PROJECT

팀 프로젝트 발표

캘리포니아 통신사 고객 이탈률

우리는다함께조

박초연 | 신혜원 | 안태영 | 윤상혁 | 허상호

목차

프로젝트 소개

- **선정 주제** | 캘리포니아 통신사 고객 이탈률 예측
- **데이터 수** | 행 7,044개, 열 66개
- **목표** | 캘리포니아 A통신사 고객 이탈률 예측을 통해 통신 시장에 대한 고찰

한국 통신시장 점유율

SKT 점유율 40% 턱걸이...더 팽팽해진 '통신 삼국지'

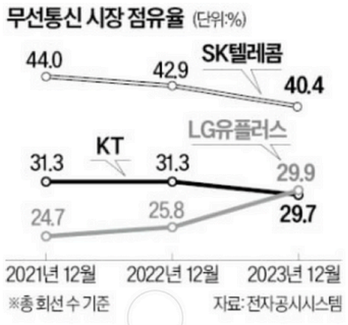
정지은 기자

입력 2024.03.07 17:56 수정 2024.03.08 02:06 지면 A13

작년말 SKT 점유율 40.4%
LG유플 29.9%·KT 29.7%
4대 3대 3 구도로 시장 재편

통신업계 부동의 1위로 꼽히던 SK텔레콤이 점유율 40%대에 턱걸이한 것으로 나타났다. LG유플러스는 0.2%포인트 차이로 KT를 앞서 2위에 올랐다. 수년간 '5 대 3 대 2'로 굳어 있던 무선통신서비스 시장 지형이 지난해를 기점으로 크게 바뀌었다는 평가가 나온다.

7일 LG유플러스가 공시한 사업보고서에 따르면 지난해 말 무선서비스 시장 점유율(총회선 기준)은 SK텔레콤 40.4%, LG유플러스 29.9%, KT 29.7%를 기록했다. ×. SK텔레콤의 는 수준까지 게 주요



세계일보

2024년에만 13만8000명 이탈… SK텔레콤 떠난 고객들, 왜?

입력 2024.09.30. 오후 2:10

올 들어 SK텔레콤 휴대폰 가입자 이탈이 눈에 띄게 늘고 있다. 소비자들이 가성비가 좋은 알뜰폰을 찾는 흐름이 계속되는 가운데 SK텔레콤 서비스에 대한 만족도가 떨어지는 게 주요인으로 분석된다. 이동통신 시장점유율 1위를 지켜온 SK텔레콤 위상이 흔들리는 것 아니냐는 지적이 나온다.

분석 프로세스

STEP 01
데이터 탐색

EDA
- 기초 통계량 확인
- 데이터 시각화 실시
- 변수 간 상관관계 확인

1차 변수 설정
- 분석 대상 변수 선정

STEP 02
데이터 전처리

기초 전처리
- 결측치
- 이상치 처리

모델링 준비
- train/test 분리
- SMOTE Sampling
- Standardization

STEP 03
모델학습

피처 분석
- Tree Bagging
- Random Forest

파라미터 튜닝
- Grid Search

STEP 04
성과평가

사용 알고리즘
- Decision Tree
- Random Forest
- SVM, XGBoost

성과비교
- Accuracy
- Precision
- Recall
- F1 score

STEP 05
모델해석

변수 중요도
- 피처 중요도
- SHAP Value

시사점
- 분석의 의의
- 시사점

데이터분석 및 전처리

데이터 개요

- **컬럼 수**
  총 66개의 컬럼 중 불필요한 37개의 컬럼 제거 후 29개 사용

- **결측치 처리**
  Offer, Total Charges 등

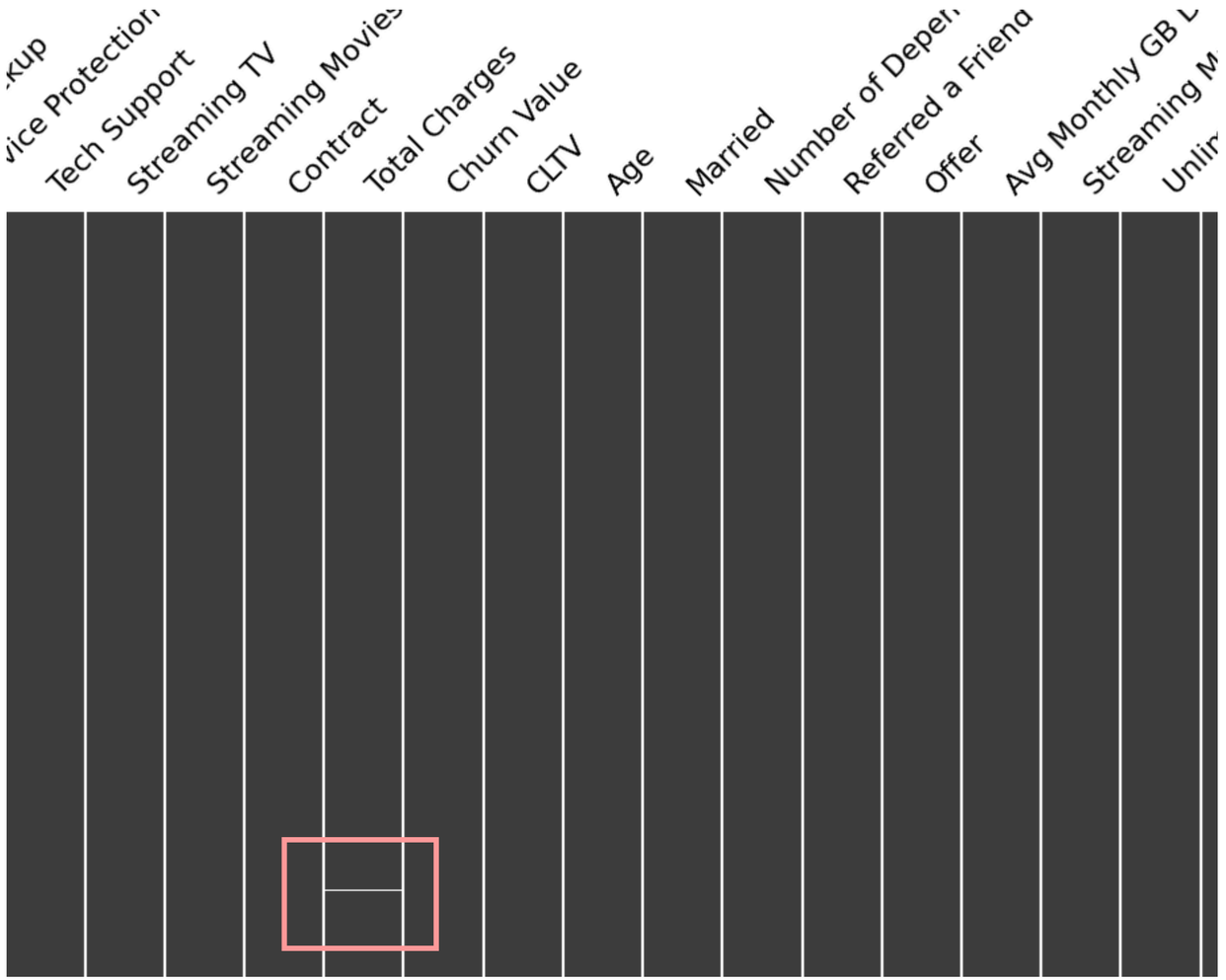
- **데이터의 특성을 고려하여 인코딩 진행**
  Age, Number of Dependents 등

