wrap up 1일차

1. 분류 (Classification) - 가장 기본적인 예측 모델링
도메인: 통신 & 이탈 방지 (Business & Churn Prevention)

---

1. 데이터셋 및 컬럼

- **데이터셋:** Kaggle의 **"Telco Customer Churn (WA_Fn-UseC_-Telco-Customer-Churn.csv)"** 데이터셋을 사용했습니다. 가상의 미국 통신사 고객 7,043명의 정보가 담겨있습니다.
- **주요 컬럼:**
  - **고객 정보:** `gender`(성별), `SeniorCitizen`(고령자 여부), `Partner`(배우자 여부), `Dependents`(부양가족 여부)
  - **계정 정보:** `tenure`(가입 기간: 개월), `Contract`(계약 유형), `PaymentMethod`(결제 방식), `PaperlessBilling`(전자 청구서), `MonthlyCharges`(월 요금), `TotalCharges`(총 요금)
  - **가입 서비스:** `PhoneService`(전화), `MultipleLines`(다회선), `InternetService`(인터넷), `OnlineSecurity`, `OnlineBackup`, `DeviceProtection`, `TechSupport`, `StreamingTV`, `StreamingMovies`(각종 부가 서비스)
  - **타겟 변수 (Y):** `Churn`(이탈 여부: Yes/No)

---

2. 데이터 클리닝 및 전처리 (EDA 단계)

시각화(EDA)를 위해 데이터를 정리하는 몇 가지 중요한 작업을 수행했습니다.

1. **결측치 확인:** `missingno` 라이브러리의 `matrix` 플롯으로 시각화한 결과, 눈에 띄는 결측치(흰색 줄)는 **없었습니다.**
2. **데이터 타입 오류 수정 (`TotalCharges`):**
   - `TotalCharges`(총 요금) 컬럼이 숫자(`float`)가 아닌 문자열(`object`)로 되어 있었습니다. (빈칸 `''` 등 포함)
   - `pd.to_numeric(..., errors='coerce')`를 사용해 숫자로 강제 변환하고, 이 과정에서 생긴 `NaN` 값을 `dropna()`로 제거했습니다.
3. **데이터 값 변환 (`SeniorCitizen`):**
   - `SeniorCitizen` 컬럼이 `0`(비고령자), `1`(고령자)로 되어 있어 의미 파악이 어려웠습니다.

- `.map({0: "No", 1: "Yes"})`를 사용해 시각화 및 분석이 쉽도록 **문자열 "No"/"Yes"로 변환**했습니다.

4. **이상치 제거 (`tenure`):**
    - `tenure`가 0인 데이터(신규 고객)가 일부 존재하여, 이 데이터도 `drop()`으로 제거했습니다.

---

3. 시각화를 통한 고객 인사이트 (EDA)

다양한 시각화를 통해 고객 이탈에 영향을 미치는 핵심 요인들을 파악했습니다.

📊 분포 확인 (도넛 차트 / 히스토그램)

- **이탈률 (Churn):** 전체 고객 중 **26.6%가 이탈(Yes)**, 73.4%가 유지(No)했습니다. 이는 '이탈'이 '유지'보다 훨씬 적은 "불균형 데이터셋"임을 의미합니다.
- **성별 (Gender):** 남녀 비율이 50.5% vs 49.5%로 거의 1:1이었습니다.
- **결제 방식 (PaymentMethod):** 'Electronic check'(전자 수표)가 33.6%로 가장 많았습니다.

📈 이탈(Churn) 관련 핵심 인사이트 (막대/박스 플롯)

1. **계약 유형 (Contract) - 🌟 가장 강력한 요인**
    - **"Month-to-month" (월간 계약)** 고객이 이탈 고객의 **대부분을 차지**했습니다.
    - 반면, **"Two year" (2년 약정)** 고객은 거의 이탈하지 않았습니다.
2. **가입 기간 (Tenure) - 🌟 두 번째 요인**
    - **이탈(Yes) 고객**의 가입 기간 중간값은 약 **10개월**이었습니다.
    - **유지(No) 고객**의 가입 기간 중간값은 약 **38개월**이었습니다.
    - **결론: 가입한 지 1년이 안 된 신규 고객이 이탈 위험이 매우 높습니다.**
3. **요금 (Charges)**
    - **월 요금(MonthlyCharges):** 이탈 고객은 **높은 월 요금** 구간에 많이 분포했습니다.
    - **총 요금(TotalCharges):** 이탈 고객은 **낮은 총 요금** 구간에 많이 분포했습니다.
    - **(종합): 비싼 요금제를 쓰는 신규 고객**이 이탈 위험군입니다.
4. **부가 서비스 (Services)**

- `OnlineSecurity`나 `TechSupport` (기술 지원) 서비스에 **가입하지 않은('No') 고객**의 이탈률이 가입한 고객보다 훨씬 높았습니다.

5. **인구 통계 (Demographics)**
   - *`SeniorCitizen` (고령자)이 비고령자보다 이탈률이 높았습니다.
   - `Partner` **(배우자)가 없거나** `Dependents` **(부양가족)가 없는** 고객의 이탈률이 더 높았습니다.

