Airbnb Price Prediction Model 개발 및 유의 변수 탐색

2015-16828 김민규
2015-15162 김태현
2015-15711 민준기

목차

1. 데이터 설명

우리가 사용하는 데이터는 Kaggle의 'Airbnb Price Prediction' Dataset(https://www.kaggle.com/stevezhenghp/airbnb-price-prediction)으로, 각 숙박 업소의 가격 및 숙박 업소 관련 정보들이 존재하는 데이터이다. 총 74,111건의 데이터가 존재하고 각 데이터는 29개의 변수들로 이루어져 있으며, 변수들의 목록은 아래와 같이 정리할 수 있다.

Quantitative Variables	Continuous	log_price
	Discrete	accommodates, bathrooms, number_of_reviews, bed_type
		room_type, cleaning_fee, review_scores_rating, beds
Qualitative Variables	Nominal	host_has_profile_pic, host_identity_verified, instant_bookable, property_type, amenities, city
	Ordinal	cancellation_policy
Not Used	thumbnail_url, zipcode, latitude, longitude, host_response_rate, host_since	
	id, description, first_review, last_review, name, neighbourhood	

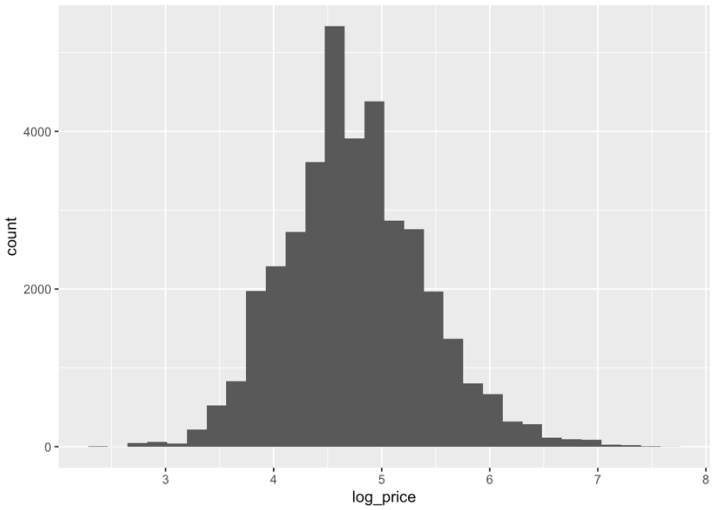
위의 변수들 중에서 'thumbnail_url', 'zipcode', 'latitude', 'longitude', 'host_response_rate', 'host_since', 'id', 'description', 'first_review', last_review', 'name', 'neighbourhood'와 같은 변수들은 이미 다른 변수들에 포함된 정보거나(latitude, longitude는 city와 관련), 'log_price'에 무관한 변수(id, name 등)라고 판단하여 분석에서 제외하였다.

각 변수들에 대한 기초 탐색적 분석 결과 및 자세한 설명은 아래와 같다.

1-1. Quantitative Variables

1-1-1. Continuous Variables

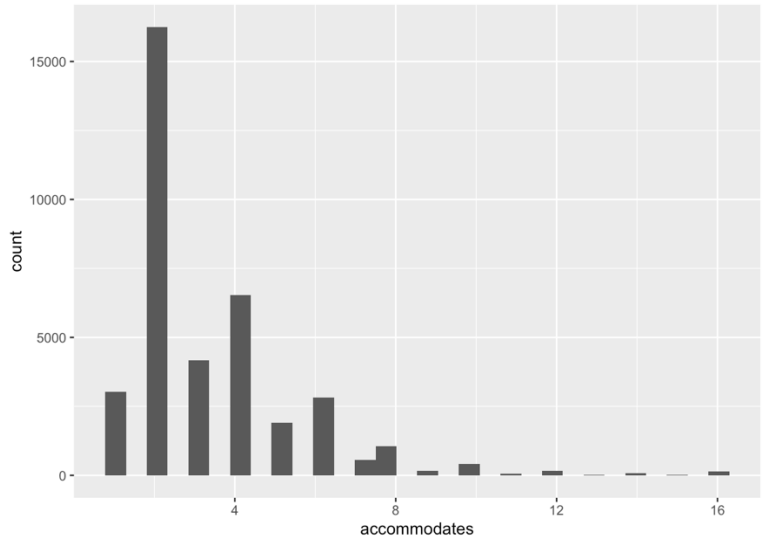
1-1-1-1. log_price



분석하고자 하는 Response Variable인 'log_price'이다. 이는 전반적으로는 정규분포와 형태가 유사한 것으로 보이지만, Shapiro Test 결과로는 정규성을 띠지 않는 것으로 확인되었다. 그럼에도 불구하고 형태가 유사하다는 점에서 우선은 정규분포에 기반한 모형들을 시도해보았다.

