Airbnb

에어비엔비를 시작하는 사람들에게

19조 박현서, 이채은

목차

분석 목표

주제 선정 배경

에어비앤비 사업자 입장에서, 숙소사업 성공여부를 예측하고자 함
성공 여부를 판단하기 위해서 평점을 기준으로 좋은 숙소와 안좋은 숙소로 구분

분석 대상

좋은 숙소(GOOD)
안 좋은 숙소(BAD)
구분

분석 방법

1. KNN
2. LDA/QDA
3. 로지스틱 회귀분석
4. Neural Network
5. Decision Tree
6. Ensemble

변수 및 데이터 설명

● 관측치(숙소): 53,736개, 변수: 20개 출처 : http://insideairbnb.com/get-the-data.html

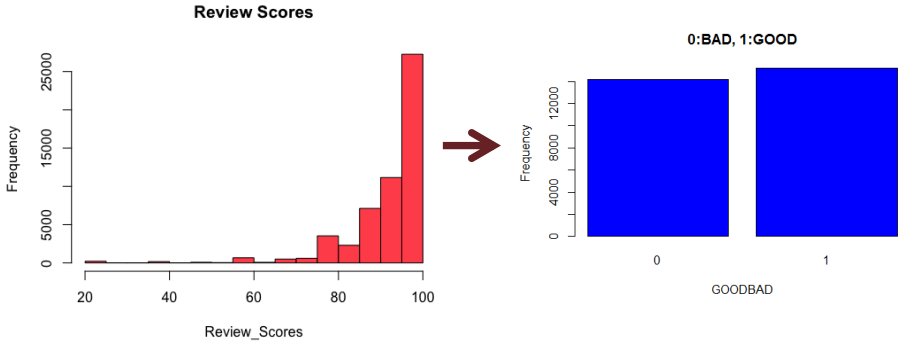
	변수	변수설명
반응변수	후기 점수 가격	좋은 숙소와 안좋은 숙소의 기준으로 사용 binary(1:good, 0:bad)
수치형변수	수용가능인원 침실 개수 욕실 개수 침대 개수 최소 숙박일 이용가능일 (기준:1년) 총후기 개수 월 평균 후기 개수	
범주형변수	슈퍼호스트 여부 제안된 숙소 유형 호스트의 프로필 사진 유무 호스트의 신상 확인 여부 숙소 종류 침대 종류 즉시 예약 가능여부 예약 취소 정책 에어비앤비 숙소 등록 년도 위치	binary(1:true, 0:false) 가족, 비즈니스, 커플, 사회, 해당없음 binary(1:true, 0:false) binary(1:true, 0:false) 집전체, 개인실, 공유객실 에어배드, 소파, 이불포대, 침대소파, 침대 binary(1:true, 0:false) 유연, 일반, 엄격(유예기간), 엄격, 매우엄격30일, 매우엄격60일 08,09,10,11,12,13,14,15,16,17,18년도 33개 지역구 이름

