Airbnb

에어비엔비를 시작하는 사람들에게

19조 박현서, 이채은

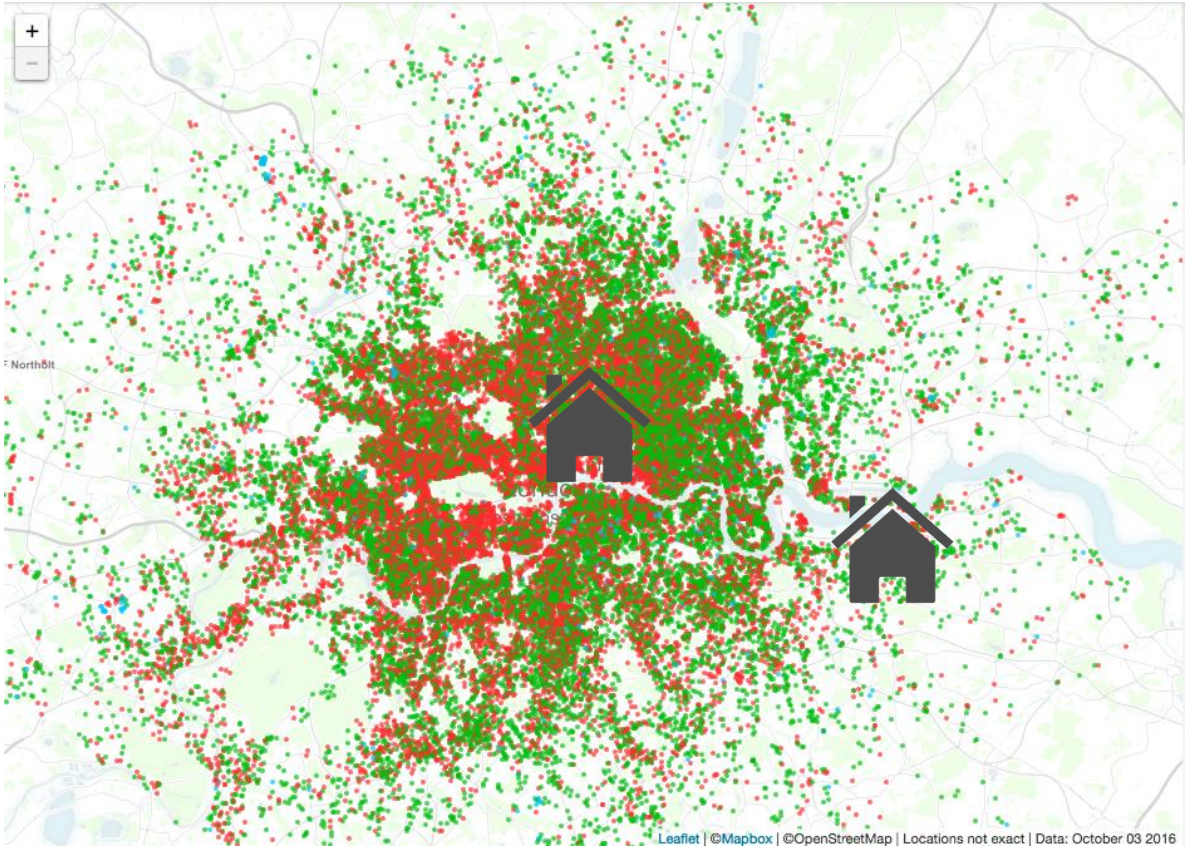
목차

53,736 | 20 | Airbnb data | London

데이터 선정 배경

London

서론



데이터 선정 배경

에어비앤비 호스트 입장에서

서론

1. 에어비앤비 사업을 새로 시작하고 싶은데, 내 집이 좋은 평점을 받을 수 있을까?
➡ KNN, LDA/QDA, 로지스틱 회귀분석으로 좋은 숙소와 안좋은 숙소를 구별