전처리

- **결측치 처리**
  Offer, Total Charges 등

```
***** Check NA *****
Churn Reason                5174
Offer                       3877
Internet Type               1526
Churn Category              5174
```

- ✓ None
- ✓ Offer A
- ✓ Offer B
- ✓ Offer C



전처리

- 이상치 및 데이터 확인

***** DESCRIPTIVE STATISTICS *****

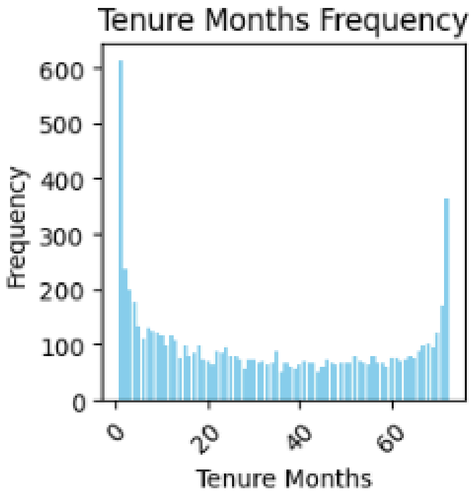
	count	mean	std	min	25%	50%
Unnamed: 0	7043.0	3521.000000	2033.283305	0.000000	1760.500000	3521.000000
Count	7043.0	1.000000	0.000000	1.000000	1.000000	1.000000
Zip Code	7043.0	93521.964646	1865.794555	90001.000000	92102.000000	93552.000000
Latitude	7043.0	36.282441	2.455723	32.555828	34.030915	36.391777
Longitude	7043.0	-119.798880	2.157889	-124.301372	-121.815412	-119.730885
Tenure Months	7043.0	32.371149	24.559481	0.000000	9.000000	29.000000
Monthly Charges	7043.0	64.761692	30.090047	18.250000	35.500000	70.350000
Churn Value	7043.0	0.265370	0.441561	0.000000	0.000000	0.000000
Churn Score	7043.0	58.699418	21.525131	5.000000	40.000000	61.000000
CLTV	7043.0	4400.295755	1183.057152	2003.000000	3469.000000	4527.000000
LoyaltyID	7043.0	550382.651001	260776.118690	100346.000000	323604.500000	548704.000000
Tenure	7043.0	32.371149	24.559481	0.000000	9.000000	29.000000
Age	7043.0	46.509726	16.750352	19.000000	32.000000	46.000000
Number of Dependents	7043.0	0.468692	0.962802	0.000000	0.000000	0.000000
ID	7043.0	795.888542	491.448525	1.000000	354.000000	780.000000
Number of Referrals	7043.0	1.951867	3.001199	0.000000	0.000000	0.000000
Tenure in Months	7043.0	32.386767	24.542061	1.000000	9.000000	29.000000
Avg Monthly Long Distance Charges	7043.0	22.958954	15.448113	0.000000	9.210000	22.890000
Avg Monthly GB Download	7043.0	20.515405	20.418940	0.000000	3.000000	17.000000
Monthly Charge	7043.0	64.761692	30.090047	18.250000	35.500000	70.350000
Total Refunds	7043.0	1.962182	7.902614	0.000000	0.000000	0.000000
Total Extra Data Charges	7043.0	6.860713	25.104978	0.000000	0.000000	0.000000
Total Long Distance Charges	7043.0	749.099262	846.660055	0.000000	70.545000	401.440000
Total Revenue	7043.0	3034.379056	2865.204542	21.360000	605.610000	2108.640000
Satisfaction Score	7043.0	3.244924	1.201657	1.000000	3.000000	3.000000

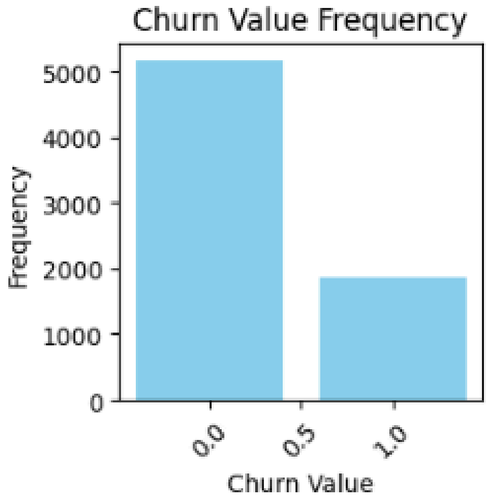
***** NUNIQUE *****

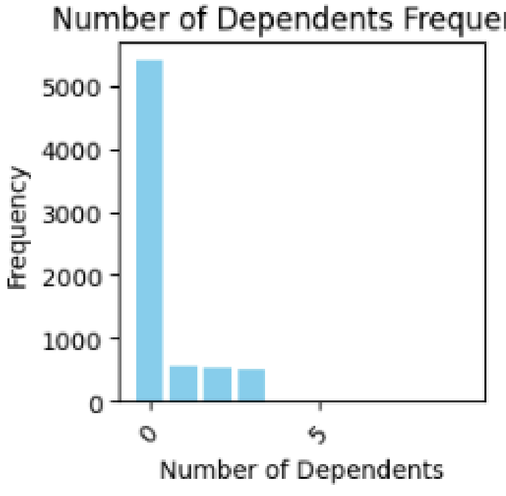
Unnamed: 0	7043
Customer ID	7043
City	1129
Zip Code	1652
Lat Long	1652
Latitude	1652
Longitude	1651
Location ID	7043
ID	1652
Population	1592
Service ID	7043
Total Long Distance Charges	6068
Total Revenue	6975
Status ID	7043
Payment Method	4
Monthly Charges	1585
Total Charges	6531
Churn Score	85
CLTV	3438
Churn Reason	20
LoyaltyID	7021
Tenure	73
Churn	2
Age	62
Under 30	2