데이터 전처리 및 변수 재정의

29,362 obs

좋음/안좋음

후기 점수를 잘 주는 경향
'후기 점수'변수를 '좋음/안좋음'변수로 새롭게 정의
(중앙값 95점 기준) => GOOD/BAD 비슷한 비율



지역

위치에 따라 숙소의 특성이 달라질 것이라 예상
예) 1박당 100파운드가 넘는 숙소의 비율이 매우 차이남

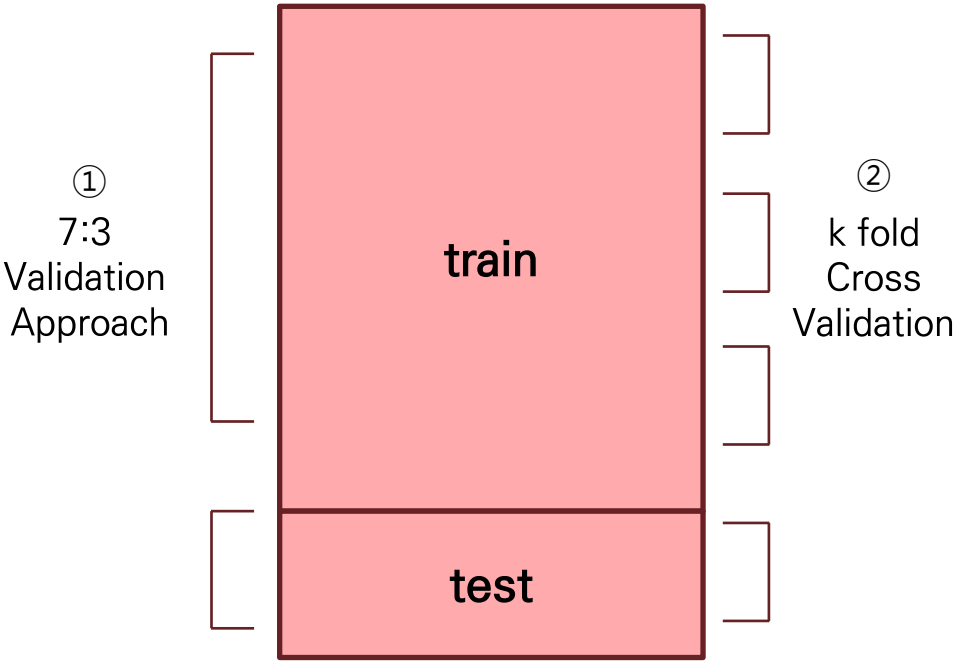
중심부에 더 많은 관측치, 분석결과를 다른 대도시에도 적용해볼 수 있도록 중심부만 추출

총 후기 개수

후기가 너무 적은 관측 값들의 점수는 신빙성이
떨어지므로 후기개수 3 이상인 관측치 추출 후 분석 진행

분석개요

Model Fitting



Step 1

Validation set Approach (7:3)로
민감도, 특이도, 예측 정확도 등을 고려

Step 2

K fold Cross Validation로
cv test error 확인하여 최종 모델 선택

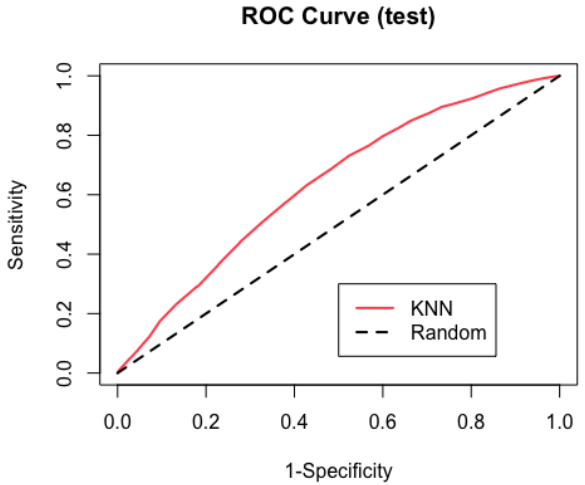
1. KNN

수치형 변수로만 모델 적합
(수용가능인원, 침실 개수, 욕실 개수, 침대 개수, 최소 숙박일,이용 가능 일, 월 후기 개수, 가격)
따라서 범주형 변수의 정보를 잃는다는 한계가 있음

적절한 k값 선택

	Cutoff =0.5	K=10	K=40	K=50	K=60
Validation set 7:3	Sensitivity	0.619	0.700	0.708	0.715
	Specificity	0.553	0.509	0.498	0.489
	Misclassification	0.413	0.392	0.392	0.394
	Prediction Accuracy	0.587	0.608	0.607	0.606
Fold=5	Test error	0.412	0.392	0.390	0.390