2. 숙소의 조건에 따른 1박당 합리적인 가격은 얼마일까?
➡ 다중회귀분석으로 가격 해석

데이터 및 변수 설명

서론

● 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

관측치 제외

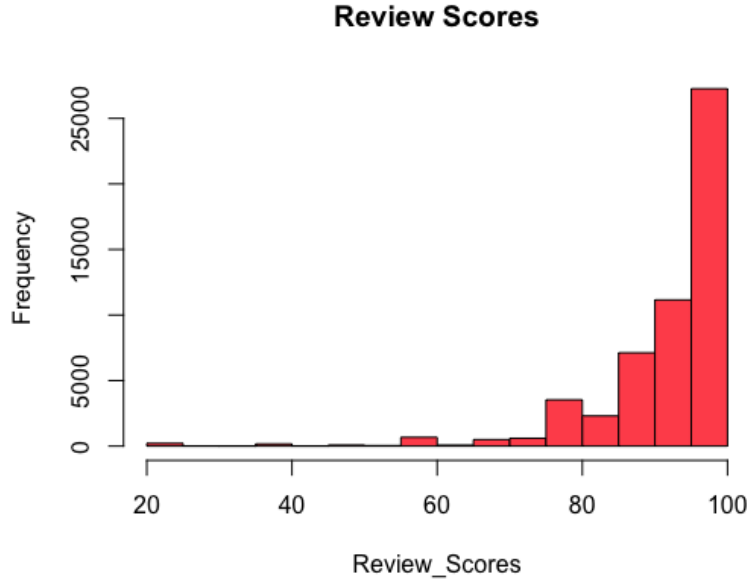
변수 정의

총 후기 개수

후기가 너무 적은 관측 값들의 평점은 너무 적은 수의 값들에 의해 정해지므로 제외한 후 모델을 적합함

좋음/안좋음

후기 점수가 98점 이상인 숙소를 '좋음'으로 코딩하여 반응변수 '후기 점수'를 새롭게 정의

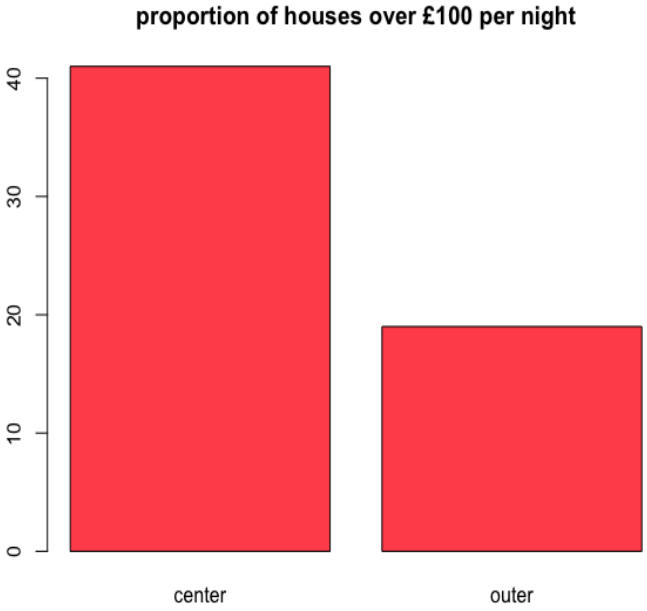


리뷰를 잘 주는 경향이 강함
'좋음'의 기준을 보수적으로 잡음
*98점은 중위수와 Q3의 중간이다.

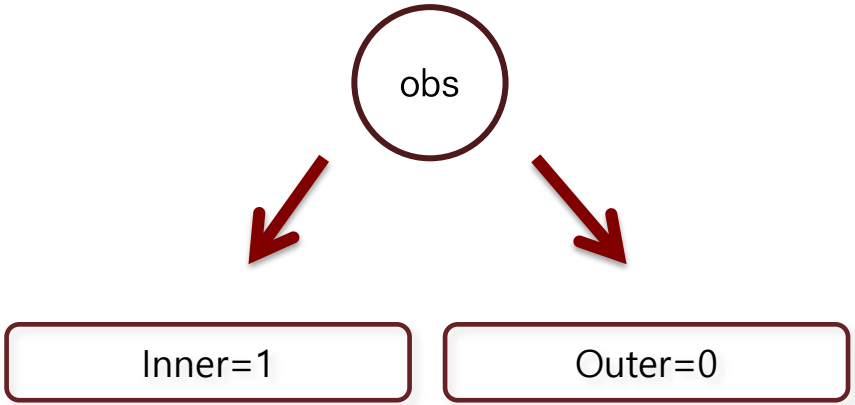
데이터 전처리

새로운 변수 추가

중심부/외곽



위치에 따라 숙소의 특성이 달라질 것이라 예상하여 중심부와 외곽을 구분하여 분석 진행
예) 1박당 100파운드가 넘는 숙소의 비중이 런던 중심부는 41%, 외곽은 19%이다.