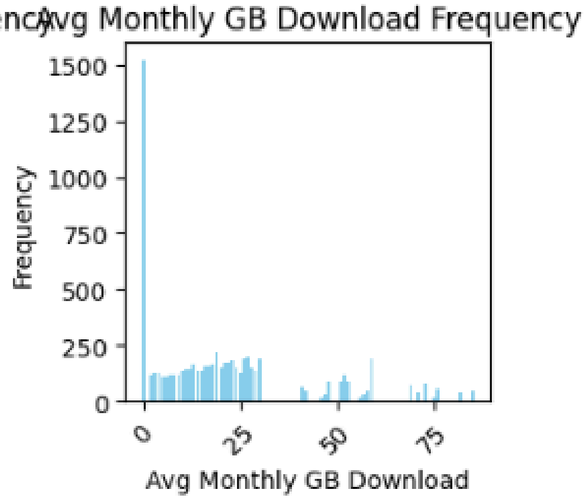
전처리

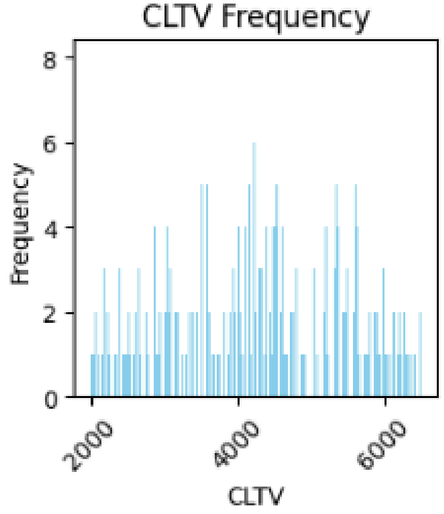
- **연속형 데이터의 분포 확인**

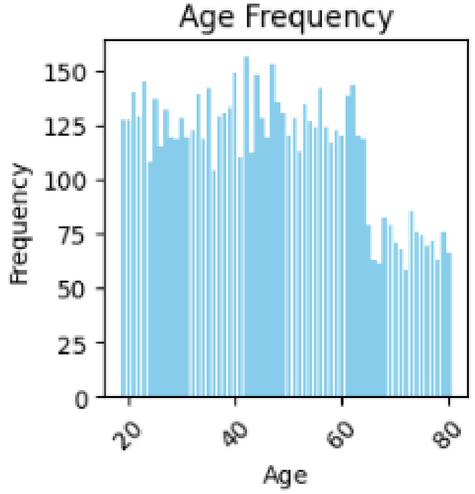


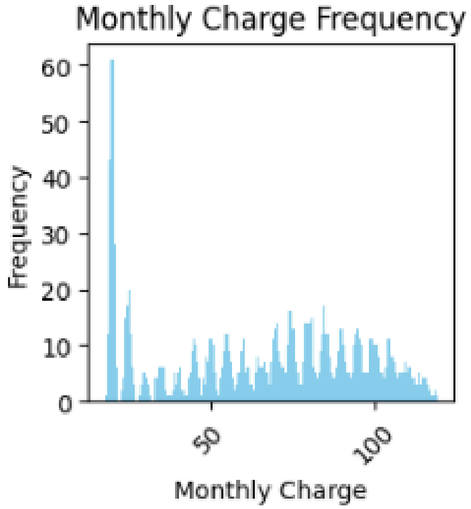


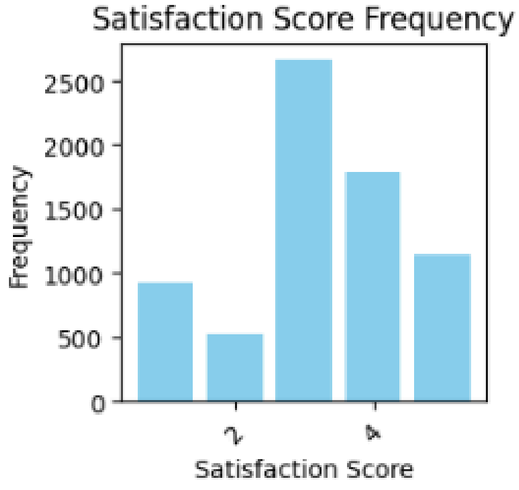












전처리

- 데이터 인코딩

```
# 데이터를 인코딩하는 함수 선언

def encode_data(data):
    #Tenure Months 열 인코딩
    data['Tenure Months'] = pd.cut(data['Tenure Months'], bins=3, labels=['Basic', 'Standard', 'Premium'])
    data['Tenure Months'] = data['Tenure Months'].replace({'Basic': 0, 'Standard': 1, 'Premium': 2}).astype(int)

    # Total Long Distance Charges 열 인코딩
    data['Total Long Distance Charges'] = pd.cut(data['Total Long Distance Charges'], bins=3, labels=['Low', 'Mediume', 'High'])
    data['Total Long Distance Charges'] = data['Total Long Distance Charges'].replace({'Low': 0, 'Mediume': 1, 'High': 2}).astype(int)

    # CLTV 열 인코딩
    data['CLTV'] = pd.cut(data['CLTV'], bins=3, labels=['Low', 'Mediume', 'High'])
    data['CLTV'] = data['CLTV'].replace({'Low': 0, 'Mediume': 1, 'High': 2}).astype(int)

    # Age 1~17:0, 18~44:1, 45~64:2, 65~:3
    bins_age = [-1, 17, 44, 64, 200]  # 구간 설정
    labels_age = [0, 1, 2, 3]  # 인코딩할 값

    data['Age'] = pd.cut(data['Age'], bins=bins_age, labels=labels_age)
    data['Age'] = data['Age'].astype(int)

    # Number of Dependents
    bins_dep = [-1, 1, 2, 200]   # 구간 설정
    labels_dep = [0, 1, 2]  # 인코딩할 값

    data['Number of Dependents'] = pd.cut(data['Number of Dependents'], bins=bins_dep, labels=labels_dep)
    data['Number of Dependents'] = data['Number of Dependents'].astype(int)
```

```
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Gender                     7032 non-null   int64
 1   Tenure Months              7032 non-null   int64
 2   Phone Service              7032 non-null   int64
 3   Multiple Lines             7032 non-null   int64
 4   Internet Service           7032 non-null   int64
 5   Online Security            7032 non-null   int64
 6   Online Backup              7032 non-null   int64
 7   Device Protection          7032 non-null   int64
 8   Tech Support               7032 non-null   int64
 9   Streaming TV               7032 non-null   int64
 10  Streaming Movies           7032 non-null   int64
 11  Contract                   7032 non-null   int64
 12  Total Charges              7032 non-null   int64
 13  Churn Value                7032 non-null   int64
 14  CLTV                       7032 non-null   int64
 15  Age                        7032 non-null   int64
 16  Married                    7032 non-null   int64
 17  Number of Dependents       7032 non-null   int64
 18  Referred a Friend          7032 non-null   int64
 19  Offer                      7032 non-null   int64
 20  Avg Monthly GB Download    7032 non-null   int64
 21  Streaming Music            7032 non-null   int64
 22  Unlimited Data             7032 non-null   int64
 23  Monthly Charge             7032 non-null   int64
 24  Total Refunds              7032 non-null   int64
 25  Total Extra Data Charges   7032 non-null   int64
 26  Total Long Distance Charges 7032 non-null  int64
 27  Total Revenue              7032 non-null   int64
 28  Satisfaction Score         7032 non-null   int64
```

