이탈 고객 분류 모델
(통신사 이탈률 분석)

제출일	2021, 06, 06	**전공**	데이터사이언스융합
과목	Advanced Machin Learning	**학번**	2020710189
담당교수	김재광 교수님	**이름**	고준석

문제 정의

카드 회사나 통신사 등 많은 기업에서 고객의 이탈을 막기위해 막대한 비용을 쏟고 있다. 통신사의 경우 이러한 이탈 고객을 전담 대응하는 '해지방어'라는 부서가 따로 존재할 정도이다. 하지만 이는 고객이 먼저 해지한다는 사실을 통신사에 고지했을 때 수동적으로 대응할 뿐이다. 많은 고객은 이러한 고지 없이 바로 이탈해 버리기 때문에 이를 막기위해 이탈고객을 예측하는 분석을 진행하였다.

원래 목표는 신용카드 사용정보로 이탈고객을 분류하려 했으나, 사용내역 데이터는 쉽게 구할 수 있지만 타겟 변수인 '이탈' 정보가 포함된 데이터 셋은 찾을 수 없었다. 그래서 UCI에서 실제와 유사하게 만든 통신사 이탈 dataset을 사용하였다.

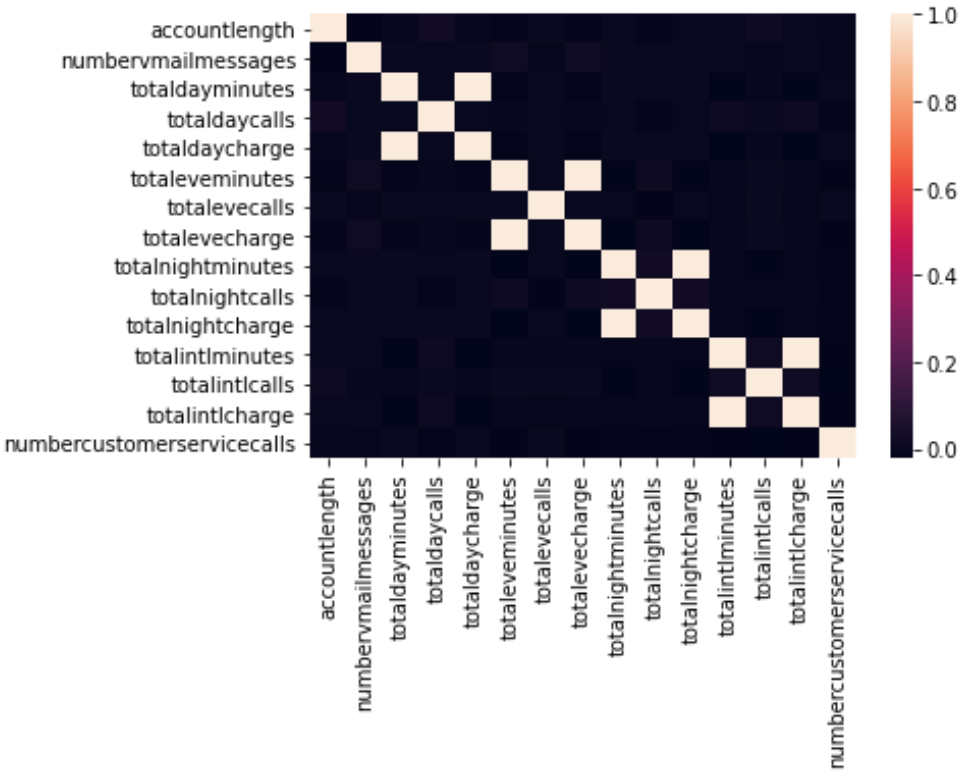
구현 방법

1. Data preview & preprocessing

먼저 data.world에서 이탈고객 데이터 5천건을 가져와서 각 변수를 확인한다.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   churn                   5000 non-null   object
 1   accountlength           5000 non-null   int64
 2   internationalplan       5000 non-null   object
 3   voicemailplan           5000 non-null   object
 4   numbervmailmessages     5000 non-null   int64
 5   totaldayminutes         5000 non-null   float64
 6   totaldaycalls           5000 non-null   int64
 7   totaldaycharge          5000 non-null   float64
 8   totaleveminutes         5000 non-null   float64
 9   totalevecalls           5000 non-null   int64
 10  totalevecharge          5000 non-null   float64
 11  totalnightminutes       5000 non-null   float64
 12  totalnightcalls         5000 non-null   int64
 13  totalnightcharge        5000 non-null   float64
 14  totalintlminutes        5000 non-null   float64
 15  totalintlcalls          5000 non-null   int64
 16  totalintlcharge         5000 non-null   float64
 17  numbercustomerservicecalls  5000 non-null  int64
dtypes: float64(8), int64(7), object(3)
```

데이터는 총 18개의 변수로 구성되어 있고, churn은 이탈 고객인지 아닌지를 나타내는 타겟 변수이다. 변수의 데이터 타입을 보면 churn과 international plan, voice mail plan 중에는 yes, no로 구성된 범주형 변수가 있기 때문에 yes를 1, no를 0으로 각각 변환하였다.



