금이 더 높았습니다.
- 인터넷 보안, 온라인 백업, 디바이스 보호, 기술 지원, TV 스트리밍, 영화 스트리밍 6개의 인터넷 부가 서비스의 사용 유무에 따른 월별 요금이 그래프에서는 큰 차이를 보이지 않았지만, 다수의 서비스를 이용하면 더해지는 값도 늘어날것 이므로 부가 서비스가 월별 요금에 영향을 끼치지 않는다는 결론은 내리지 않았습니다.

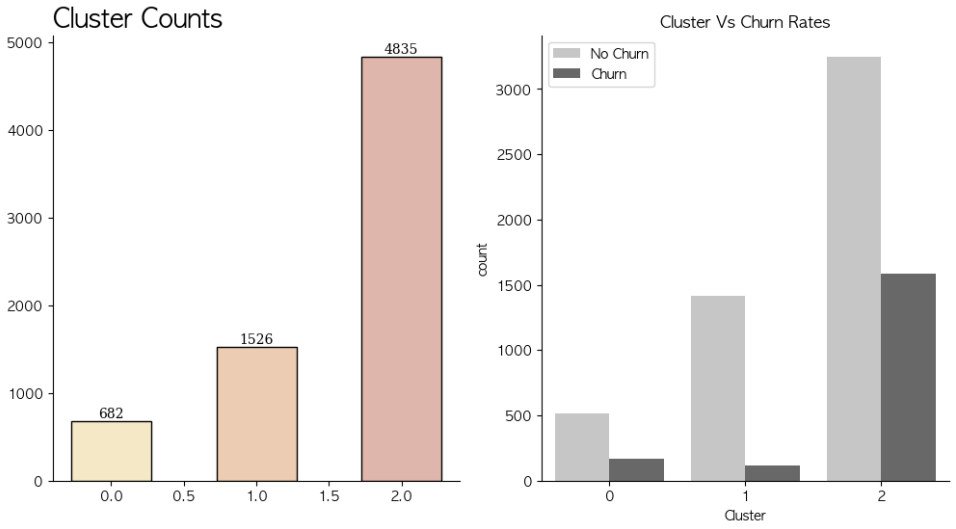
4-4. K-Means 클러스터링 및 PCA를 이용한 시각화



- 'customerID'와 'Churn' 변수를 제외한 데이터의 정규화 후 k-means 알고리즘을 시행했습니다.
- 최적의 k 값을 찾기 위해 이너셔를 이용한 엘보우 방법론을 수행한 결과, k=3 일 때 최적의 클러스터 개수라고 판단했습니다.



- k=3의 클러스터링 적용 후 PCA를 이용해 시각화한 결과, 클러스터 0과 2가 살짝 겹치긴 하지만 클러스터 1과 뚜렷이 구분되는 형태를 볼 수 있습니다.



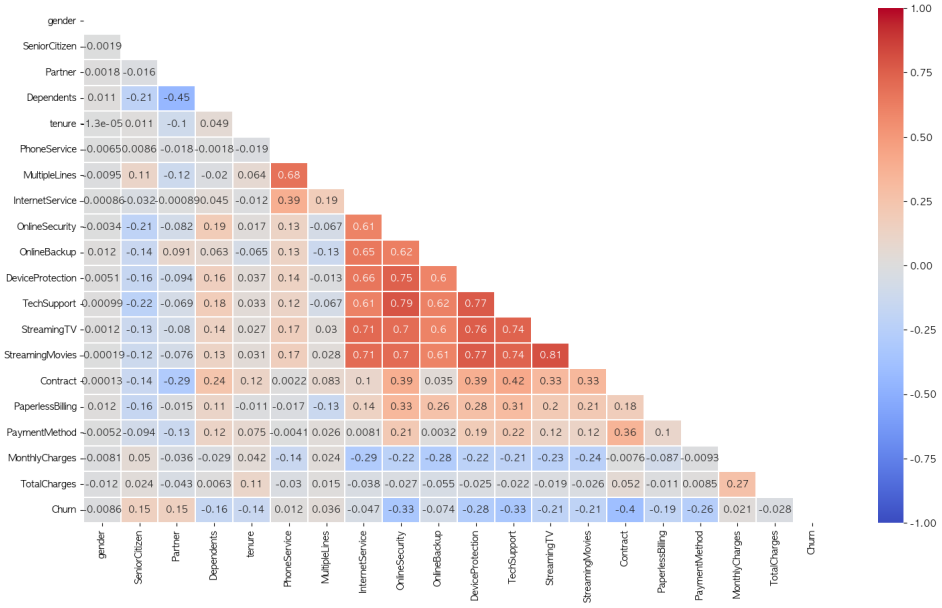
- 클러스터별 데이터 개수와 이탈 여부 변수 'Churn'의 카운트 그래프를 분석한 결과,

  클러스터 0과 1에 비해 클러스터 2에서 이탈한 고객들이 더 많이 나타났습니다.

- 클러스터링 분석을 통해 도출된 클러스터들이 고객 이탈 예측에 중요한 요소로 활용될 수 있음을 시사합니다. 이를 추가적인 파생 변수로 활용하여 모델 성능을 개선하는 데 유용할 것으로 판단합니다. (모델링 완료 후에 시행한 EDA)

4-5. 상관관계 분석 및 예측 모델링을 위한 추가 전처리

상관관계 분석



- **타겟 변수 'Churn' 분석**
  - 'Churn' 변수는 대부분의 다른 변수들과 음의 상관관계를 보입니다. 그 중에서도 상대적으로 상관관계가 높은 변수들은 'Contract', 'OnlineSecurity', 'TechSupport', 'DeviceProtection' 등이 있습니다.
  - 'Contract' 변수는 고객의 계약 기간을 의미하며, 이를 통해 계약 기간이 짧을수록 고객 이탈 확률이 높아진다는 것을 알 수 있습니다.
  - 'OnlineSecurity', 'TechSupport', 'DeviceProtection' 등의 부가 서비스를 이용하지 않는 고객일수록 이탈 확률이 높다는 것을 확인 가능합니다.