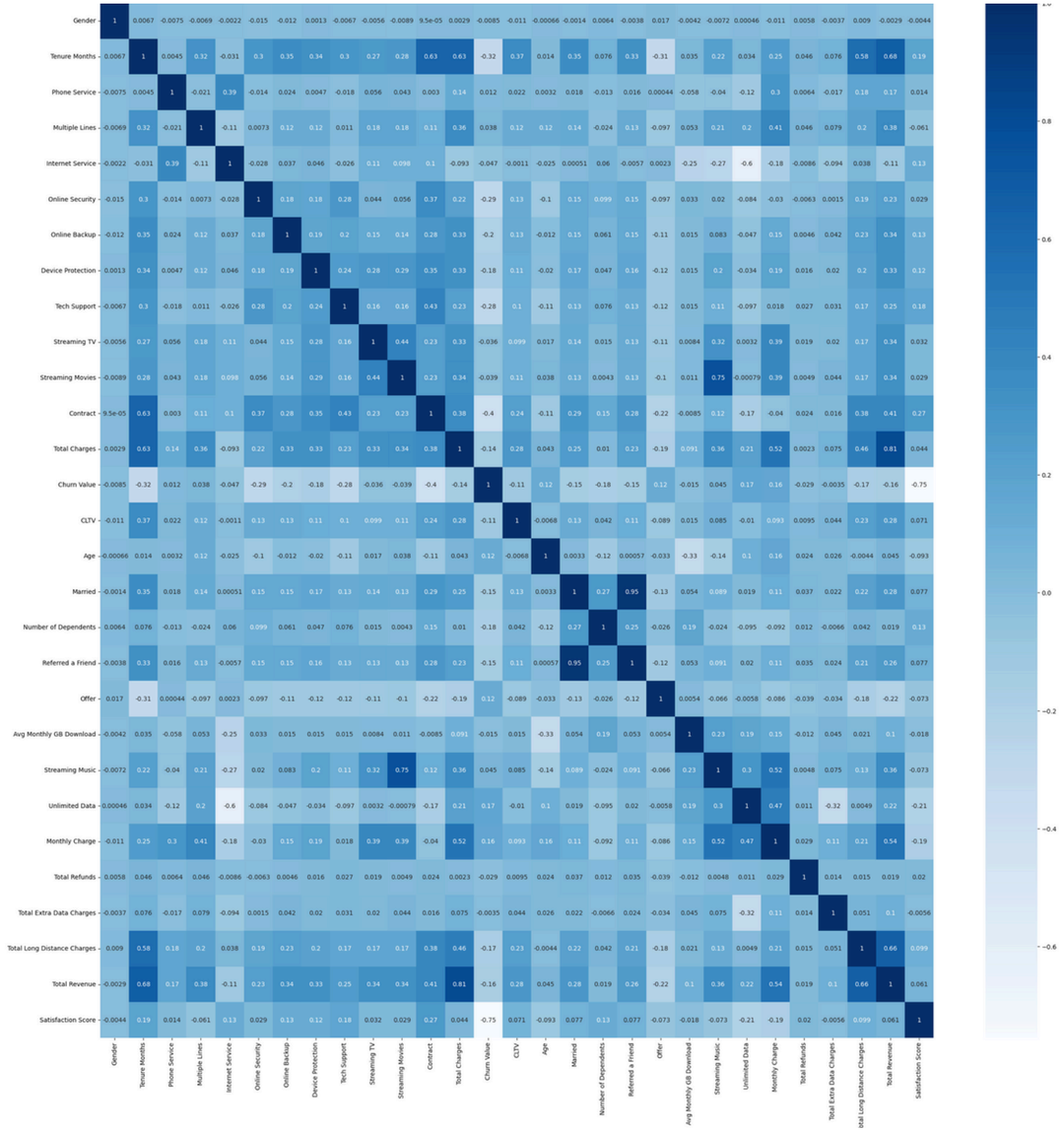
상관관계 분석

- **중요한 상관관계 발견:**

Total Charges와 Monthly Charges 사이에
0. 53의 비교적 높은 양의 상관관계가 나타남
-〉 총 청구 금액이 월 청구 금액과 밀접한 관련 있음 보임

Internet Service와 Streaming Movies,
Streaming TV 모두 상대적으로 강한 양의 상관관계
-〉 인터넷 서비스가 스트리밍 서비스 사용에 중요한 역할 함

Churn Value와 Total Charges는 -0.19로
약한 음의 상관관계를 보임
-〉 청구 금액이 클수록 이탈 가능성이 다소 낮아질 수 있음



주요 컬럼

컬럼명	설명	컬럼명	설명	컬럼명	설명
Gender	성별	Payment Method	요금 납부 방법	Age	나이
Senior Citizen	노년층	Monthly Charges	월 요금	Number of Referrals	추천인
Partner	동거인	Total Charges	총 요금	Offer	멤버십 서비스
Tenure Month	가입기간 (개월)	Churn Value	이탈여부	Total Revenue	총 통신요금
Phone Service	유선 전화 서비스	Churn Score	이탈 가능성 점수	Satisfaction	만족도
Contract	계약조건	CLTV	고객이 기업과 계약기간 기업에 지출하는 총 금액	Customer Status	고객 상태

최종 선정 컬럼

- Gender : 고객 설명(남성/여성)
- Tenure Months : 가입 기간(개월 수)
- Phone Service : 전화 서비스 사용 여부(예/아니오)
- Multiple Lines : 여러 전화 회선 사용 여부
- Internet Service : 인터넷 서비스 유형(DSL,광섬유,없음)
- Online Security : 온라인 보안 서비스 사용 여부(예/아니오/없음)
- Online Backup : 온라인 백업 사용 여부(예/아니오/없음)
- Device Protection : 디바이스 보호 서비스 사용 여부(예/아니오/없음)
- Tech Support : 기술 지원 사용 여부(예/아니오/없음)
- Streaming TV : 스트리밍 TV 사용 여부(예/아니오/없음)
- Streaming Movies : 영화 스트리밍 사용 여부(예/아니오/없음)
- Contract : 계약 유형(월별,1년,2년)
- Total Charges : 총 청구 요금
- Churn Value : 이탈 여부(1/0)

최종 선정 컬럼

- CLTV : 고객 생애 가치
- Age : 고객 나이
- Married : 결혼 여부
- Number of Dependents : 부양가족 수
- Referred a Friend : 친구 추천 여부(예/아니오)
- Offer : 제공된 혜택 정보
- Avg Monthly GB Download : 월평균 데이터 다운로드(GB)
- Streaming Music : 스트리밍 음악 사용 여부
- Unlimited Data : 무제한 데이터 사용 여부
- Monthly Charge : 월 청구 요금
- Total Refunds : 총 환불 금액
- Total Extra Data Charges : 총 추가 데이터 요금
- Total Long Distance Charges : 총 장거리 요금
- Total Revenue : 총 수익
- Satisfaction Score : 고객 만족도 점수(1~5)

가설에 따른 이탈률

고객 만족도 만족도가 높을수록 이탈률 ↓

고객 혜택 제공된 혜택 종류에 따라 이탈률 변동

계약 조건 계약조건이 길 수록 이탈률 ↓

가입 기간 가입기간이 길수록 이탈률 ↓

온라인 보안 부가서비스 이용 여부 이탈률 ↓

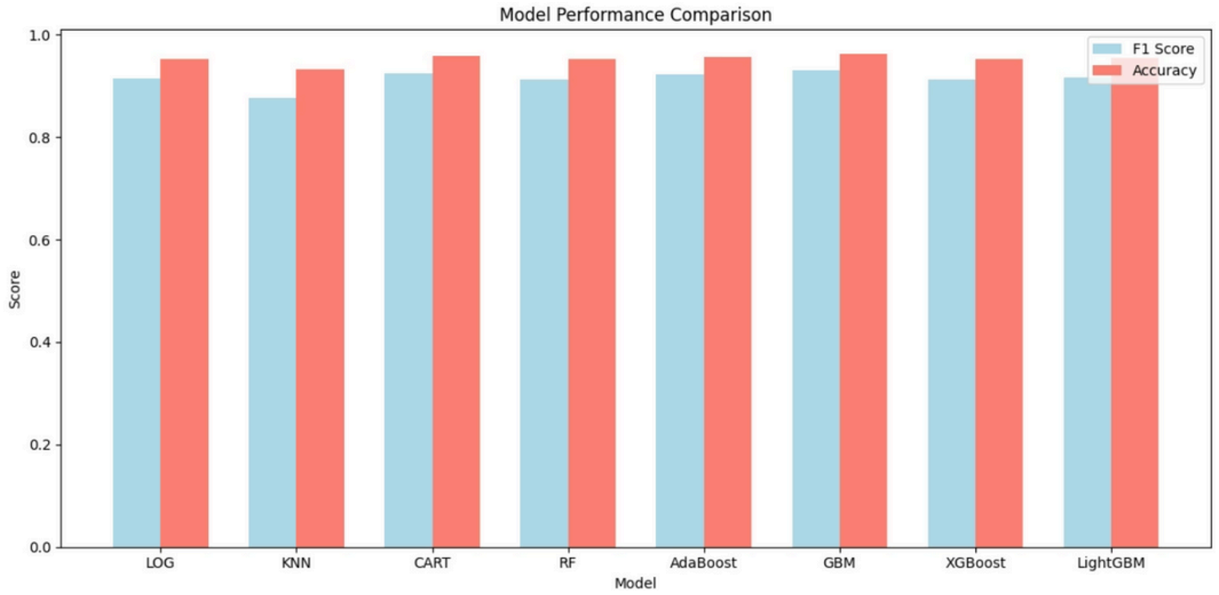
기술지원 기술지원 경험 이탈률 감소 ↓

모델 성능비교 및 해석

LOG, KNN, CART, RF, AdaBoost, GBM, XGBoost ...