각 변수가 독립인지 확인하기 위해 상관관계 히트맵을 확인하였다. 히트맵을 보면 total day minutes과 total day charge, total eve minutes과 total eve charge, total night minutes과 total night charge, total intl minutes과 total intl charge에 의존관계가 있다.

따라서 total day charge, total eve charge, total night charge, total intl charge 변수를 제거하고, 타겟변수 churn을 y로, 다른 변수들을 X로 설정했다.

타겟 변수 churn의 데이터를 확인해보니, 이탈 유저가 약 14%로 데이터 불균형이 존재하였다.

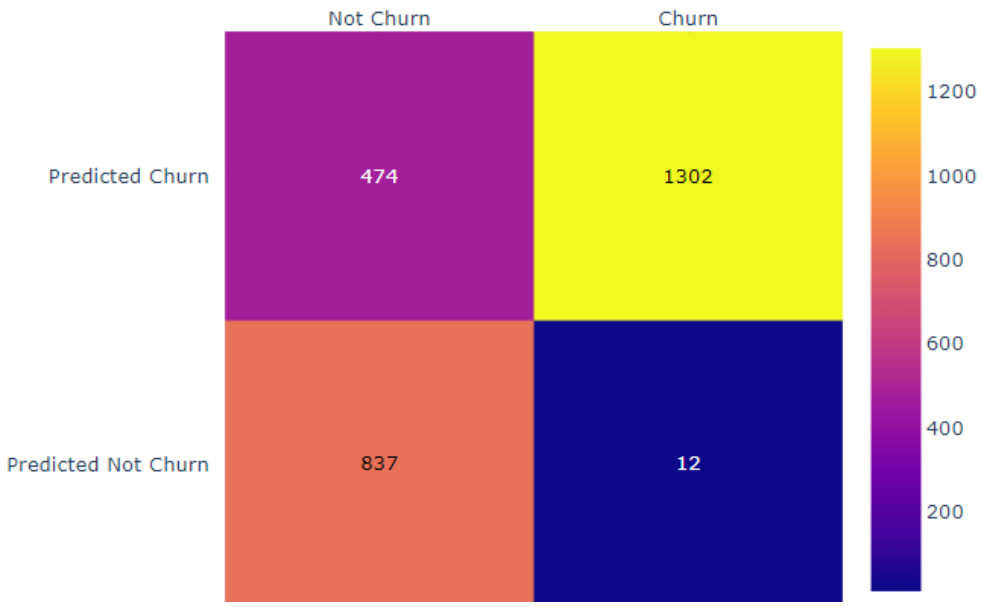
```
0    0.8586
1    0.1414
Name: churn, dtype: float64
```

관심있는 데이터는 이탈한 고객이기 때문에, 이탈 고객 데이터를 adasyn 기법으로 over sampling 하였지만, 데이터 불균형이 모델에 성능을 얼마나 영향을 끼치는지 확인하기 위해 불균형 데이터 그대로 사용하여 모델링 해본 후, resampled data로 다시 진행하여 성능 차이를 확인하였다. 아래에 나온 모델링 결과는 resampled data의 분석 결과이다.