- **부가 서비스 간의 상관관계**
  - 'InternetService' 변수부터 'StreamingTV' 변수까지는 양의 상관관계가 매우 높음을 알 수 있습니다. 이는 한 가지 부가 서비스를 이용하는 고객들이 다른 부가 서비스도 함께 이용하는 경향이 있음을 보입니다.

파생 변수 및 더미 변수 생성

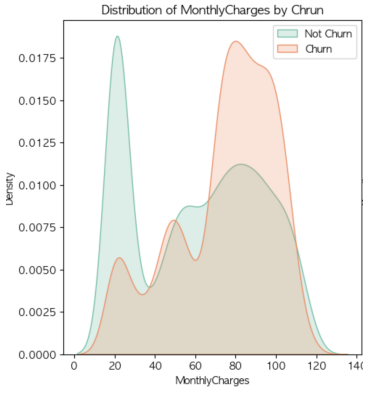
1. **통신사 이용 개월 수를 나타내는 'tenure'의 파생 변수**

```
1 df['tenure'].describe()
✓ 0.0s
count    7043.000000
mean       32.371149
std        24.559481
min         0.000000
25%         9.000000
50%        29.000000
75%        55.000000
max        72.000000
Name: tenure, dtype: float64
```

```python
def create_tenure_bin(x):
    if x < 24:
        return 'short'
    elif x < 48:
        return 'mid'
    else:
        return 'long'
```

- 최소 0개월, 최대 72개월의 이용 개월 수에 따라 0개월부터 24개월씩 'short', 'mid', 'long'의 세 구간으로 나누어 범주형 변수를 생성했습니다.

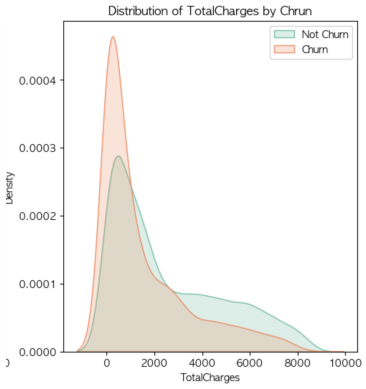
2. **월별 요금을 나타내는 'MonthlyCharges'의 파생 변수**




```python
def create_monthly_charge_group(x):
    # low mid high very high
    if x < 30:
        return 'low'
    elif x < 60:
        return 'mid'
    elif x < 90:
        return 'high'
    else:
        return 'very high'
```

- 위 시각화 결과를 바탕으로, 고객 이탈 ('Churn') 유무를 구분하는 범위에 따라 'low', 'mid', 'high', 'very high'의 4개의 범주형 변수를 생성했습니다.

3. **서비스 가입 후 지불한 총 금액을 나타내는 'TotalCharges'의 파생 변수**




```python
def create_Total_charge_group(x):
    if x < 3000:
        return 'low'
    elif x < 6000:
        return 'mid'
    else:
        return 'high'
```

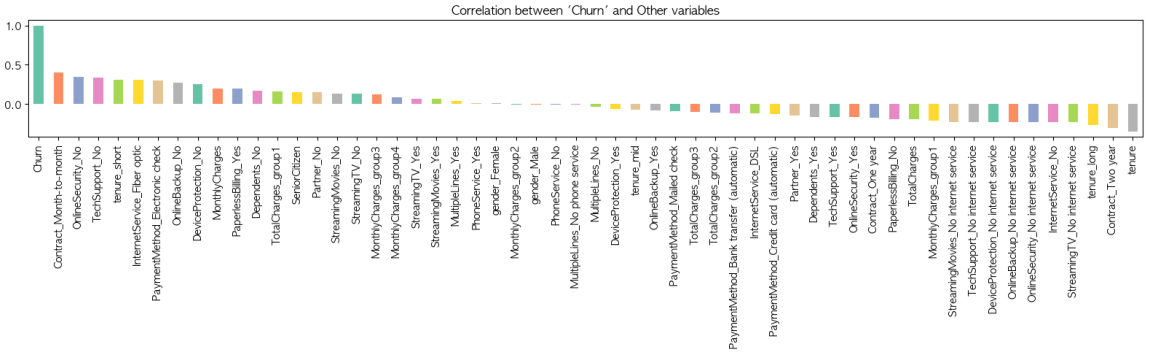
- 위 시각화 결과를 바탕으로, 고객 이탈 ('Churn') 유무를 구분하는 범위에 따라 'low', 'mid', 'high'의 3개의 범주형 변수를 생성했습니다.

4. 더미 변수 생성

```
# get_dummies
ramain_vars = ['customerID', 'Churn',
               'tenure_short', 'tenure_mid', 'tenure_long', 'MonthlyCharges_group1',
               'MonthlyCharges_group2', 'MonthlyCharges_group3',
               'MonthlyCharges_group4', 'TotalCharges_group1', 'TotalCharges_group2',
               'TotalCharges_group3', 'tenure', 'MonthlyCharges', 'TotalCharges',
               'tenure_bin', 'MonthlyCharges_group', 'TotalCharges_group']
categorical_vars = df.drop(ramain_vars, axis=1)
cat_df = pd.get_dummies(categorical_vars).astype(int)
df['Churn'] = df['Churn'].replace({"No": 0, "Yes": 1})
df = pd.concat([cat_df, df[ramain_vars]], axis=1)
```

- pandas의 `get_dummies` 함수를 이용해 범주형 변수에 대한 더미 변수를 생성했습니다.