하이퍼파라미터 그리드 설정

```
param_grids = {
    'LOG': {'C': [0.001, 0.01, 0.1, 1, 10]},
    'KNN': {'n_neighbors': [3, 5, 7, 9]},
    'CART': {'max_depth': [5, 10, 20, None]},
    'RF': {'n_estimators': [100, 200], 'max_depth': [10, 20, None]},
    'AdaBoost': {'n_estimators': [50, 100, 200]},
    'GBM': {'learning_rate': [0.01, 0.1, 0.2], 'n_estimators': [100, 200]},
    'XGBoost': {'n_estimators': [100, 200], 'max_depth': [3, 5, 7]},
    'LightGBM': {'learning_rate': [0.01, 0.1, 0.2], 'n_estimators': [100, 200]}
}
```

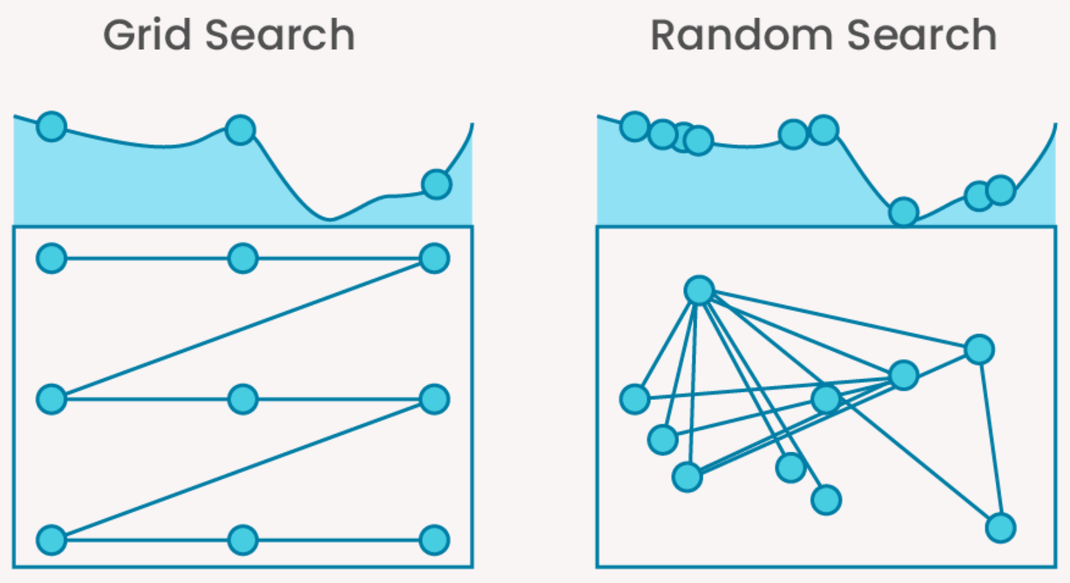


Voting, Stacking

```
# 스태킹
base_models = [
    ('xgb', XGBRFClassifier(random_state=42, use_label_encoder=False,
     eval_metric='mlogloss', n_jobs=-1, colsample_bytree = 0.7, gamma = 0,
     learning_rate = 1, max_depth = 19, n_estimators = 200)),
    ('rf', RandomForestClassifier(random_state=42, n_jobs=-1)),]

# soft voting
base_models = [
    ('lr', LogisticRegression(random_state=42, max_iter=10000)),
    ('rf', RandomForestClassifier(random_state=42, n_jobs=-1)),
    ('xgb', XGBRFClassifier(random_state=42, use_label_encoder=False,
     eval_metric='mlogloss', n_jobs=-1, colsample_bytree = 0.7, gamma = 0,
     learning_rate = 1, max_depth = 19, n_estimators = 200)),
    ("XGBoost", XGBClassifier(objective='reg:squarederror')),]

# hard voting
base_models = [
    ('rf', RandomForestClassifier(random_state=42, n_jobs=-1)),
    ('xgb', XGBRFClassifier(random_state=42, use_label_encoder=False,
     eval_metric='mlogloss', n_jobs=-1, colsample_bytree = 0.7, gamma = 0,
     learning_rate = 1, max_depth = 19, n_estimators = 200)),
    ("XGBoost", XGBClassifier(objective='reg:squarederror')),]
```

하이퍼파라미터 튜닝

- 모델의 파라미터 조합을 순차적으로 입력하면서 가장 성능이 좋은 조합을 탐색

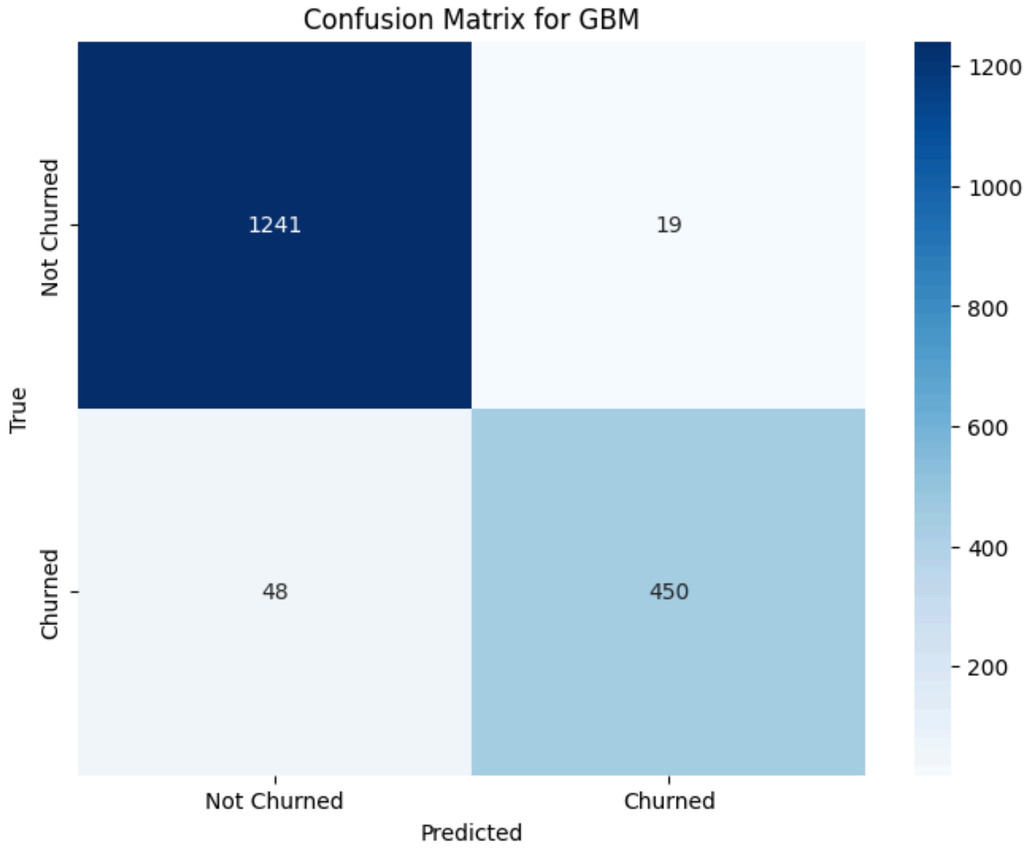


✓ 입력 파라미터 및 사용 코드

```
gbm = GradientBoostingClassifier()

param_grid = {
    'learning_rate': [0.01, 0.1, 0.2],
    'n_estimators': [100, 200],
}

grid_search = GridSearchCV(
    estimator=gbm,
    param_grid=param_grid,
    scoring='f1_macro',
    n_jobs=-1,
    cv=3,
    verbose=1
)

grid_search.fit(X_train, y_train)

best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test)
print(classification_report(y_test, y_pred))
print("Best Parameters: ", grid_search.best_params_)
```

