모델 설계 과정.

1. 어떤 변수를 선택해야 모델의 성능을 높일 수 있을까?

과정1) 모델링에 사용할 수 없는 변수를 버리자!

=> 삭제한 컬럼 : date 형식으로 되어 있는 컬럼(연, 월, 일) => "도착 월" 데이터만 범주형으로 남기고 모두 삭제.(모델링에 사용할 수 없기에 삭제.)

=> 누가 봐도 쓸데없는데, 범주형이어서 더미화하면 컬럼이 너무 늘어나는 것들을 없애자! => reserved\_room\_type, asigned\_room\_type.

과정2) 변수의 속성을 변경하자!

=> 자녀들의 특성을 children, babies로 나누어 놓았는데, 자녀의 수 보다는 자녀의 유무 특성이 더 유효할 것으로 판단, children과 babies속성을 삭제하고 자녀의 유무를 0과 1로 나타내는 kids 컬럼을 추가.

=> 총 방문인원을 나타내는 total\_guest 열을 추가하고, 각 방문객의 연령층을 나타내는 adults, children, babies열은 삭제.

=> agent열은 숫자형이지만, 여행사의 id를 나타내는 사실상 범주형 특성임. 또한 여행사의 개수가 많아 이를 그대로 사용하는 것은 모델링에 적합하지 않다고 판단하여, agent를 여행사 이용 여부를 0과 1로 나누는 특성으로 변경하였음.

과정 3) 후진제거법을 활용하자! => RFE 방법을 활용하였으나, 이 방법을 통해 선택한 변수들을 채택하여 모델링 하였음에도 성능에 큰 변화가 없었음.

과정 4) rf모델링 후 변수중요도 순으로 특성을 제거해 보자!

=> 변수 중요도 별로 특성을 drop하며 성능 체크를 하였으나, 성능에 큰 변화가 없었음.

최종적으로 결정한 내용: 특성들 간의 상관관계가 크지 않고, 모듈을 활용한 방법을 몇 가지 적용하여 보았으나 의미있는 수준의 성능 개선이 이루어 지지 않았음.

=> 따라서 최종적으로, 우리가 생각했을 때 유의미하다고 판단되는 변수를 선택하여 일반 사람들에게 설명력이 있으면서도 어느 정도 성능을 이끌어 낼 수 있는 모델을 구현하고자 함.

==> 최종 특성선택 과정:

특성 선택 기준

1. 가장 직관적으로 취소에 영향이 있을 것이라고 생각되는 데이터들을 선정.

2. 모델에 반영하기 쉬운 이산형 데이터를 우선적으로 고려하였음.

3. 범주형일 경우, unique 값이 너무 많은 특성은 탈락시킴.

=> 이 세 가지 기준에 따라 변수 중요도를 상, 중, 하로 나누고, 최종 회의를 통해 16개의 설명변수를 선택

2. 어떤 식으로 데이터 스케일링을 할 것인가?

target 값의 분포가 비슷하지 않아 모델의 일반화 성능을 높이기 위해서 오버샘플링 진행.

train 값 0.8, test 값 0.2 의 비율로 맞춰 놓은 후 진행.

수학적 모델의 성능을 높이기 위해 스케일링 => standard scaling

2. 어떤 분류기 모델을 사용할 것인가?

다양한 분류기 모델을 적용해 봄.

=> 트리 계열 분류기(dt, rf, gbrt)에서는 매개변수값에 따라 오버피팅 문제 발생.

=> 최근접 이웃 모델(knn)역시 오버피팅 문제 발생.

=> 서포트 벡터 머신 분류기(SVC)는 오버피팅이 크게 발생하지 않고, 어느 정도의 성능을 내지만, 분류에 걸리는 시간이 너무 오래 걸림. 따라서 svc는 모델 활용에 있어서 문제가 있을 것으로 예측됨.

=> lr은 오버피팅이 발생하지 않고, 모델링에 걸리는 시간도 길지 않지만, 성능이 다른 모델들에 비해 낮았음.

=> GridSearch를 통하여 각 모델의 매개변수값 조정.

=> 교차검증을 통하여 각 모델별 최적 매개변수 모델의 안정성 확인.

=> 결론: 트리 계열 분류기의 성능이 준수하면서 시간도 적정하게 걸렸으므로, 매개변수 값의 조정을 통해 오버피팅 문제를 해결하기로 결정.

=> 처음에는 매개변수 그리드서치를 통해 찾으려고 했으나, 성능이 높아지는 동시에 오버피팅이 발생하는 현상 발생.

=> 매개변수값에 따른 훈련 스코어/검증 스코어를 확인하면서 오버피팅이 발생하지 않으면서 성능을 끌어올리는 매개변수값 선택.

### 각 모델에 대한 평가

1. 수학적 분류기 모델(수학적 분류기 모델들은 standard scaling 된 값을 기준으로 함.)
2. svc

훈련 데이터 점수 : 0.7790054749471478

평가 데이터 점수 : 0.7998157299606332

장점: 오버피팅 별로 안 나면서도 성능이 준수함.

단점: 모델이 너무 무거움.(시간이 오래 걸림.)

특별한 매개변수값 조정 없이도 데이터 스케일링만으로도 충분한 성능을 보여 주었기 때문에 수학적 분류기 모델에서는 SVC를 best모델로 선정하였다.

1. KNN

훈련 데이터 점수 : 0.869486655945648

평가 데이터 점수 : 0.7869587067593601

장점: 매개변수 값 따로 주지 않고서도 잘 돌아감.