'Churn' 변수와의 상관관계



- **'Churn' 변수와 각 변수와의 상관관계 크기에 따라 정렬한 막대 그래프**
  - 양의 상관관계가 높은 변수
    - 달 단위로 계약한 고객, 'OnlineSecurity'와 'TechSupport' 서비스를 이용하지 않는 고객, 총 이용 개월 수가 짧은 고객, Fiber optic의 인터넷 서비스를 이용하는 고객, 전자 수표를 이용한 결제 방식을 이용하는 고객일수록 이탈 확률이 높음을 알 수 있습니다.
  - 음의 상관관계가 높은 변수
    - 총 이용 개월 수가 짧은 고객, 2년 단위의 계약 기간을 이용하는 고객, 부가 서비스들을 이용하지 않는 고객일수록 이탈 확률이 낮음을 알 수 있습니다.

5. 후보 예측 모델 선정 및 성능 비교

위에서 전처리 한 데이터를 기반으로 sklearn의 Standard Scaler를 통한 정규화 후, 다양한 머신러닝 모델과 딥러닝 모델을 활용하여 예측 결과를 비교했습니다.

총 8가지 모델을 조사하여 각 모델의 장단점, 사용 목적, 튜닝 방향 등을 비교한 후 후보 예측 모델의 초기 성능을 비교하였습니다.

그 후 초기 결과를 기반으로, 최적 모델을 선정하고 하이퍼라라미터 튜닝을 통해 성능 개선을 진행했습니다.

5-1. 후보 예측 모델 선정 및 설명

모델	장점	단점	사용 목적	튜닝 방향
로지스틱 회귀 (Logistic Regression)	• 이해 및 구현이 간단 • 계산이 빠름 • 출력이 확률로 해석될 수 있어 직관적 • 정규화와 같은 기법을 통해 과적합 방지 가능	• 선형적 관계를 가정하므로 비선형 문제에 적합하지 않음 • 다중 공선성 문제 발생 가능	• 이진 분류 문제	• 정규화 파라미터 C • 다양한 정규화 기법 (L1, L2)
의사결정 나무 (Decision Tree)	• 시각화가 가능해 해석이 쉬움 • 비선형 데이터도 잘 처리 • 전처리가 거의 필요 없음	• 과적합 가능성이 높음 • 작은 변화에도 민감하여 불안정	• 분류 및 회귀 문제 • 데이터 해석이 중요한 경우	• 최대 깊이 (max_depth) • 최소 샘플 수 (min_samples_split,min_samples_leaf) • 분할 품질 기준 (criterion)
랜덤 포레스트 (Random Forest)	• 과적합 방지 • 안정적이고 높은 성능 • 변수 중요도 제공	• 많은 메모리와 시간 소모 • 매우 큰 데이터셋에 대해 비효율적일 수 있음	• 분류 및 회귀 문제 • 다양한 피처를 가진 데이터셋	• 트리의 개수 (n_estimators) • 최대 깊이 (max_depth) • 최소 샘플 수 (min_samples_split, min_samples_leaf)
그래디언트 부스팅 머신 (Gradient Boosting Machine, GBM)	• 높은 예측 성능 • 과적합 방지 • 다양한 손실 함수 지원	• 많은 메모리와 시간 소모 • 많은 하이퍼파라미터 튜닝 필요	• 분류 및 회귀 문제 • 고성능 모델이 필요한 경우	• 학습률 (learning_rate) • 트리의 개수 (n_estimators) • 최대 깊이 (max_depth) • 최소 샘플 수 (min_samples_split,min_samples_leaf)
XGBoost	• 높은 예측 성능 • 과적합 방지 • 빠른 학습 속도	• 많은 메모리와 시간 소모 • 많은 하이퍼파라미터 튜닝 필요	• 분류 및 회귀 문제 • 매우 큰 데이터셋	• 학습률 (learning_rate) • 트리의 개수 (n_estimators) • 최대 깊이 (max_depth) • 최소 샘플 수 (min_samples_split, min_samples_leaf) • 정규화 파라미터 (lambda, alpha)
LightGBM	• 높은 예측 성능 • 빠른 학습 속도 • 대용량 데이터 처리 가능	• 작은 데이터셋에 비효율적일 수 있음 • 많은 하이퍼파라미터 튜닝 필요	• 분류 및 회귀 문제 • 매우 큰 데이터셋	• 학습률 (learning_rate) • 트리의 개수 (n_estimators) • 최대 깊이 (max_depth) • 최소 샘플 수 (min_samples_split, min_samples_leaf) • 리프 노드 수 (num_leaves)
CatBoost	• 높은 예측 성능 • 범주형 데이터 자동 처리 • 빠른 학습 속도	• 많은 메모리 소모 • 많은 하이퍼파라미터 튜닝 필요	• 분류 및 회귀 문제 • 범주형 데이터가 많은 경우	• 학습률 (learning_rate) • 트리의 개수 (iterations) • 최대 깊이 (depth) • 정규화 파라미터 (l2_leaf_reg)
MLP (Multi-Layer Perceptron, 다층 퍼셉트론)	• 복잡한 비선형 문제 해결 가능 • 다양한 문제에 유연하게 적용 가능	• 많은 계산 자원 소모 • 과적합 가능성 높음 • 많은 데이터 필요	• 분류 및 회귀 문제 • 복잡한 패턴 인식이 필요한 경우	• 은닉층 수 및 노드 수 (hidden_layer_sizes) • 학습률 (learning_rate) • 활성화 함수 (activation) • 정규화 파라미터 (alpha)