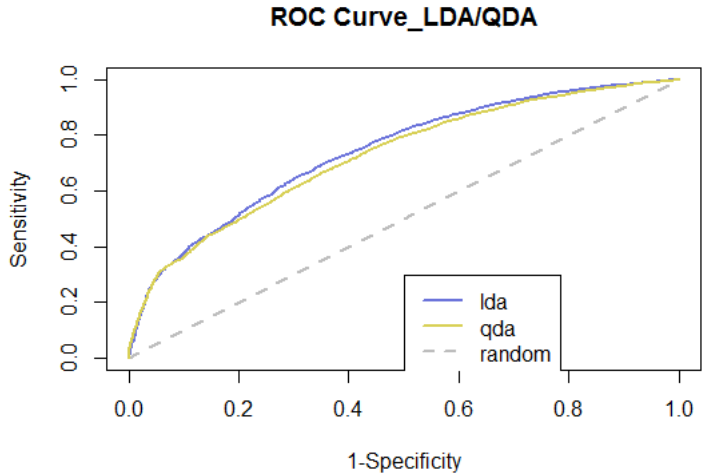
- 특이도가 많이 낮아지지 않으면서 test error가 줄어드는 K=50으로 선택
  test AUC = 0.635



2. LDA/QDA

수치형 변수 + binary 변수로 모델 적합
(수용가능인원, 침실 개수, 욕실 개수, 최소 숙박일, 이용 가능 일, 월 후기 개수, 가격
+ 슈퍼호스트 여부, 프로필 사진 유무, 신상 확인 여부, 즉시 예약 가능 여부)

		LDA	QDA
Validation set 7:3	Sensitivity	0.626	0.576
	Specificity	0.714	0.730
	Misclassification	0.330	0.348
	Prediction Accuracy	0.670	0.651
Fold=5	Test error	0.332	0.387



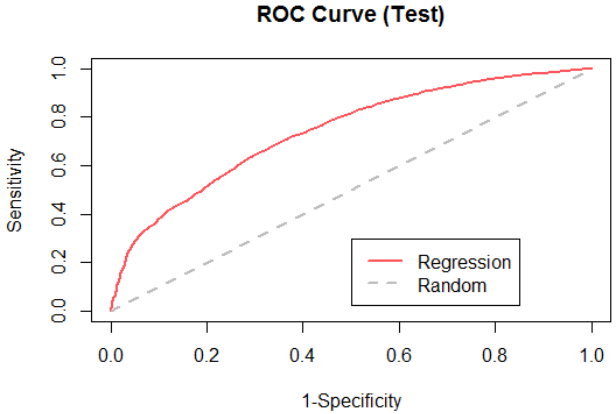
- Prediction Accuracy : LDA 〉 QDA
  test error : LDA 〈 QDA
  test AUC : LDA 〉 QDA

- 민감도, 예측 정확도, test error 모두 고려하여 LDA 모델을 채택
  test AUC_LDA = 0.740, test AUC_QDA = 0.725

3. 로지스틱 회귀분석

모든 변수로 모델 적합
(제안된 숙소 유형, 슈퍼호스트여부, 프로필사진 유무, 신상 확인 여부, 숙소 종류, 수용가능인원, 욕실개수, 침실 개수, 침대 종류, 가격, 최소 숙박 일, 이용가능일(기준:1년), 즉시 예약 가능여부, 예약 취소 정책, 월 후기 개수)
변수 제외 – 욕실 수, 프로필 사진 유무 제외

		Cutoff=0.5	Cutoff=0.3
Validation set 7:3	Sensitivity	0.638	0.937
	Specificity	0.705	0.262
	Misclassification	0.329	0.395
	Prediction Accuracy	0.670	0.605
Fold=5	Test error	0.331	