→ 총 18회 탐색

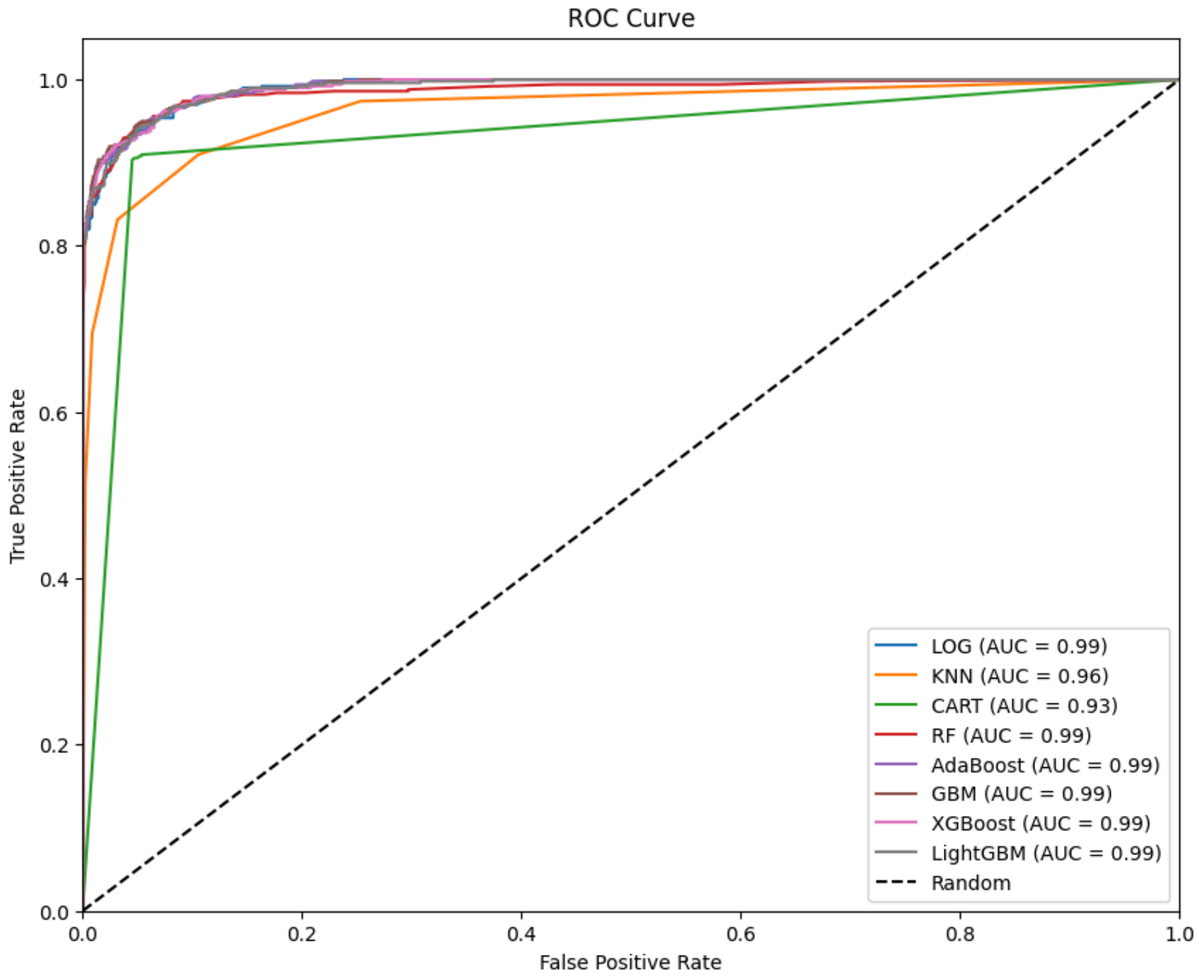
GBM



Training GBM...
Classification Report for GBM:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	1260
1	0.96	0.90	0.93	498
accuracy			0.96	1758
macro avg	0.96	0.94	0.95	1758
weighted avg	0.96	0.96	0.96	1758

ROC 커브 시각화 함수



- Logistic Regression과 트리 기반 모델들 (Random Forest, AdaBoost, GBM등) 모두 높은 AUC 기록함

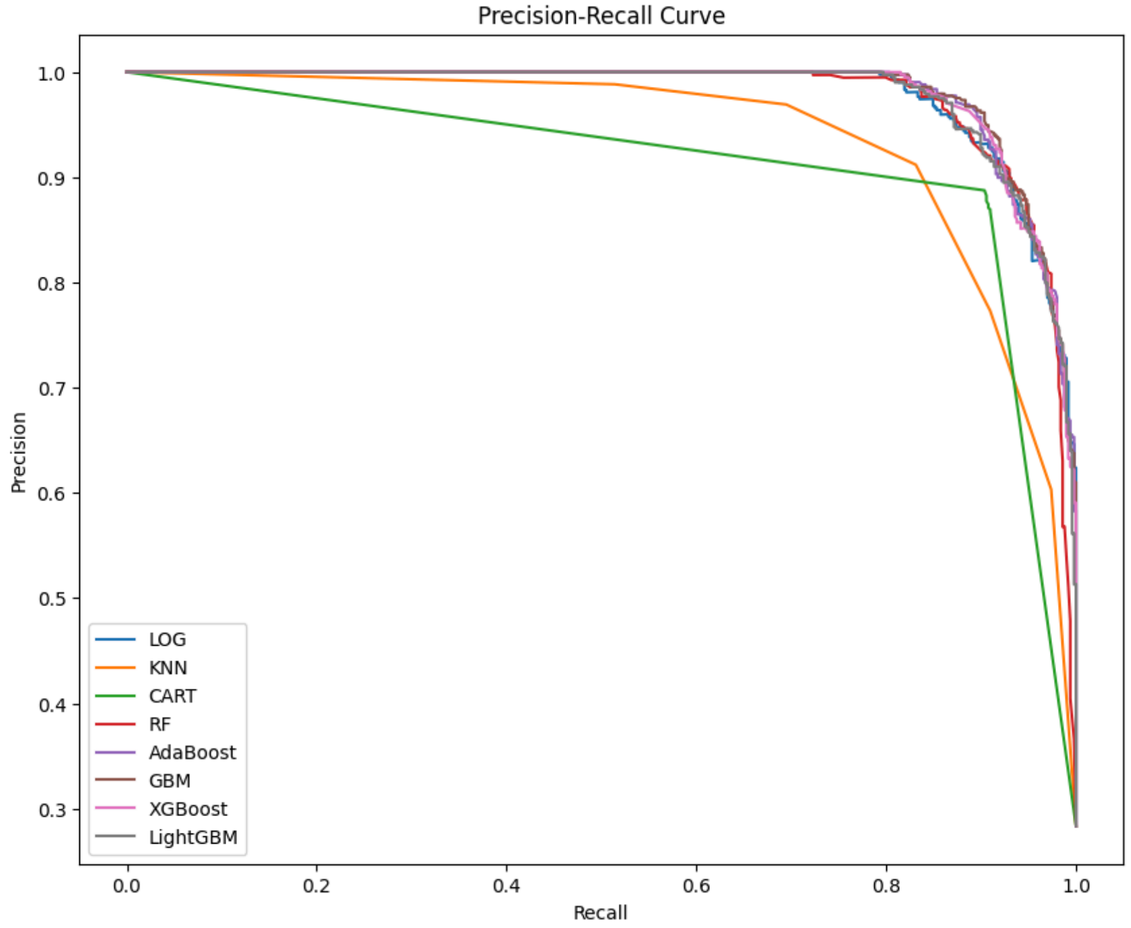
-> 이탈 예측에서 매우 신뢰할 수 있는 모델!