5-2. 전처리 전후 데이터의 초기 성능 비교 및 결과 분석

원본 데이터를 이용한 초기 모델 성능 평가

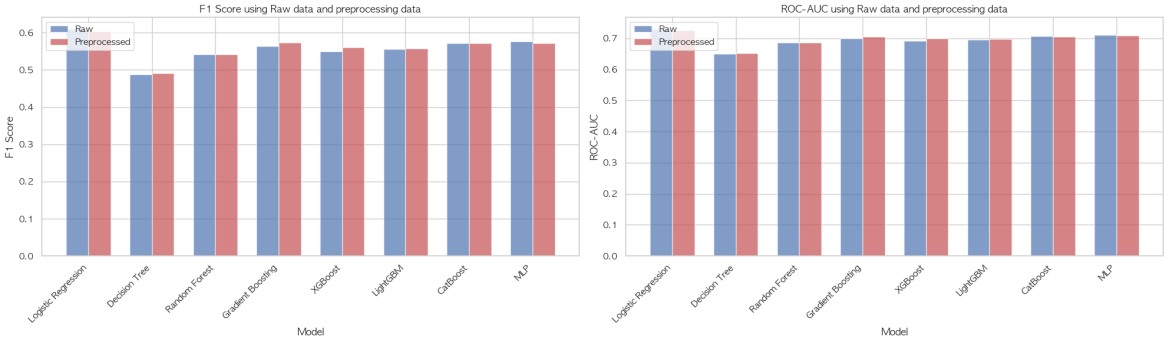
모델 이름	Accuracy	Precision	Recall
Logistic Regression	0.81	0.67	0.56
Decision Tree	0.73	0.48	0.49
Random Forest	0.78	0.62	0.48
Gradient Boosting	0.80	0.65	0.50
XGBoost	0.78	0.59	0.52
LightGBM	0.79	0.64	0.49
CatBoost	0.79	0.64	0.52

전처리 후 데이터를 이용한 초기 모델 성능 평가

모델 이름	Accuracy	Precision	Recall
Logistic Regression	0.73	0.66	0.55
Decision Tree	0.73	0.49	0.49
Random Forest	0.79	0.61	0.49
Gradient Boosting	0.80	0.66	0.51
XG Boost	0.79	0.61	0.52
LightGBM	0.79	0.63	0.50
CatBoost	0.79	0.63	0.52

모델 이름	Accuracy	Precision	Recall	모델 이름	Accuracy	Precision	Recall
MLP	0.78	0.59	0.57	MLP	0.771	0.59	0.56

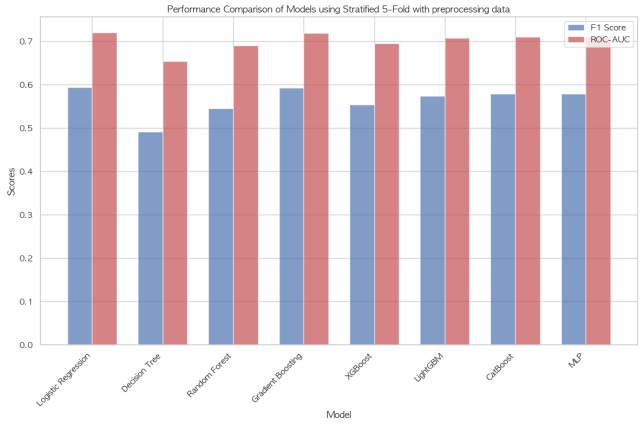
- **성능 비교 및 분석**



- 전처리 전후에 대해 각 모델의 성능을 평가한 결과, 일부 모델에서 성능 개선이 있음을 확인하였습니다.

  특히, 랜덤 포레스트와 XGBoost 모델에서 유의미한 성능 향상이 있었습니다. **그러나 전반적으로 로지스틱 회귀 모델의 성능이 가장 우수하게 나타났습니다.**

5-3. Stratified K-Fold 교차 검증 분석



이후 전처리 후 데이터로 Stratified K-Fold 교차 검증 분석을 진행하였습니다.

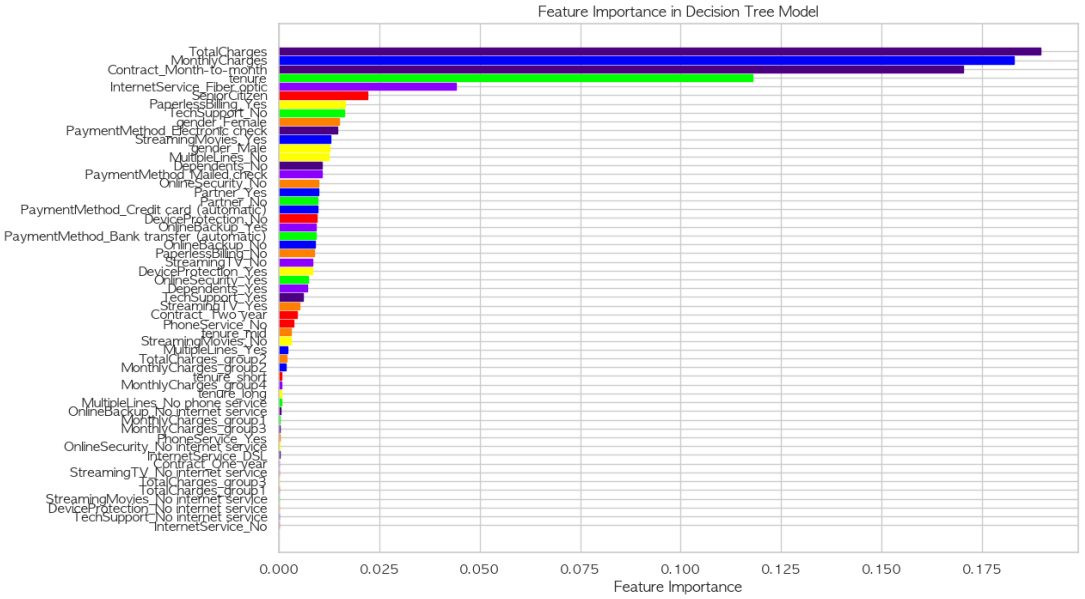
Stratified K-Fold는 불균형 클래스 문제를 해결하기 위해 각 폴드에서 클래스의 비율이 동일하게 유지되도록 데이터를 분할하는 방법입니다.