- 민감도와 특이도가 균형적으로 높고, 예측정확도가 높은 cutoff=0.5모델 선택
- test AUC = 0.740

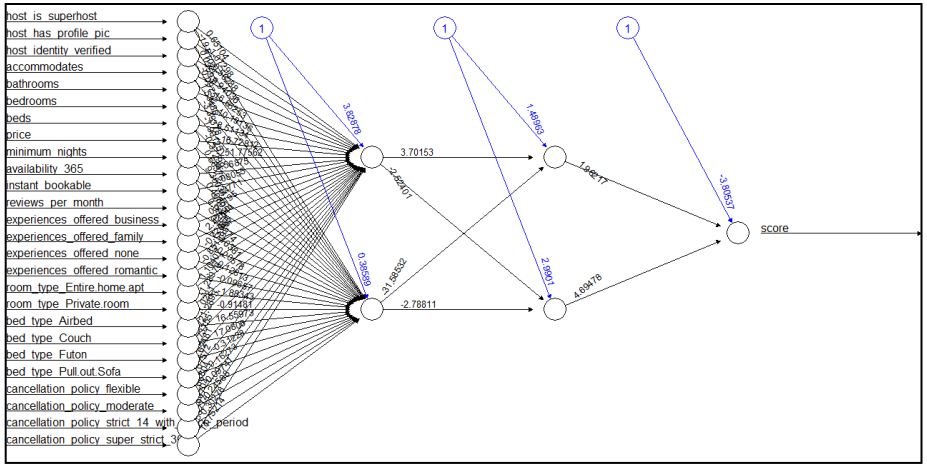
〈Fitted Model〉
침실 수 많을수록 GOOD
슈퍼호스트 일수록 GOOD
Pull-out sofa나 futon (소파와 침대 겸용) 일수록 GOOD

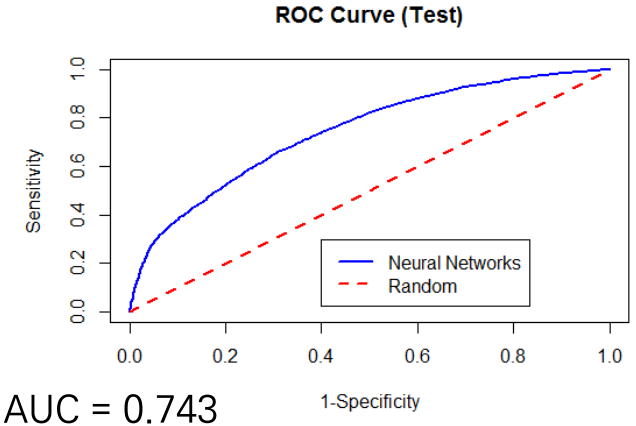
4. Neural Network

	(layer, node,threshold)	(2,2,0.1)	(2,4,0.1)	(2,2, 0.09)	(3,4, 0.09)
Validation set 7:3	Sensitivity	0.696	0.690	0.715	0.688
	Specificity	0.650	0.657	0.628	0.650
	Misclassification	0.326	0.326	0.328	0.331
	Prediction Accuracy	0.674	0.674	0.672	0.669
Fold=5	Test error			0.326	