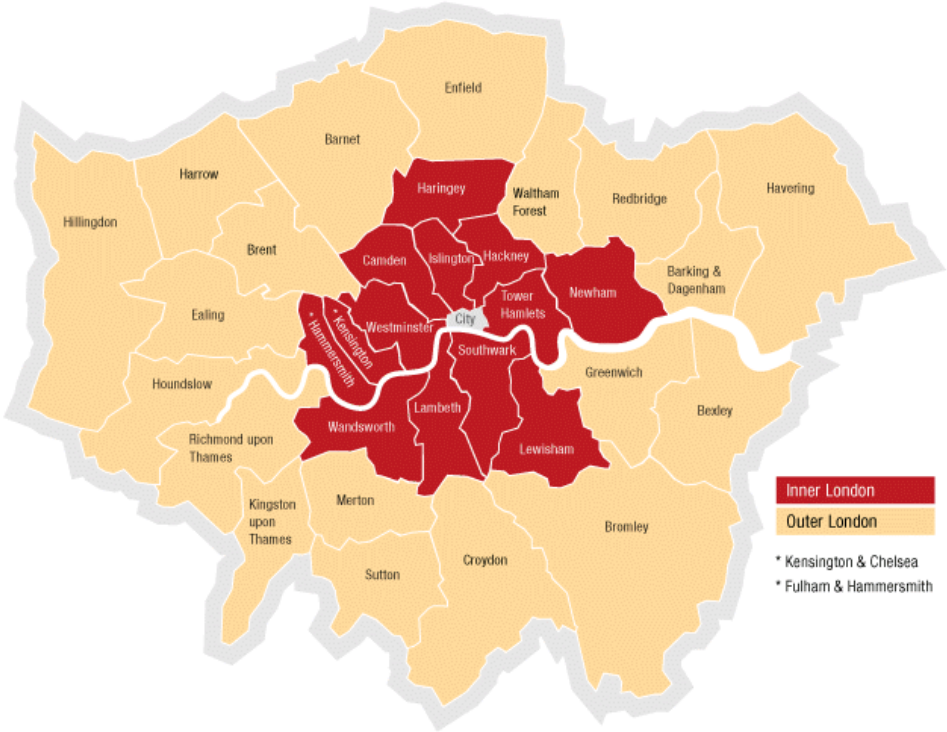


서론

데이터 전처리

새로운 변수 추가

중심부/외곽

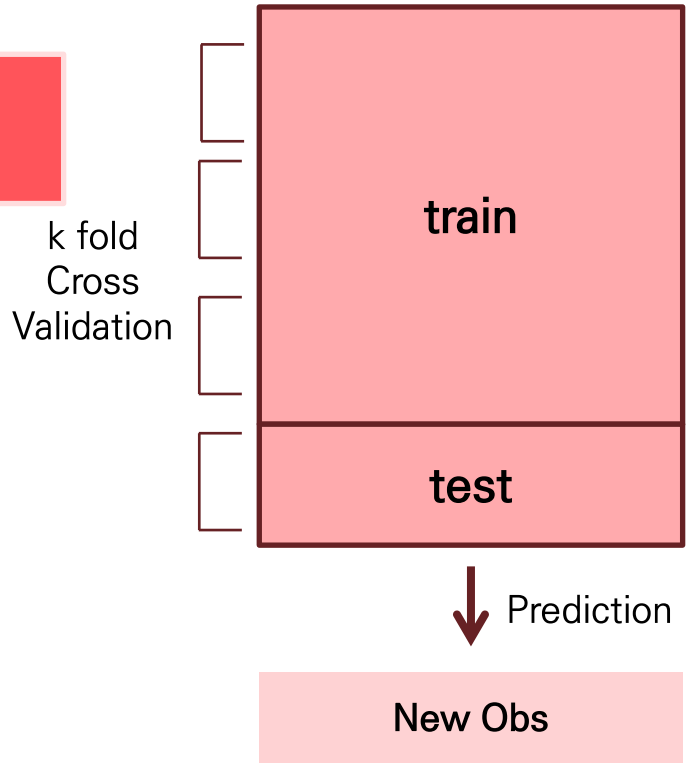



분석

분석 개요

분석

Model fitting (Validation set Approach) → Prediction



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

2. K fold Cross Validation 실시하여 cv test error 확인하여 최종 모델 선택

3. 후기 개수가 적은 (3이하인) 관측값들은 후기 점수의 신빙성이 떨어지므로 평점 데이터를 없앤 후 예측 데이터로 사용

KNN

KNN

변수선택

분석

수치형 변수만 추출

X's

수용가능인원
침실 개수
욕실 개수
침대 개수
최소 숙박일
이용 가능 일
월 후기 개수
가격

Binary 변수

Y

좋음/안좋음

KNN

좋음/안좋음

Inner London

분석

- 적절한 k값 선택

	Prediction Accuracy	Sensitivity	Specificity	CV.error.test