이를 통해 모델의 성능을 더욱 정확하게 평가할 수 있었습니다. 이 분석 결과, **ROC-AUC가 f1-score보다 성능이 더 높게 나타났습니다.**

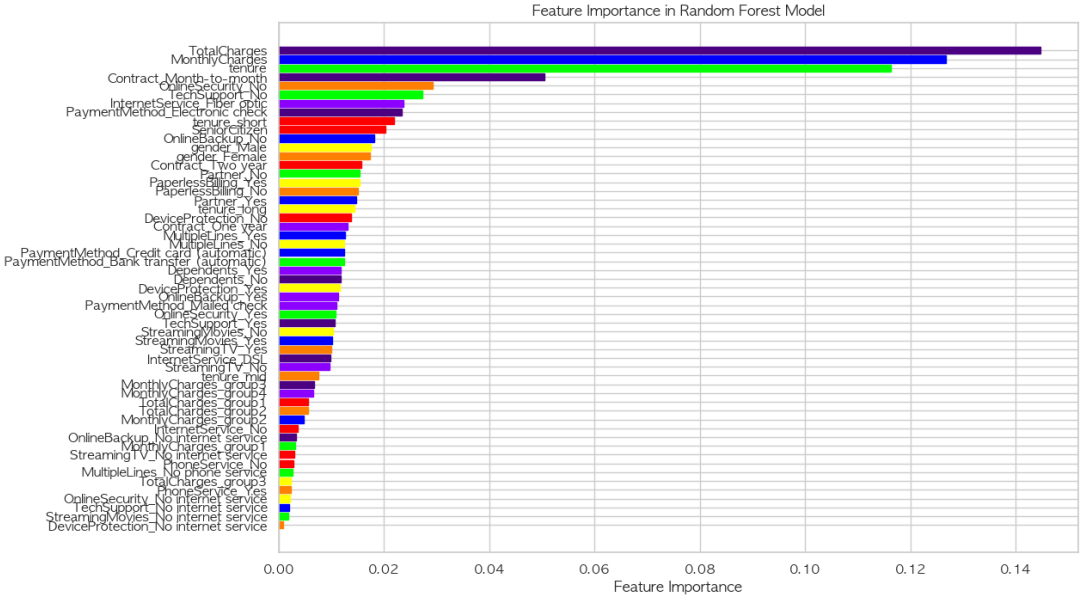
5-3. Feature Importance 분석

교차 검증 분석을 수행한 모델에 대해 Feature Importance 그래프를 바탕으로 어떤 변수들이 중요한지에 대한 통계를 분석하였습니다.