Computing power를 고려했을 때, 예측 정확도가 떨어지지 않는 선에서 민감도를 높이려 함

Actual \ Predicted	0	1
0	2707	1606
1	1271	3194

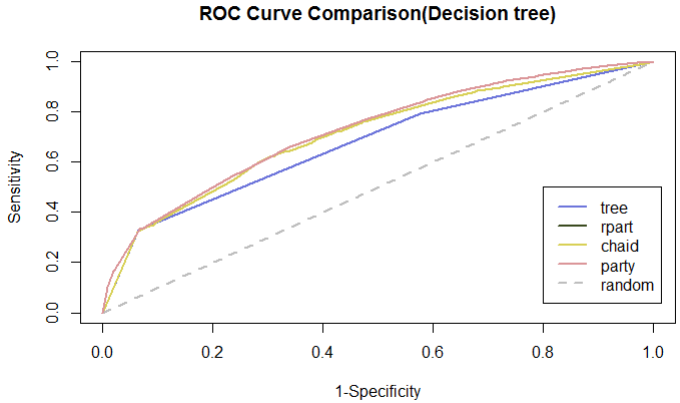




● test AUC = 0.743

5. Decision Tree

	Cutoff =0.5	ctree	rpart	chaid	party
Validation set 7:3	Sensitivity	0.328	0.640	0.512	0.664
	Specificity	0.935	0.672	0.790	0.656
	Misclassification	0.380	0.344	0.354	0.340
	Prediction Accuracy	0.620	0.655	0.646	0.660
Fold=5	Test error	0.380	0.330	0.351	0.338

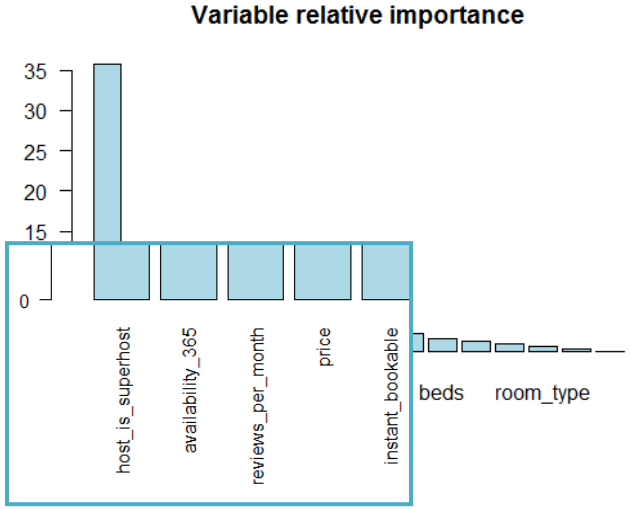


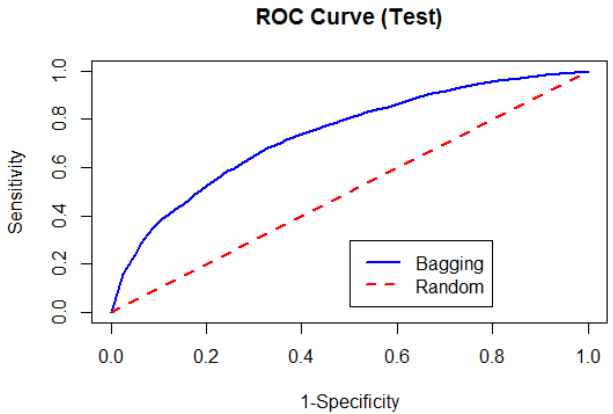
ctree: ***슈퍼호스트여부, 즉시 예약가능 여부만 사용됨*** train AUC = 0.686, test AUC = 0.678
rpart: ***호스트 프로필사진 등록 여부만 제외하고 모두 사용됨*** train AUC = 0.742, test AUC = 0.725
chaid: ***범주형 변수만 사용됨*** train AUC = 0.719 ,test AUC = 0.714
party: ***모든 변수가 사용됨*** train AUC = 0.742, test AUC = 0.725

- rpart 알고리즘의 cv test error가 가장 낮고 민감도와 특이도가 균형 있게 높아 최적이라고 판단
- test AUC = 0.725

6. Ensemble_Bagging

		Cutoff=0.5
Validation set 7:3	Sensitivity	0.690
	Specificity	0.656
	Misclassification	0.327
	Prediction Accuracy	0.673
Fold=5	Test error	0.323