k=1	62.5%	0.323	0.738	37.1%
k=5	67.8%	0.209	0.855	32.2%
k=10	69.6%	0.154	0.900	30.5%
k=20	71%	0.076	0.952	28.8%

- 대체적으로 민감도 (실제로 좋은 숙소를 좋은 숙소로 예측)가 낮게 나오는 경향
  민감도를 높이면서 test error가 많이 커지지 않는 선에서 k선택
  k=5로 선택

KNN

좋음/안좋음

Outer London

분석

- 적절한 k값 선택

	Prediction Accuracy	Sensitivity	Specificity	CV.error.test
k=5	59.5%	0.326	0.747	40.1%
k=10	61.7%	0.285	0.802	39.2%
k=20	62.8%	0.243	0.842	37.7%
k=30	62.9%	0.176	0.881	37.3%

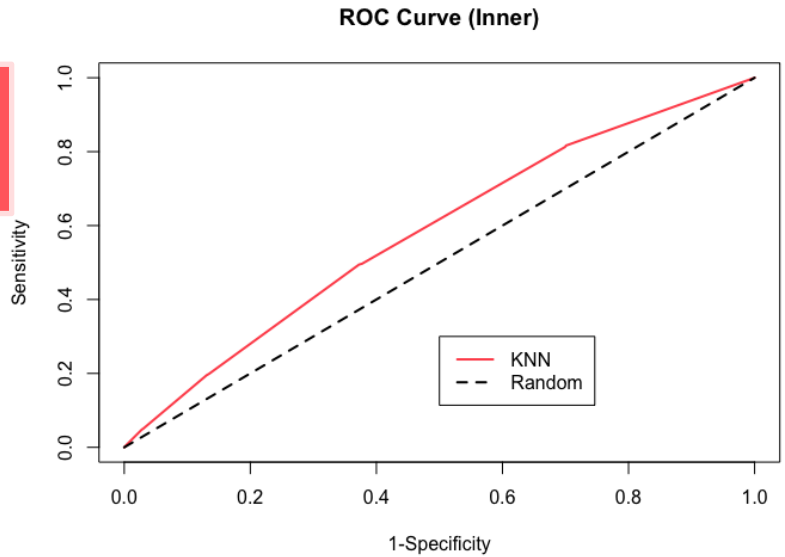
- 런던 시내의 경우와 같이 대체적으로 민감도가 낮게 나오고 시내보다 test error가 높다
  test error를 낮추면서 민감도가 너무 작아지지 않는 선에서 k 선택
  k=20으로 선택