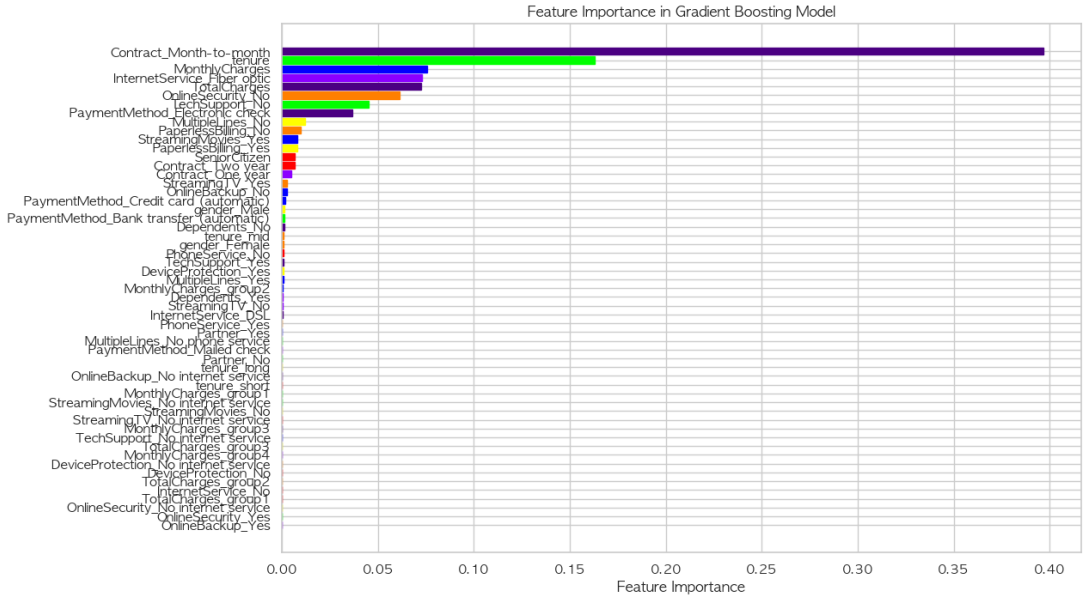
- **Decision Tree**



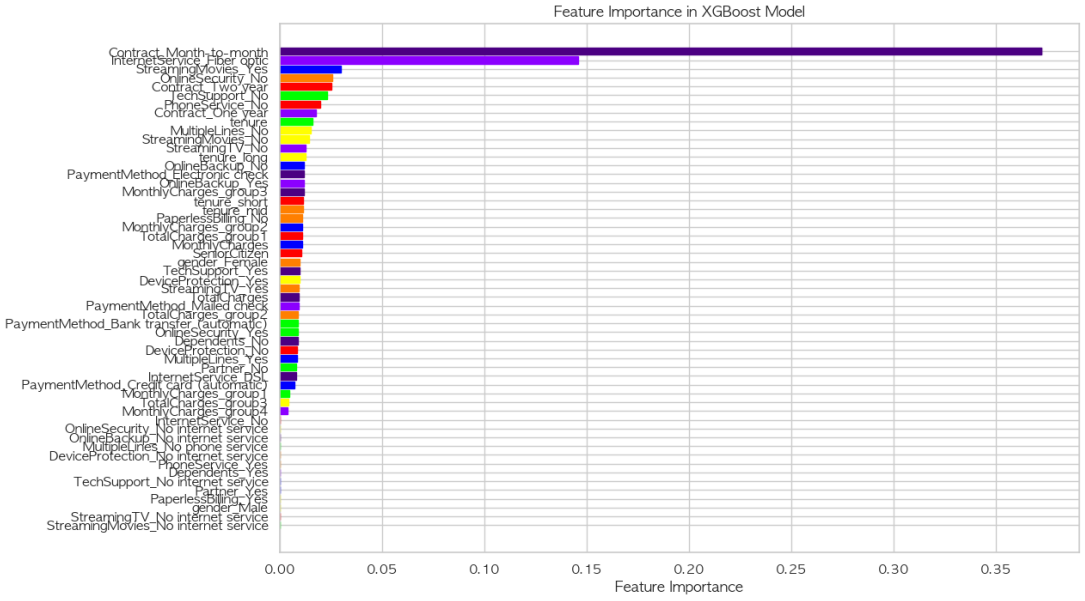
- **Random Forest**



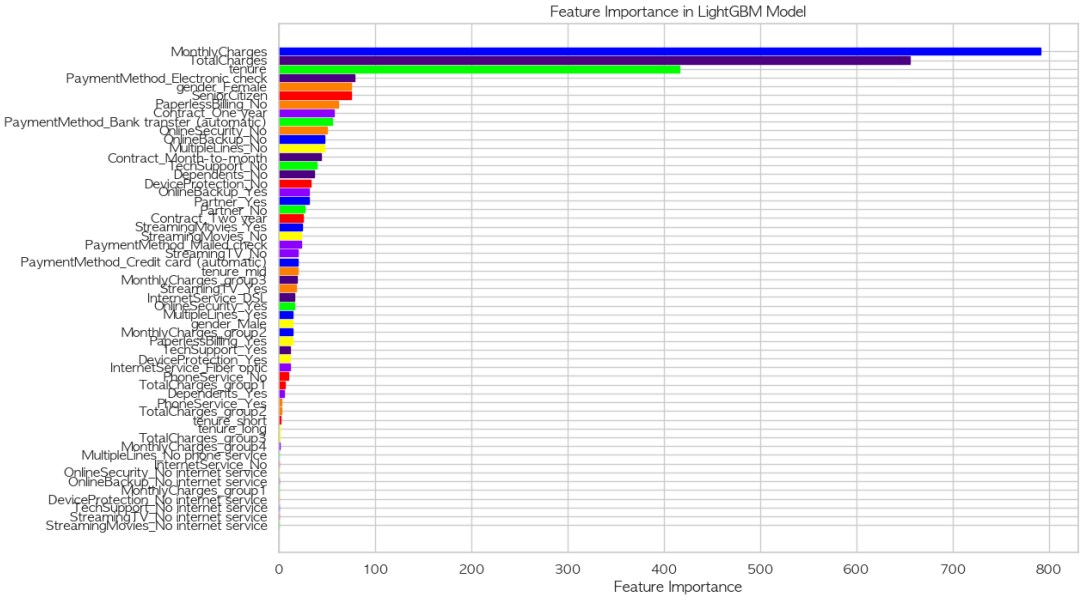
- **Gradient Boosting Model**



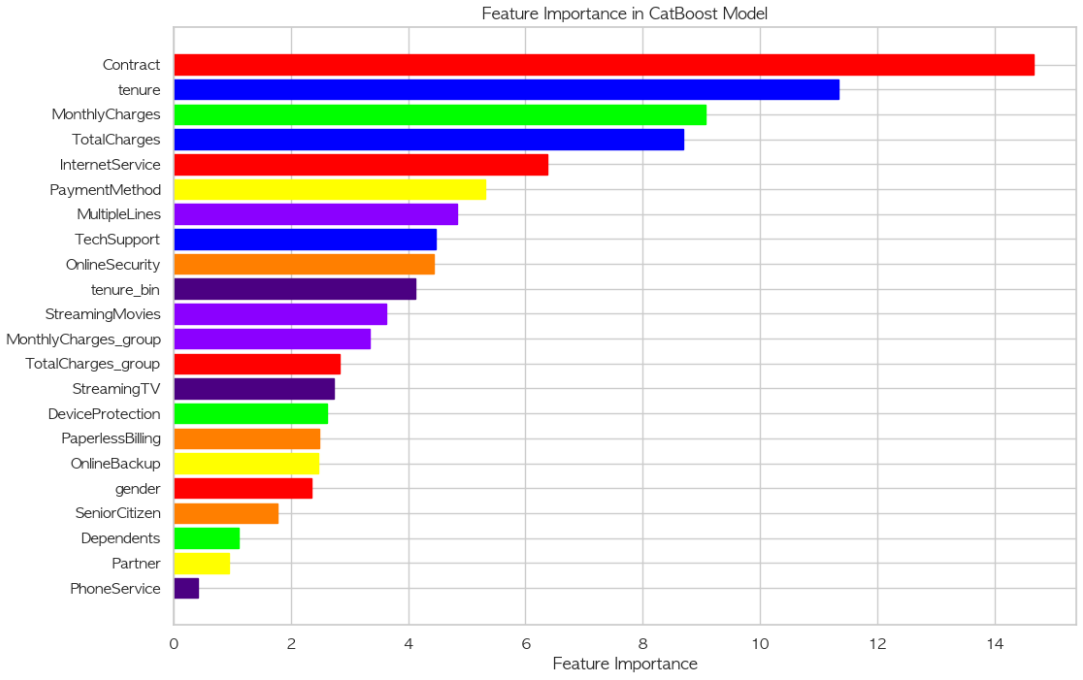
- **XGBoost Model**



- **LightGBM Model**



- **CatBoost Model**



- **Feature Importance 분석 결과**
  - 각 예측 모델들이 고객 이탈 예측에 있어 공통적으로 중요하다고 생각한 변수로는 'TotalCharges', 'MonthlyCharges', 'Contract', 'tenure' 등으로

    해당 통신사를 이용한 계약 기간과 이용 개월 수 및 월별 요금과 누적 요금이 중요한 것으로 알 수 있습니다.

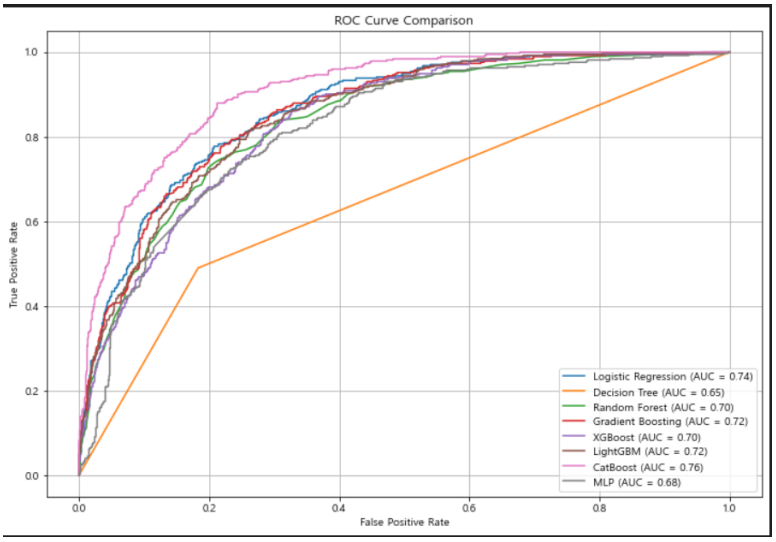
중간 결론

앙상블 모델이나 Boosting 모델이 가장 높은 성능을 보일 것으로 예상했지만, 실제 모델 예측 결과는 로지스틱 회귀가 가장 높은 것으로 나타난 것을 확인했습니다.

이에 대해 최종 모델을 조금 더 상세하게 간추려 보고자, 모든 후보 모델에 대해 GridSearchCV를 이용한 하이퍼파라미터 튜닝을 진행했습니다.

5-4. 후보 예측 모델의 하이퍼파라미터 튜닝

모델 이름	Accuracy	Precision	Recall	F1 Score	ROC-AUC	Confusion Matrix
Logistic Regression	0.8197	0.6913	0.5764	0.6287	0.7419	[[940 96] [158 215]]