- CART 모델은 가장 낮은 AUC 기록

-> 잘못된 양성 예측이 다른 모델에 비해 많아 정확한 분류 어렵다는 것 의미

- KNN 역시 성능 괜찮지만, 다른 고성능 모델에 비해 일관된 예측이 어려울 가능성 있음

Precision-Recall 커브 시각화 함수



- 정밀도와 재현율간의 관계를 시각화함
- 데이터 불균형이 있는 경우 모델 성능 평가 유용

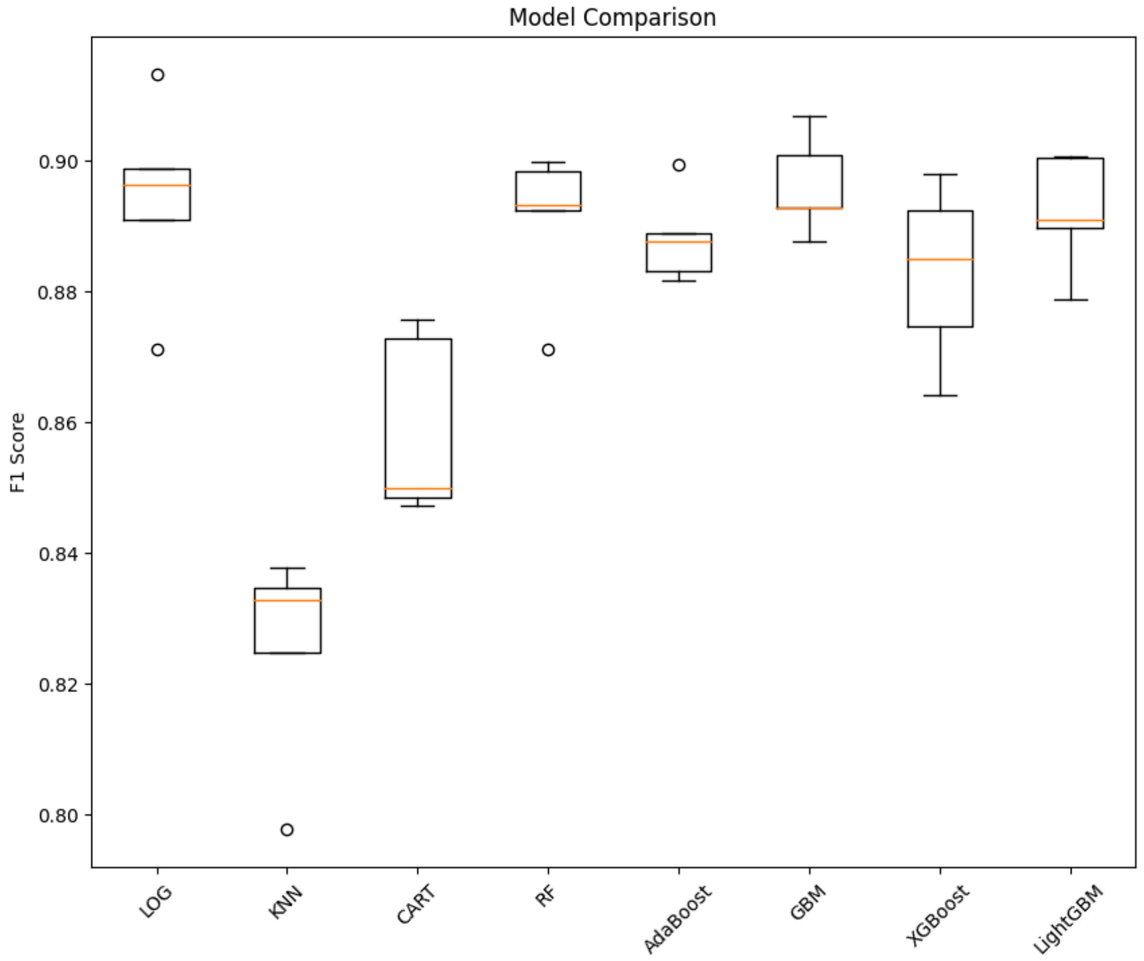
-> 특히 양성 클래스(고객 이탈) 예측의 성능 확인!