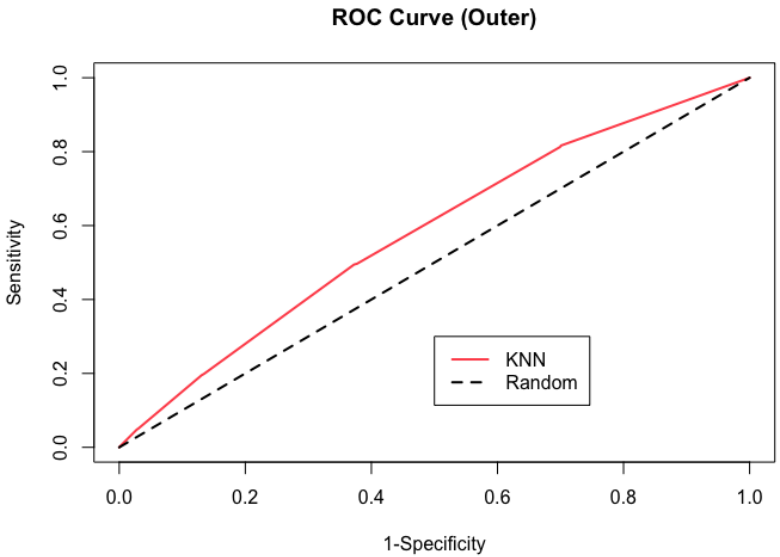
분석

KNN

좋음/안좋음

KNN 결론





- 런던 시내 AUC = 0.58, 런던 외곽 AUC = 0.60으로 예측력이 낮은 편
- 모델 특성상 범주형 변수의 정보를 잃음

LDA/QDA

LDA/QDA

변수선택

수치형 변수 + binary 변수

X's

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

Binary 변수

Y

좋음/안좋음

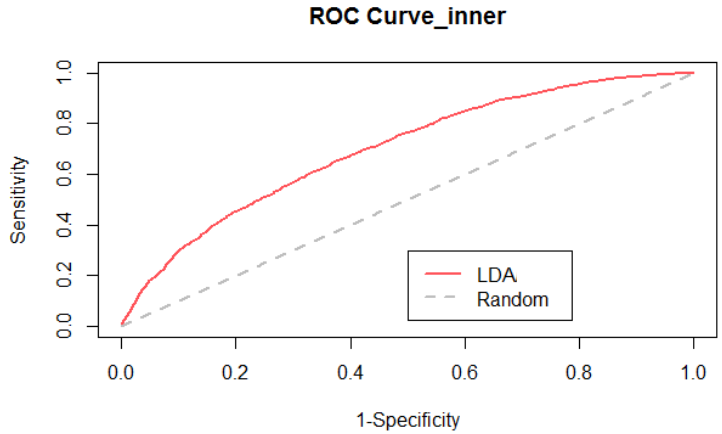
LDA/QDA

좋음/안좋음

Inner London

분석

Fold 개수		LDA	QDA
Validation set 7:3	Sensitivity	0.23	0.36
	Specificity	0.93	0.85
	Misclassification	0.26	0.28
	Prediction Accuracy	0.74	0.72
Fold=5	Test error	0.27	0.28
Fold=10	Test error	0.27	



- Prediction Accuracy : LDA 〉 QDA
  Test error : LDA 〈 QDA
- LDA의 민감도가 더 낮았지만 test error와 예측 정확도를 기준으로 LDA 모델을 채택하였다.
- AUC = 0.7

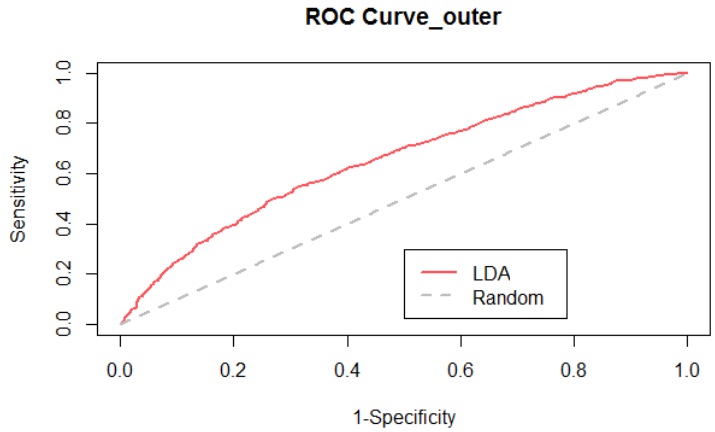
LDA/QDA

좋음/안좋음

Outer London

분석

Fold 개수		LDA	QDA
Validation set 7:3	Sensitivity	0.27	0.5
	Specificity	0.88	0.71
	Misclassification	0.34	0.37
	Prediction Accuracy	0.66	0.63
Fold=5	Test error	0.34	0.51
Fold=10	Test error	0.34	