Decision Tree	0.7303	0.4906	0.4879	0.4892	0.6528	[[847 189] [191 182]]
Random Forest	0.7977	0.6594	0.4879	0.5609	0.6986	[[942 94] [191 182]]
Gradient Boosting	0.8091	0.6745	0.5389	0.5991	0.7226	[[939 97] [172 201]]
XG Boost	0.7857	0.6127	0.5174	0.5610	0.6998	[[914 122] [180 193]]
LightGBM	0.7999	0.6454	0.5416	0.5889	0.7172	[[925 111] [171 202]]
CatBoost	**0.8453**	**0.7739**	**0.5871**	**0.6677**	**0.7627**	[[972 64] [154 219]]
MLP	0.7892	0.6508	0.4397	0.5248	0.6774	[[948 88] [209 164]]



클래스 불균형 분포에 대응하기 위해 precision과 recall을 조합한 F1-score와 전체 성능을 평가하는 ROC-AUC 지표를 중점적으로 모델 성능을 비교했습니다.

GridSearchCV를 통한 하이퍼파라미터 튜닝 결과, CatBoost 알고리즘이 모든 성능 지표에서 가장 높은 성능을 보이는 것으로 나타났습니다.

이 결과를 바탕으로, Target 변수를 제외한 나머지 20개의 변수 중 17개의 변수가 범주형 데이터임에 따라, 범주형 변수를 효과적으로 처리할 수 있는 CatBoost가 가장 적합한 알고리즘으로 예상됩니다. 다른 알고리즘의 경우, 17개의 범주형 변수에 대해 더미 변수를 생성해야 하므로, 이로 인해 모델이 불필요한 변수에 집중하게 되어 성능이 저하될 가능성이 있습니다.