단점: 오버피팅, 이것도 모델이 무거움.(시간이 오래 걸림.)

오버피팅 문제가 발생하고, 그리드 서치를 수행하기에 모델이 무거움.

1. 로지스틱 회귀

훈련 데이터 점수 : 0.7401837026210012

평가 데이터 점수 : 0.7609096239216014

장점: 훈련/평가 데이터 점수간의 차가 크지 않으면서도, 다른 수학적 분류기 모델들에 비해 빠르게 돌아감.

단점: 매개변수값 조정해도 성능 항상 제자리.

들어가는 비용(시간, 메모리) 대비 효율이 좋은 모델이어서 채택하려고 했으나, 매개변수 값을 조정해도 성능에 변화가 없어 본 연구데이터에 최적의 모델이 아니라고 판단하였음.

2. 트리 계열 분류기

1. 결정트리

train set score is 0.990

test set score is 0.784

장점: 높은 분류예측 성능.

단점: 오버피팅 심하게 발생

매개변수 조정을 통해 오버피팅 문제 해결 가능.

1. 랜덤포레스트

train set score is 0.990

test set score is 0.812

장점: 높은 분류 예측 성능. dt보다는 오버피팅 덜 발생함.(하지만 그래도 오버피팅 문제 심함.)

단점: 오버피팅.

1. 그래디언트 부스팅(?)

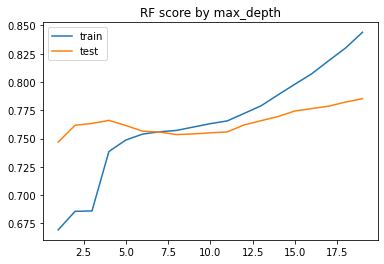
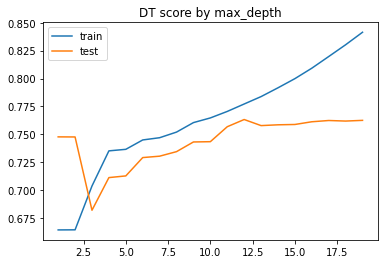
0.689942881811842

0.765516374905771

장점: 오버피팅 발생 x

단점: 다른 트리 계열 분류기 모델보다 성능 낮음. n\_estimator 값 300까지 올려도 성능이 높지 않음

DT, RF모델의 매개변수값 조정



RF모델의 max\_depth가 12~15 정도에서 가장 준수한 성능을 보이는 것을 확인 가능. dt보다 rf 모델이 근소하게 더 높은 test 성능을 보임.

max\_depth = 15, 나머지 값은 default로 두고 랜덤포레스트 모델 분류예측 진행 했을 시 성능

train set score is 0.798

test set score is 0.774

와 이상적이다!

cross\_val\_score활용한 교차 검증 검수 확인

교차 검증 점수 : [0.7856876 0.78983853 0.78389373 0.79032794 0.78742217]

교차 검증 점수 평균: 0.7874339931413028

교차 검증 점수의 편차 크지 x.

일반화도 어느 정도 잘 되고 있다는 판단 가능.

dt 도 마찬가지.

그리드 서치 결과 선택된 criterion = ‘gini’, ‘splitter = ‘best’ 로 두고,

max\_depth는 위 그래프에서 확인된 훈련/테스트 점수 차이가 크지 않은 12로 두고 돌렸을 때,다음과 같은 결과 도출.

train set score is 0.777

test set score is 0.763

교차 검증 점수 : [0.76800465 0.76812918 0.76309672 0.77065172 0.77102532]

교차 검증 점수 평균: 0.7681815181915658

랜덤포레스트와 비슷함. dt가 훨씬 가벼운 모델이라는 점을 고려했을 때, 상당히 좋은 모델이라고 할 수 있음.

결론: dt, rf, svc를 best 취소예측 분류모델로 선정. 단, svc는 비용 대비 효율이 떨어지므로 dt, rf 모델이 최적 모델로 여겨집니다.

마무리 =>

우리가 만든 호텔 예약 취소 예측 분류모델이 호텔 경영에 도움이 될 것.

취소 분류 예측이 가능해 질 경우 창출 가능한 취소 방지 서비스

1. 예약 취소 확률이 높은 고객에게 예약 확인 연락을 더 자주 넣기.
2. 예약 취소 확률이 높은 고객에게 재방문 할인 쿠폰 등 반드시 방문할 수 있도록 하는 유인책 제시하기.
3. 예약 취소를 막기 위해, 예약 취소 확률이 높은 고객에게 호텔 무료 조식 등의 이벤트 제시하기.
4. 취소율이 높은 고객을 많이 데리고 오는 여행사 또는 그룹에 대하여는 예약 시 예치금을 올리거나, 캔슬 발생시 환불 정책을 다르게 적용하는 등의 대비 가능.
5. 예약 취소 확률이 높은 고객의 예약이 대량 발생하는 경우에 미리 방 예약률을 높이기 위한 이벤트를 기획함으로써 공실 발생률 낮추기.

=> 이런 서비스를 가능하게 하는 분류모델의 디자인이 가능하니, 이 분류모델을 기반으로 하여 호텔 경영관리 향상에 도움을 주는 서비스 개발도 가능할 것.(ex. 취소가 예상되는 손님에게 자동으로 예약 확인 문자를 더 자주 보내는 자동화 시스템 등.)