2. Modelling

모델은 KNN, Naive Bayes, SVM, Decision Tree, Random Forest까지 수업시간에 다루었던 모델들을 사용하여 각각 성능을 측정하였다. 불균형 데이터이기 때문에 성능 측정 방법으로 Accuracy를 사용하지 않고, Confusion Matrix와 F1 Score, AUC-ROC score를 사용하였다.

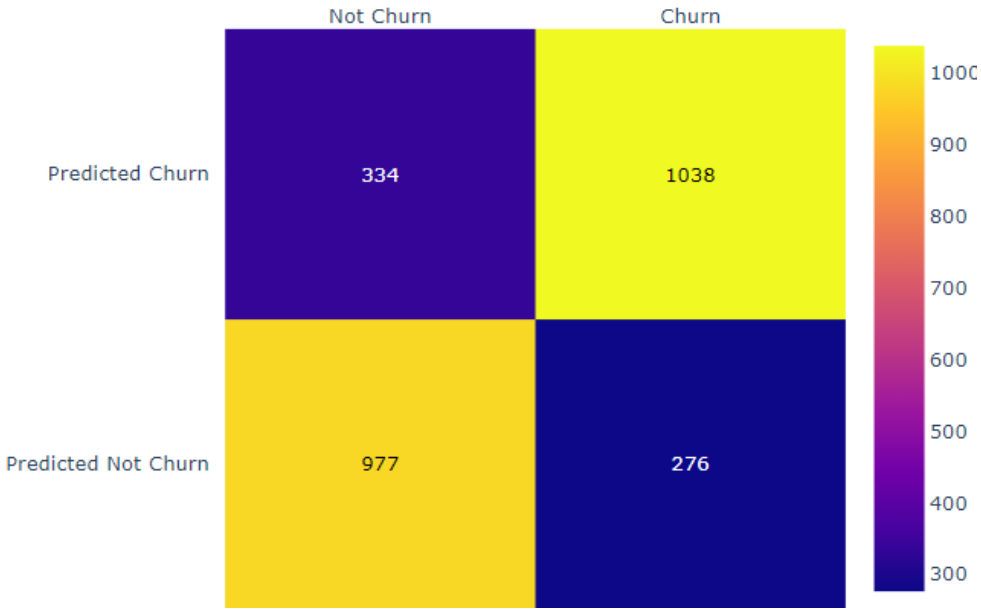
2.1. KNN



roc auc score: 0.81

f1 score: 0.84

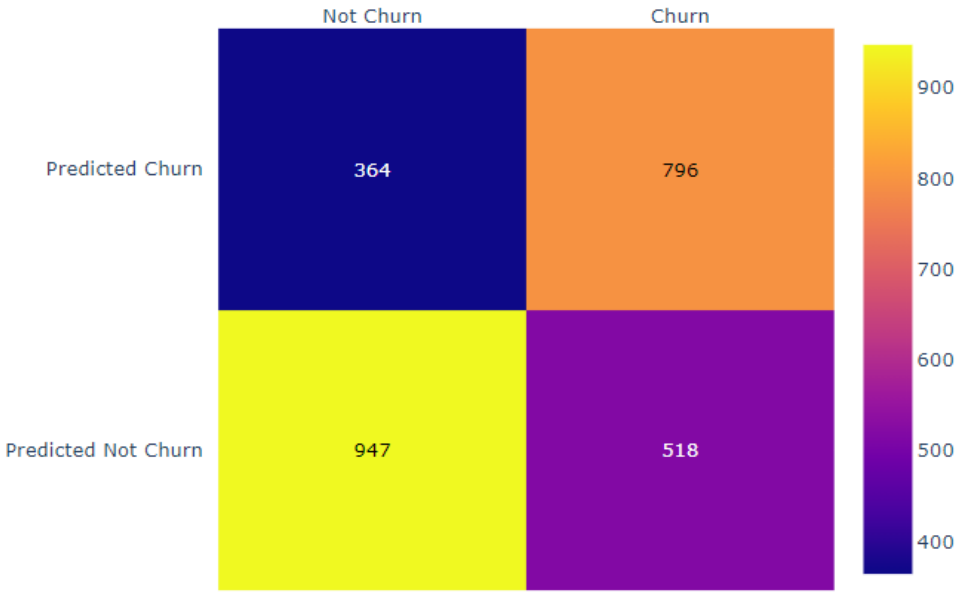
2.2. Naïve Bayes



auc score: 0.77

f1 score: 0.77

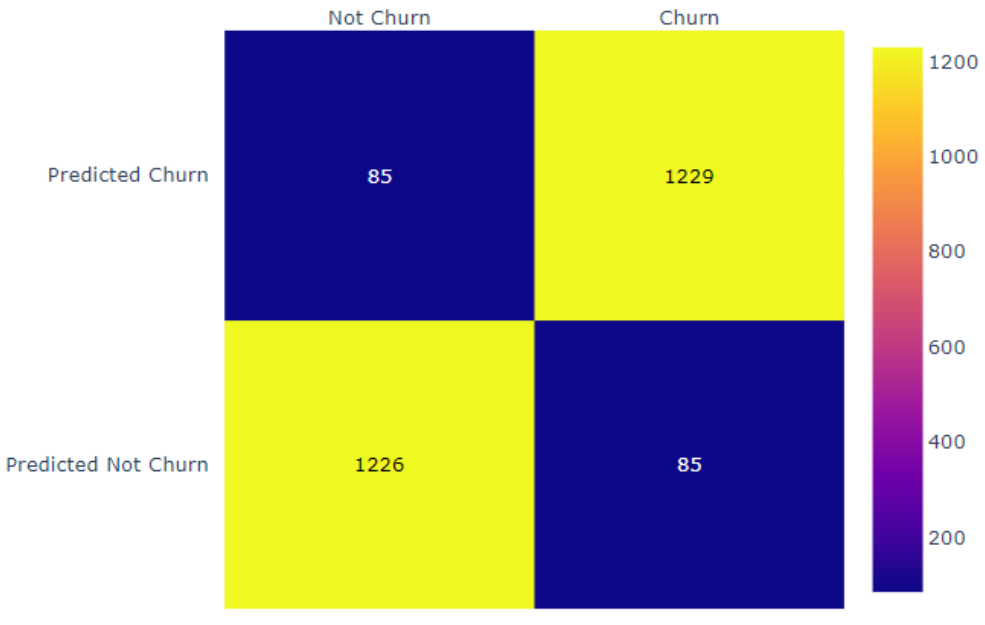
2.3. SVM



auc score: 0.66

f1 score: 0.64

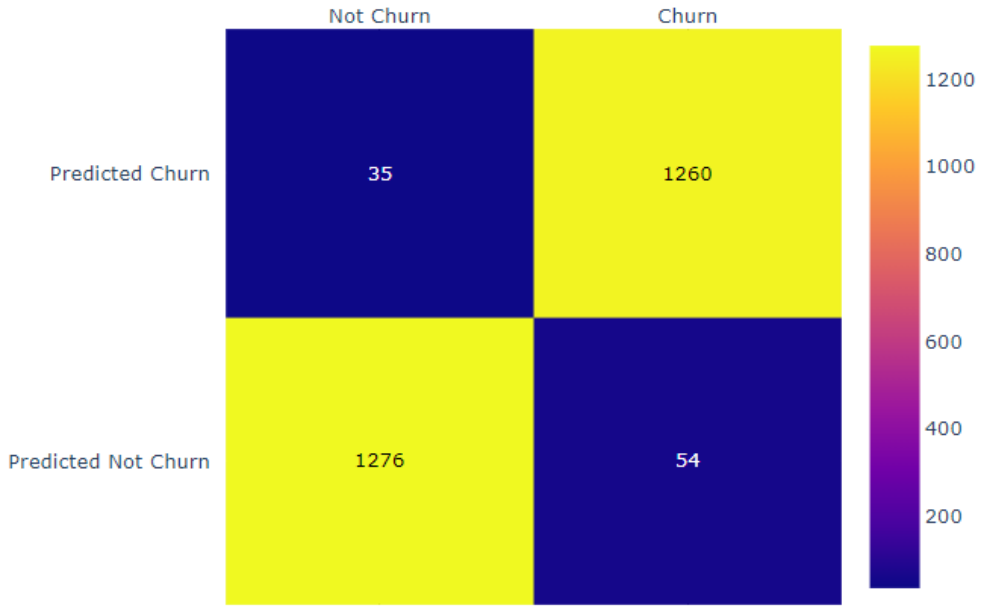
2.4. Decision Tree



auc score: 0.94

f1 score: 0.94

2.5. Random Forest



auc score: 0.97

f1 score: 0.97

3. Feature importance