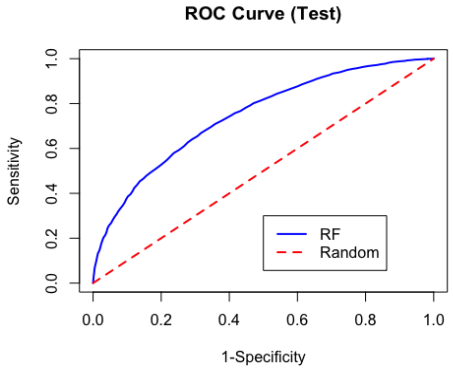


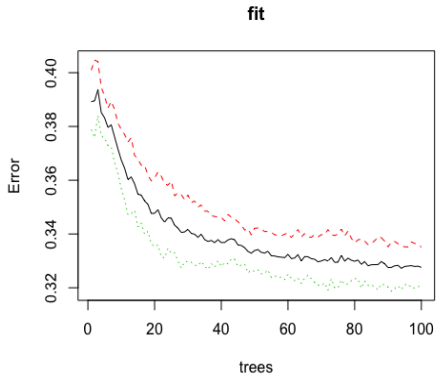


- 민감도가 0.69이고 예측 정확도가 0.673으로 비교적 적정수준 이상이라 판단
- 중요한 변수: 슈퍼호스트 여부, 이용가능일, 월 후기 개수, 가격
- test AUC = 0.736

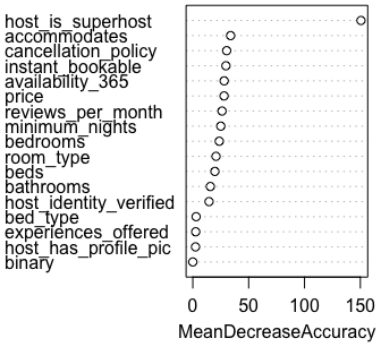
6. Ensemble_Random Forest

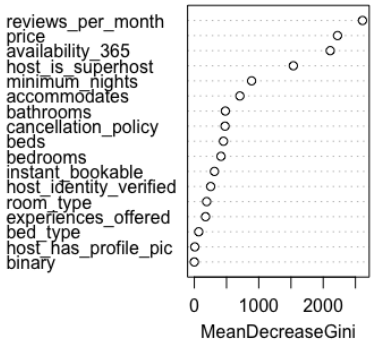
		Cutoff=0.5
Validation set 7:3	Sensitivity	0.682
	Specificity	0.668
	Misclassification	0.324
	Prediction Accuracy	0.676
Fold=5	Test error	0.326





Variable Importance Plot

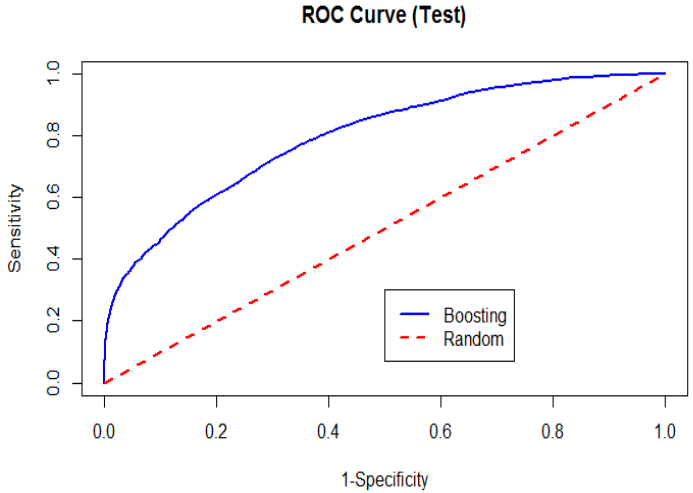




- 민감도, 특이도 모두 균형있게 높고, cv test error는 0.326으로 낮은 편
- 중요한 변수: 슈퍼호스트여부, 월 후기개수, 가격, 이용가능일 ,최소숙박일, 수용가능인원
- test AUC = 0.745

6. Ensemble_Boosting

	(maxdepth, mfinal)	(6,50)	(8,50)	(10,50)
Validation set 7:3	Sensitivity	0.696	0.713	0.759
	Specificity	0.691	0.709	0.730
	Misclassification	0.306	0.289	0.255
	Prediction Accuracy	0.694	0.711	0.745
Fold=5	Test error	0.319	0.319	0.323