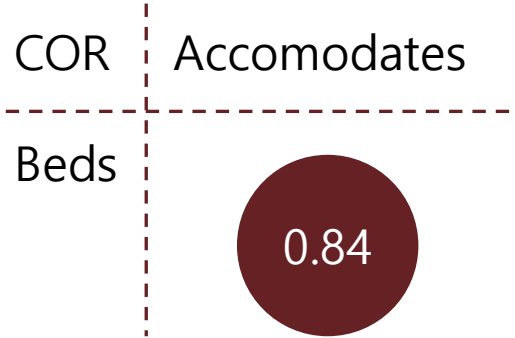
- Prediction Accuracy : LDA 〉 QDA
  Test error : LDA 〈〈 QDA
- 민감도가 LDA가 더 낮았지만 test error 와 예측 정확도를 기준으로 LDA 모델을 채택하였다.
- AUC = 0.66

회귀 분석

회귀 분석

변수선택

분석



'침대개수' 변수 제외
(다중공선성 문제)

- **설명변수**
  제안된 숙소 유형, 슈퍼호스트여부, 프로필사진 유무, 신상 확인 여부, 숙소 종류, 수용가능인원, 욕실 개수, 침실 개수, 침대 종류, 가격, 최소 숙박 일, 이용가능일(기준:1년), 즉시 예약 가능여부, 예약 취소 정책, 월 후기 개수, 중심부/외곽, 좋음/안좋음

  Inner : – 침대유형, 최소 숙박일, 욕실 수, 프로필 사진 유무
  Outer : – 침대유형, 월평균 후기 개수, 최소숙박일, 숙소 유형

- **반응변수**
  좋음/안좋음 — Binary 로지스틱 회귀의 Y변수
  가격 — 수치형 선형회귀 분석의 Y변수

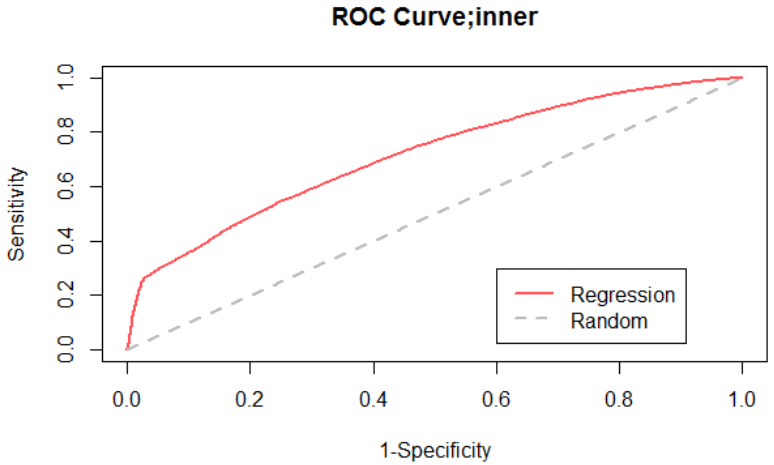
로지스틱 회귀

Y=좋음/안좋음

Inner London

분석

		Cutoff=0.5	Cutoff=0.3
Validation set 7:3	Sensitivity	0.20	0.56
	Specificity	0.94	0.72
	Misclassification	0.26	0.33
	Prediction Accuracy	0.74	0.67
Fold=5	Test error	0.26	0.33
Fold=10	Test error	0.26	