LOG, AdaBoost, LightGBM 등
높은 정밀도와 재현율을 유지해 좋은 성능 보임

KNN과 CART 는 다른 모델에 성능이 떨어지며,
정밀도와 재현율이 모두 감소하는 부분이 명확하게 나타남

Logistic Regression은 비교적 좋은 성능
but 다른 복잡한 모델들에 비해 유연성 부족 가능성

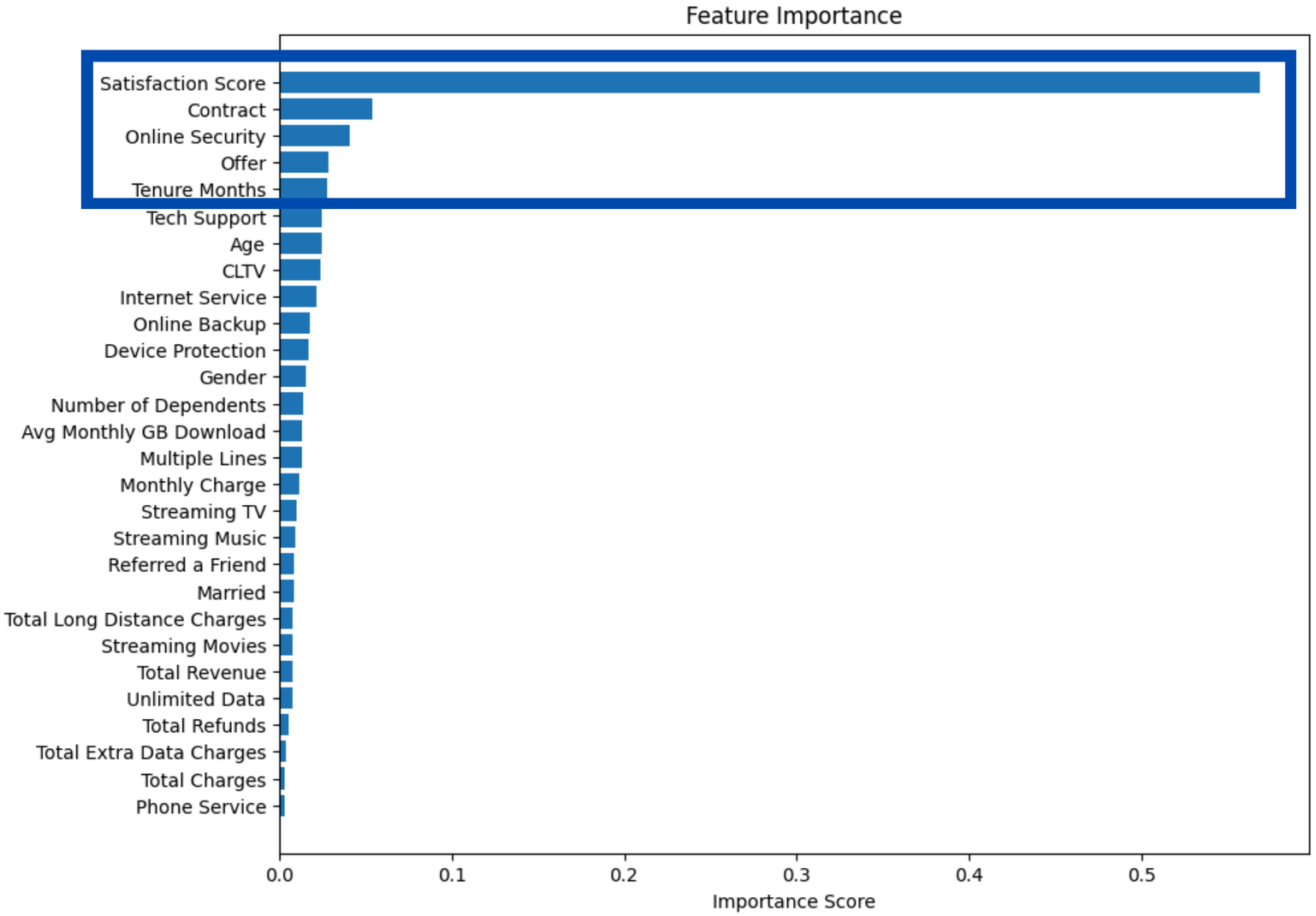
박스플롯 비교 함수



- LightGBM과 LOG 모델은 높은 F1 점수의 중앙값과 상대적으로 적은 분산을 보여줘,
  -> 대체로 안정적이고 일관된 성능 보임
- KNN과 CART 모델은 F1 점수가 낮고,
  특히 CART는 분산이 커서 성능의 일관성이 떨어질 가능성이 있음
  -> 예측 성능이 크게 달라질 가능성 존재
- RF, AdaBoost, GBM등 다른 트리 기반 모델들은 비교적 높은 F1 점수를 기록해
  -> 고객 이탈 예측에서 우수한 성능 보여줌

시사점

데이터 분석 결과, 고객 만족도가 통신사 이탈에 가장 큰 영향을 미치는 주요 요인



1. 고객 만족도
2. 통신사와 계약 조건
3. 온라인 보안
4. 부가 제공 서비스
5. 가입 기간

현재 통신사가 놓치고 있는 점

헤럴드경제

“영화 무료 연 24→3회로 싹둑” 유재석 ‘분노’ 폭발 했더니

입력 2024.09.06. 오후 9:39 · 수정 2024.09.07. 오전 10:31

통신사 장기 고객 혜택 축소를 작심하고 비판한 유재석 [유튜브 채널 캡처]

매경ECONOMY

단골이 호구? 발등 찍는 ‘배째라’ 통신사

입력 2024.10.15. 오후 9:01

> ‘가격 역전’부터 ‘선택약정할인’ 꼼수까지

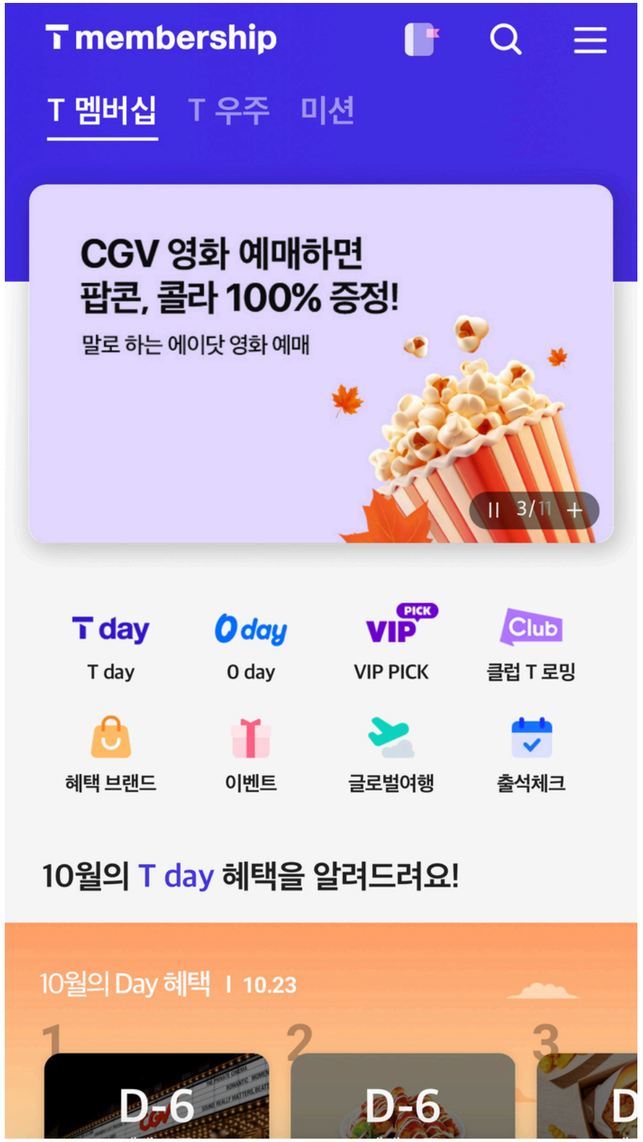
아시아경제

통신료도 비싼데 제대로 못 챙기는 할인

입력 2024.09.29. 오전 7:00 · 수정 2024.09.30. 오전 7:18

> (39)등급따라 혜택 받을 수 있는 폭 달라져
> 앱 자주 확인 안 하면 놓칠 수 있는 혜택 多

고객 만족도 향상 전략



이뉴스투데이 SKT, 한국산업의 고객만족도(KCSI) 27년 연속 1위

SKT는 다양한 고객의 수요를 적극 반영, 다양한 서비스로 고객 만족을 실천하고 있다.

SKT는 지난 19일 자녀 안심 앱 'ZEM'의 iOS 버전 아이용 앱을 출시했다. 국내 통신사 중 안드로이드와 iOS 운영 체제에서 모두 쓸 수 있는 자녀 스마트폰 관리 앱은 ZEM 뿐이다. 여기에 아이 관련 안심 기능도 대폭 강화해 '안심지도', '안심리포트' 등의 새로운 기능을 더했다.

SKT는 'AI 기반 고객 행동 예측 모델'을 통해 고객에게 차별화된 서비스를 제공할 수 있는 개인화 시나리오를 도출, 고객이 원하는 상황에 필요한 상품을 추천받을 수 있도록 할 예정이다.

더불어, G마켓, 롯데월드, 에버랜드, 한컴독스 등 신규 제휴처를 확대하고, 그에 따라 'T 우주패스 쇼핑 G마켓', 'T 우주패스 free' 등 새로운 구독 상품도 선보였다. SKT는 여기에 그치지 않고 연내 'T 우주패스 미디어', 'T 우주패스 DIY', 'T 우주패스 Google One', 'T 우주패스 마켓컬리' 등 다양한 신규 구독 상품을 선보여 고객이 마음대로 고를 수 있는 구독 마켓의 모습을 갖춰 나갈 계획이다.