랜덤 포레스트 모델로 변수 별 중요도를 측정하였는데, 그 결과는 다음과 같다. 국제 요금제를 사용하는 고객이 가입하지 않은 고객에 비해 이탈률이 높았고, 고객 서비스 전화를 많이 하는 고객이 이탈할 확률이 컸다. 하루 통화를 오래하는 고객의 이탈률이 높다는 점이 흥미로웠는데, 그 원인에 대한 연구도 추후에 진행할 가치가 있다.

feature	importance
internationalplan	0.217400
numbercustomerservicecalls	0.195291
totaldayminutes	0.155227
totalintlcalls	0.071485
totaleveminutes	0.069428
voicemailplan	0.048539
totalintlminutes	0.046294
totalnightminutes	0.040379
accountlength	0.032907
totaldaycalls	0.030954
totalnightcalls	0.030888
numbervmailmessages	0.030716
totalevecalls	0.030491

결론 & 고찰

불균형 데이터로 성능을 평가했을 때에 비해 오버샘플링을 진행한 후에 모델의 성능이 크게 증가했다. 랜덤포레스트 모델과 의사결정트리에서 가장 좋은 성능을 보였는데, 가장 놀라운 점은 KNN이 F1 score와 AUC-LOC score는 그렇게 높지 않지만 confusion matrix에서 이탈 고객 예측에서는 가장 좋은 성능을 나타냈다는 점이다. 