- 민감도가 대체로 낮아서 예측정확도와 test error 기준
- Cutoff= 0.5인 첫 번째 모델을 선택
- AUC = 0.71

Fitted Model

Business는 BAD
침실 수 많을수록 GOOD
슈퍼호스트 일수록 GOOD

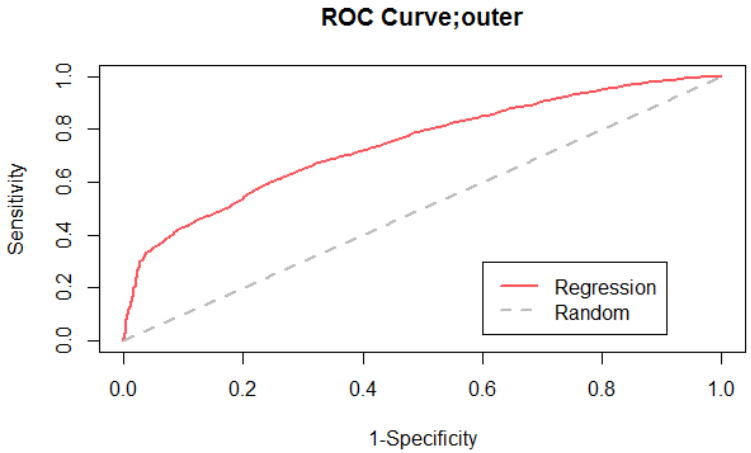
로지스틱 회귀

Y=좋음/안좋음

Outer London

분석

		Cutoff=0.5	Cutoff=0.3
Validation set 7:3	Sensitivity	0.27	0.45
	Specificity	0.88	0.74
	Misclassification	0.34	0.44
	Prediction Accuracy	0.66	0.56
Fold=5	Test error	0.33	0.43
Fold=10	Test error	0.33	



- 민감도가 대체로 낮아서 예측정확도와 test error 기준
- Cutoff = 0.5인 첫 번째 모델을 선택
- AUC = 0.66

Fitted Model

프로필 사진 게시할수록 GOOD
Private 는 GOOD P〉Share〉Entire
즉시 예약 가능하면 BAD

다중 선형 회귀 1

Y=가격

합리적인 가격은?

분석

- 전체 데이터 ; R^2=0.5624 – 침대유형, 프로필 사진 유무

$$Price = 8.1 - 6.7\,Exp_{family} - 7.3Exp_{none} + 4.4\,Exp_{roman*} - 13.7\,Exp_{social} + 5.6superhost$$
$$-47.2\,room_{private} - 1.7veri - 91.1\,room_{shared} + 24.0\,bathrooms$$
$$+17.2bedrooms + 9.2accom - 0.06minnight - 0.06avail + 8.4score_{good}$$
$$-0.2cancel_{moderate*} + 56.9\,cancel_{strict14} + 58.5\,cancel_{srict30} - 2.5cancel_{strict*}$$
$$+60.6cancel_{strict60} - 1.7review_{month} + 26.1\,region\_inner$$

1.화장실개수, 침실개수, 수용가능인원의 계수가 양수임을 보아 큰 시설을 빌릴수록 가격이 증가한다
2.후기 점수가 높을 수록, 도시 중심에 위치하면 가격이 비싸다.
–〉 평점이 좋을수록 가격이 증가하는 것은 비싼 곳 인만큼 더 좋은 장소이기 때문일 가능성이 크다.

다중 선형 회귀 2

Y=가격

합리적인 가격은?

분석

- 전체 데이터 ; R^2=0.5624

$$Price = 8.1 - 6.7\,Exp_{family} - 7.3Exp_{none} + 4.4\,Exp_{roman*} - 13.7\,Exp_{social} + 5.6superhost$$
$$-47.2\,room_{private} - 1.7veri - 91.1\,room_{shared} + 24.0\,bathrooms$$
$$+17.2bedrooms + 9.2accom - 0.06minnight - 0.06avail + 8.4score_{good}$$
$$-0.2cancel_{moderate*} + 56.9\,cancel_{strict14} + 58.5\,cancel_{srict30} - 2.5cancel_{strict*}$$
$$+60.6cancel_{strict60} - 1.7review_{month} + 26.1\,region\_inner$$

3.가격을 책정할 때 최소 숙박일수나, 1년 중 숙박 가능날짜의 영향은 크지 않기에 에어비엔비에 자신의 장소를 올릴 때 이에 따른 가격 변동은 크게 고려하지 않아도 된다.
4.한 달 이전 취소금지나 2주 이전 취소금지나 가격에 미치는 영향은 크게 다르지 않기에 개인사정에 맞게 공지하면 된다.
5. 월 평균 리뷰개수의 계수가 음수인 것은 리뷰가 증가할 수록 가격이 떨어지는 인과관계를 나타낸 것이 아니라, 가격이 저렴한 숙소일수록 많은 사용자들이 이용하는 것일 수도 있음을 고려해야 한다.