- 민감도와 특이도를 높이면서 cv error가 낮은 maxdepth=8 로 선정
  cv train error: 0.275, test error: 0.319
  큰 차이가 없어서 과적합 문제도 일어나지 않았다고 봄
  cf) Maxdepth=10인경우
  train error: 0.238, test error: 0.323로 꽤 차이가 남
- test AUC: 0.796

결론

		Prediction Accuracy	Sensitivity	Specificity	CV. error.test	Test AUC
KNN	K=50	0.607	0.708	0.498	0.390	0.635
LDA		0.670	0.626	0.714	0.332	0.740
QDA		0.651	0.576	0.730	0.387	0.725
Logistic		0.670	0.638	0.705	0.331	0.740
NN	(2, 2, 0.09)	0.672	0.715	0.628	0.326	0.743
Decision Tree	rpart	0.655	0.640	0.672	0.330	0.725
Ensemble	Bagging	0.673	0.690	0.656	0.323	0.736
	Boosting	0.711	0.713	0.709	0.319	0.796
	Random Forest	0.676	0.682	0.668	0.326	0.745

* : 다른 분석기법에 비해 상대적으로 높은 값을 의미

- LDA, Logistic 은 비슷한 수준의 에러율과 예측력을 가지며 민감도와 특이도가 균형있게 높다.
  Logistic이 변수의 영향력을 해석 할 수 있다는 점에서 이 두 모델 중에서는 ***Logistic을 선택*** 할 것이다.
- NN, Decision Tree, Ensemble을 비교하면 Boosting기법의 에러율이 가장 낮고 예측력이 가장 높다.
  따라서 예측을 위해서 모델을 선택할 때는 앙상블 중에서도 모든 부분에서 최적의 값을 가진***Boosting기법을 선택*** 할 것이다.

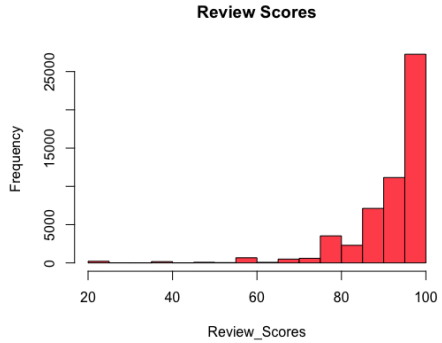
결론

- 모델을 해석하고 각 변수의 영향력을 알고자 하는 경우 최적의 모델은 ***Logistic***이다
  : 앙상블의 variable importance plot을 이용하여 변수의 중요도를 파악하는 정도는 가능하지만 로지스틱 회귀 모형은 계수 값으로 더 자세한 정보를 알 수 있기에 상대적으로 해석에 있어서 좋은 모델이다.

- 예측력이 우선적인 경우 최적의 분석기법은 앙상블 중 ***Boosting*** 이다.

- 해석이 가능한 분석 기법들을 종합해 보았을때
  슈퍼호스트여부, 월 후기개수, 가격, 이용가능 일 변수가 숙소 사업성 판단에 중요하다.

한계점



- Rating을 Good과 Bad로 나누는 기준이 임의로 결정되었다.
  : 90점 이상의 좋은 평점이 굉장히 많아 점수의 분포가 치우쳐져 있었다. 그래서 그 중에서도 누구나 만족할 만한 평점을 정하기 위해 그래프를 보고 기준을 세웠지만 이 역시 자의적인 해석이 들어갔다.

- 총 이용자 수와 유지비용 등에 대한 정보가 없어서 수익성을 예측하는데 한계가 있다.
  : 사업자의 주된 관심사일 것인 수익성에 관련한 정보가 부족하다. Rating이 수익에 영향을 미치겠지만, 수익 관련 정보가 없었기에 수익성으로 곧바로 이어지는 결과는 아니다.