---

4. 모델링 및 결과

- **전처리:** 숫자형 변수(`tenure` 등)는 `StandardScaler`로 표준화했고, 문자열 변수(`Contract`, `SeniorCitizen` 등)는 `LabelEncoder`로 0, 1, 2... 같은 숫자로 변환했습니다. (이 과정에서 "No"/"Yes"로 바꿨던 `SeniorCitizen`이 다시 0/1로 변환되었습니다.)
- **데이터 분할:** `train_test_split`을 사용해 70%(Train) / 30%(Test)로 분할했습니다. 이때 `stratify=y` 옵션을 사용하여, '이탈' 비율(26.6%)이 훈련/테스트셋 모두에 동일하게 유지되도록 했습니다.

🤖 사용한 분석 기법 및 결과 요약

모델 기법	장점 (일반적)	단점 (일반적)	노트북 결과 (Accuracy)	'이탈(1)' 예측 F1-Score (핵심)
KNN	간단, 직관적	느림, 스케일링 민감	78%	0.52 (매우 낮음)
SVC	비선형 데이터에 강함	느림, 불균형 데이터에 민감	81%	0.50 (매우 낮음)
Decision Tree	해석 용이, 시각화 가능	**과적합 (Overfitting) 매우 쉬움**	72%	0.50 (성능 최하)
Random Forest	DT의 과적합 방지, 고성능	해석 어려움	81%	0.59 (낮음)
Logistic Regression	빠름, 해석 용이 (계수), 확률 제공	선형 관계 가정, 성능 한계	81%	**0.62 (개별 모델 중 Best)**
AdaBoost	성능 좋음, 과적합 저항	이상치 민감	81%	0.60 (낮음)
Gradient Boosting	고성능 (대회 단골)	파라미터 민감, 느림	81%	0.60 (낮음)
Voting (Ensemble)	개별 모델 단점 보완 (안정성)	복잡함, 시간 오래 걸림	**82%**	**0.63 (최종 모델)**

🚨 모델링 결론: 심각한 문제 발견

- Accuracy(정확도)는 80% 전후로 그럴듯해 보이지만, 이는 **다수 클래스인 '유지(0)' 고객만 잘 맞혔기 때문**입니다.
- `classification_report`와 `confusion_matrix`를 보면, **모든 모델**이 정작 중요한 **'이탈(1)' 고객을 40~50% 가까이 놓치고 있습니다.** (Recall 값이 0.50 ~ 0.58에 불과)
- "이탈할 고객"을 "유지할 것"이라고 잘못 예측하는 심각한 한계를 보였습니다.
- 이는 모델의 성능 문제라기보다는 "불균형 데이터"를 제대로 처리하지 않았기 때문입니다.

---

5. 최종 인사이트 및 실제 적용 방안

🎯 고객 이탈 핵심 이유 (종합)

1. **계약 형태:** "월간 계약(Month-to-month)"이 이탈의 가장 강력한 원인입니다.
2. **가입 기간:** "10개월 미만의 신규 고객"의 이탈 위험이 매우 높습니다.
3. **서비스 경험: "온라인 보안", "기술 지원"** 등 핵심 부가 서비스 미가입 시 이탈률이 급증합니다.
4. **요금:** "월 요금"이 비싸다고 느끼는 고객층에서 이탈이 발생합니다.

💡 실제 비즈니스 적용 방안

이 분석 결과를 바탕으로 다음과 같은 즉각적인 조치를 취할 수 있습니다.

1. **[타겟 마케팅] "월간 계약" 고객에게 "장기 계약" 전환 유도**
    - 월간 계약 고객 중 가입 기간 10개월 미만 + 비싼 요금제 사용 고객을 "초고위험군"으로 분류합니다.
    - 이들에게 **"1년/2년 약정 시 월 요금 할인"** 또는 **"핵심 부가서비스(보안/기술지원) 6개월 무료"** 혜택을 제공하여 장기 계약으로 묶어둡니다(Lock-in).
2. **[고객 관리] "신규 고객 온보딩(On-boarding)" 프로그램 강화**
    - 가입 후 첫 10개월이 이탈 방지의 "골든 타임"입니다.
    - 이 기간에 서비스 활용법 안내, 웰컴 쿠폰 제공 등 긍정적인 경험을 집중적으로 제공하여 고객 이탈을 방지합니다.
3. **[모델 개선] "Recall" 점수 높이기 (필수)**
    - 현재 모델은 이탈 고객 2명 중 1명을 놓치므로 현업 적용이 불가능합니다.

- 데이터 전처리 단계에서 **SMOTE**나 **ADASYN** 같은 **오버샘플링(Oversampling)** 기법을 적용하여 '이탈(1)' 데이터를 인위적으로 늘린 후, 모델을 재학습시켜야 합니다. (목표: '이탈'의 Recall 점수를 80% 이상으로 끌어올리기)

6. 회고록

우선 이분은 굉장히 다양한 기법으로 시각화를 하시는게 가장 인상깊었습니다.

특히 단순히 종속변수인 이탈여부만 본 것이 아니라 이탈여부와 성별로도 묶어보는 등 여러 가지 방식으로 다양한 그래프를 그려볼 수 있었습니다.

8가지 분석기법을 사용해서 예측을 해보았는데 이탈고객 데이터의 예측률이 50-60로 유용한 모델이 도출되진 못했다.

불균형 샘플링이 문제인 것으로 보아 차후 오버샘플링으로 데이터의 이탈 고객의 데이터 수를 늘려 다시 진행해보고 어떤 결과가 나오는지 도출해봐야 될 것 같습니다.