결론

분석 방법 간 비교

16,344 obs 예측

Prediction – Inner

	Bad	Good	총합
KNN	9,737	2,269	12,006
LDA	11,529	477	12,006
Logit	11,545	461	12,006

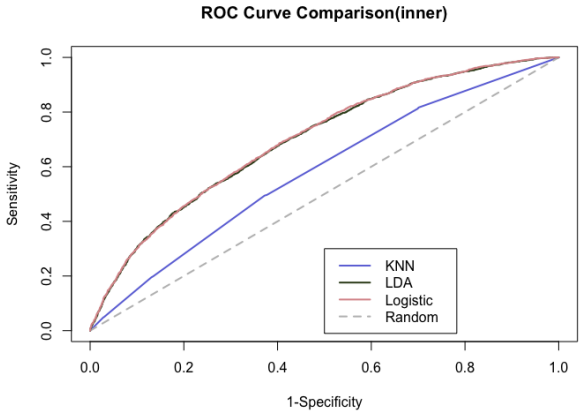
Prediction – Outer

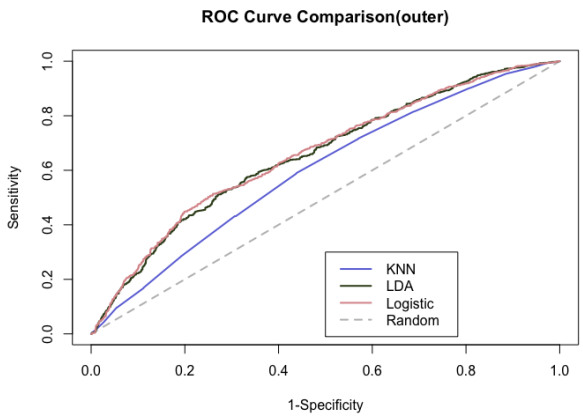
	Bad	Good	총합
KNN	3,545	793	4,338
LDA	4,074	264	4,338
Logit	4,043	295	4,338

- LDA, 로지스틱 회귀모형은 '좋음', '안좋음' 그룹에 각각 비슷한 개수로 분류

분석 방법 간 비교

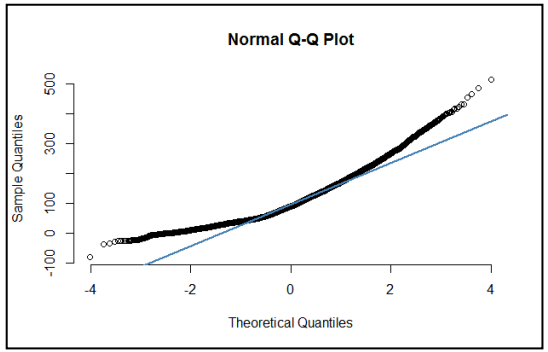
결론





- 수치형 변수만 포함시킨 KNN의 분류에 범주형 변수의 정보가 손실되어 부정확한 분류가 되었을 가능성이 있음
- 런던 시내와 외곽 모두 K fold CV test error가 낮고 예측력이 높은 LDA 혹은 로지스틱 모형을 사용하는 것이 나을 것으로 판단

한계점



정규분포 가정을 중시하는 회귀분석에서 데이터 자체가 정규성을 충족하지 못해서 가격에 대한 회귀분석 결과를 신뢰할 수 있는지는 확신할 수 없다.

- 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 있었다.
  : Rating이 수익에 영향을 미치겠지만, 수익 관련 정보가 없었기에 수익성으로 곧바로 이어지는 결과는 아니었다.

감사합니다