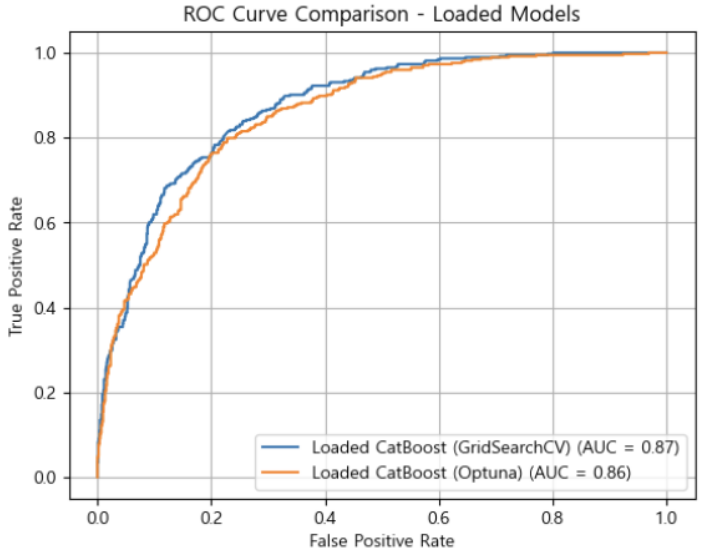
따라서 최종 예측 모델로 CatBoost 모델을 선정하였고, 이에 대해 추가적인 하이퍼파라미터 튜닝을 진행했습니다.

6. 최종 예측 모델 선정 및 하이퍼파라미터 튜닝

	Cat Boost (GridSearchCV)	Cat Boost (Optuna)
F1 Score	0.61	0.58
Accuracy	0.82	0.80
Recall	0.53	0.53
Precision	0.70	0.65
ROC AUC	0.87	0.86

```
Model Evaluation Metrics for Loaded CatBoost (GridSearchCV):
Accuracy: 0.8161816891412349
Precision: 0.7007042253521126
Recall: 0.5335120643431636
F1 Score: 0.6057838660578386
ROC AUC: 0.872677445733746
Confusion Matrix:
[[951  85]
 [174 199]]
```

```
Model Evaluation Metrics for Loaded CatBoost (Optuna):
Accuracy: 0.7998580553584103
Precision: 0.6491803278688525
Recall: 0.5308310991957105
F1 Score: 0.584070796460177
ROC AUC: 0.8560844452265364
Confusion Matrix:
[[929 107]
 [175 198]]
```



최종 모델로 선정된 CatBoost 모델을 하이퍼파라미터 튜닝을 진행하게 되면서 GridSearchCV와 Optuna를 활용하였습니다. 각각 하이퍼파라미터 튜닝을 위해 사용되지만 접근 방식과 효율성에서 큰 차이가 있습니다.

GridSearchCV는 모든 조합을 시도하기 때문에 전역 최적 해를 찾을 확률이 높지만, 조합 수가 많아질수록 계산 비용이 크게 증가합니다. 하지만 optuna는 샘플링과 프루닝(Pruning) 기법을 사용하여 비효율적인 조합을 일찍 배제하기 때문에 고차원 하이퍼파라미터 공간에서도 효과적으로 작동하지만 설정이 다소 복잡하고, 확률적 방법이기 때문에 항상 전역 최적 해를 보장하지는 않습니다.

이러한 차이와 총 연산에 걸렸던 800분을 고려하면 둘 다 높은 점수를 기록했지만 확실히 차이가 많이 난다는 점을 알 수 있습니다.

7. 예측 결과 분석 및 시사점

다양한 머신러닝 모델을 활용하여 고객 이탈 예측을 수행한 결과, 초기 테스트에서는 로지스틱 회귀 모델이 가장 높은 성능을 보였습니다.
그 후 Stratified K-Fold 교차 검증 및 Feature Importance 분석을 통해 추가적인 하이퍼파라미터 튜닝을 진행한 결과, 다른 모델에서도 성능 향상을 확인할 수 있었습니다.
특히, 20개의 컬럼 중 17개의 변수가 범주형 변수로 구성되어 있어, 범주형 변수 모델링에 유용한 CatBoost 알고리즘이 가장 적합한 모델임을 알 수 있었습니다.
앞으로의 분석에서는 더욱 정교한 파라미터 튜닝 및 모델 앙상블 기법을 도입하여 예측 성능을 향상시키는 방향으로 진행할 예정입니다.

또한, EDA에서 살펴본 결과 중에서 고객 이탈을 줄이기 위해서는 특히 인터넷 보안 서비스, 인터넷 백업 서비스, 디바이스 보호 서비스, 기술 지원 서비스와 같은 서비스를 기본서비스로 제공하거나, 부가 서비스들을 적극적으로 홍보하고, 다양한 이벤트를 통해 고객들이 해당 서비스를 더 많이 이용하도록 유도하는 전략이 필요함을 시사할 수 있습니다.

또 장기 고객을 우대 하는 것이 고객 이탈을 줄이는데 효과적일 수 있음을 알 수 있었습니다. 따라서 이용 기간에 따라 다양한 서비스 및 이벤트를 제공하여 고객들이 장기 고객으로 전환하도록 격려하는 전략을 채택하는 것이 바람직함을 알 수 있습니다.

이러한 분석 결과를 바탕으로, 통신사 기업은 고객 이탈 방지를 위한 전략을 보다 효과적으로 수립하고 실행할 수 있을 것입니다.

기술 스택

- 주요 언어: Python
- 데이터 분석: pandas, numpy, matplotlib, seaborn
- 인공지능 모델: scikit-learn, boosting 알고리즘, tensorflow, keras
- 환경: Visual Studio
- 협업: Discord

일정

- 1일차 (2024.07.05)
  - 데이터셋 검색 및 확정
  - 데이터셋 전처리 및 탐색적 데이터 분석 (EDA)
  - 기계 학습/딥러닝 후보 모델 선정 및 베이스라인 코드 작성
- 2일차 (2024.07.08)
  - 기계 학습/딥러닝 후보 모델 성능 비교 및 최종 모델 선정
  - 최종 모델 하이퍼파라미터 튜닝
  - 발표 자료 및 문서화 작업 완료 (결과 산출물 보고서, README 등)