모델링을 통해 실제 이탈 고객을 판단하는 것이 가장 큰 목표였는데, KNN 모델이 어떻게 이탈 고객을 분류했는지에 대해서도 더 구체적으로 분석해 볼 가치가 있다.

통신사 같은 레드 오션 시장에서는 새로운 고객을 유치하는 것보다 기업의 비용을 줄이는 것이 회사 성장에 더 큰 역할을 한다는 연구를 보았다. 특히 통신사 같이 고객의 풀이 매우 클 경우에는 한번 프로모션을 할 때마다 천문학적인 비용이 발생하는데, 이러한 모델링을 통해 이탈할 것으로 분류되는 고객에게만 프로모션을 진행할 수 있다면, 비용을 효과적으로 줄일 수 있을 것이라고 생각한다.

또한 이미 이탈한 고객들에게도 변수 중요도 별로 마케팅 전략을 짜서 재가입을 유도할 수도 있을 것 같다. 가령 통화를 많이 하던 이탈 고객에게 통화 요금 할인에 대한 광고를 한다면, 아예 새로운 고객에게 하는 광고보다 훨씬 효율적일 것이다.

이 분석을 진행했던 가장 큰 이유는 데이터만 바꾸면 은행이나 카드사, 보험사 등 고객을 보유한 많은 도메인에서 비슷하게 진행할 수 있다는 점이었다. 해당 데이터셋을 구하는 것이 어렵겠지만, 추후에 어떤 경로로 데이터를 얻을 수만 있다면 이 프로젝트 경험으로 쉽게 모델링을 진행할 수 있을 것 같다.