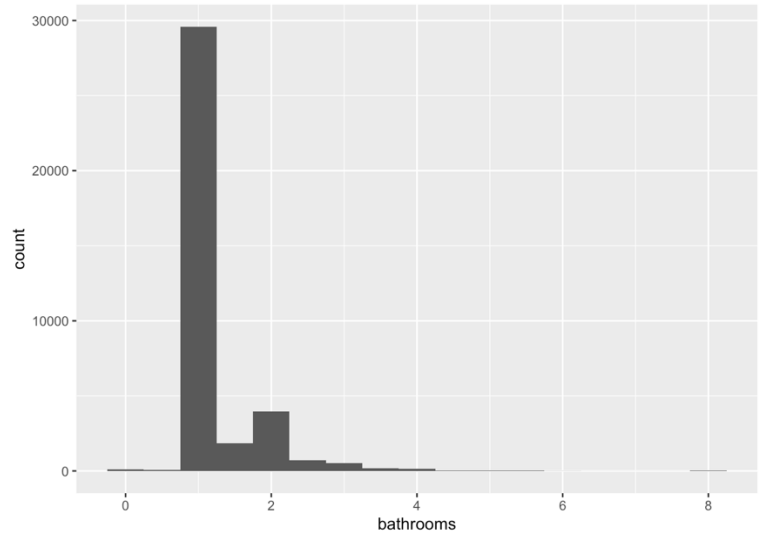
1-1-2. Discrete Variables

1-1-2-1. accommodates



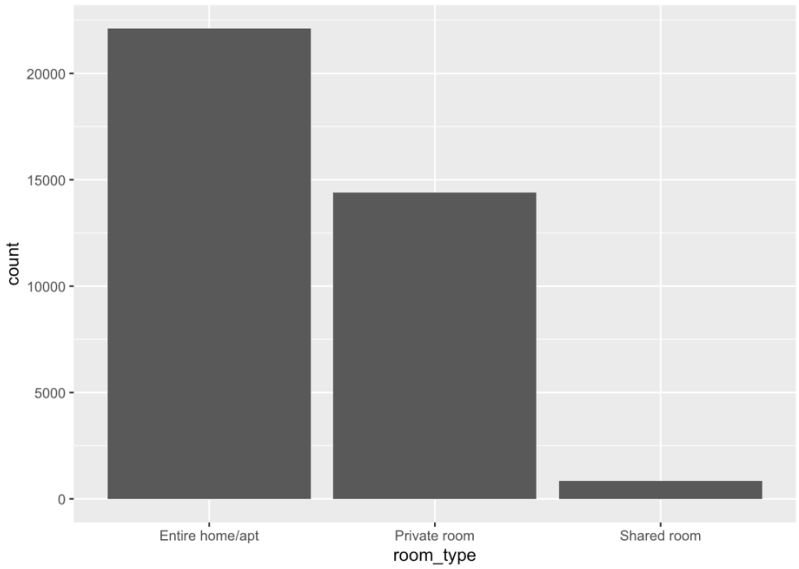
숙박 인원을 나타내는 'accommodates'이다. Discrete Quantitative Variables 중에서 하나로 전체 Data에서 Count한 것을 히스토그램으로 나타낸 것이다. 숙박 인원이 8을 초과하는 숙소는 거의 없는 것을 확인할 수 있다.

1-1-2-2. bathrooms



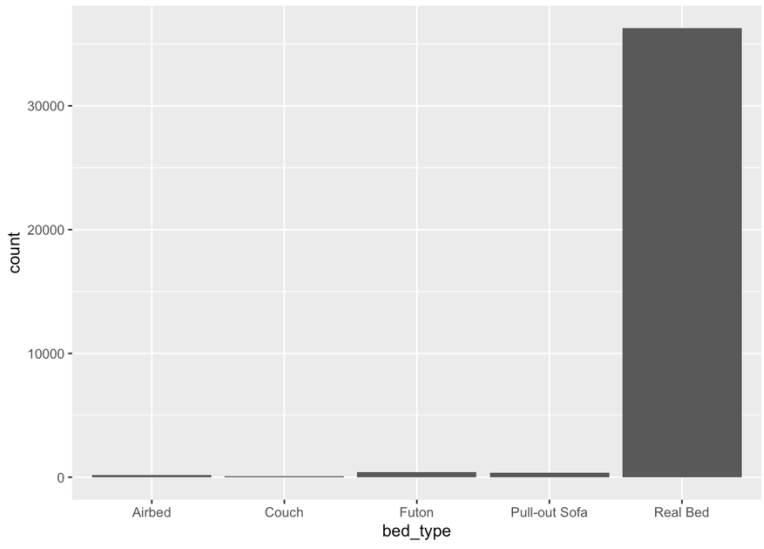
화장실의 개수를 나타내는 bathrooms 변수로, Discrete Quantitative Variable이다. 마찬가지로 대부분의 데이터가 bathrooms이 1~2개였다.

1-1-2-3. room_type



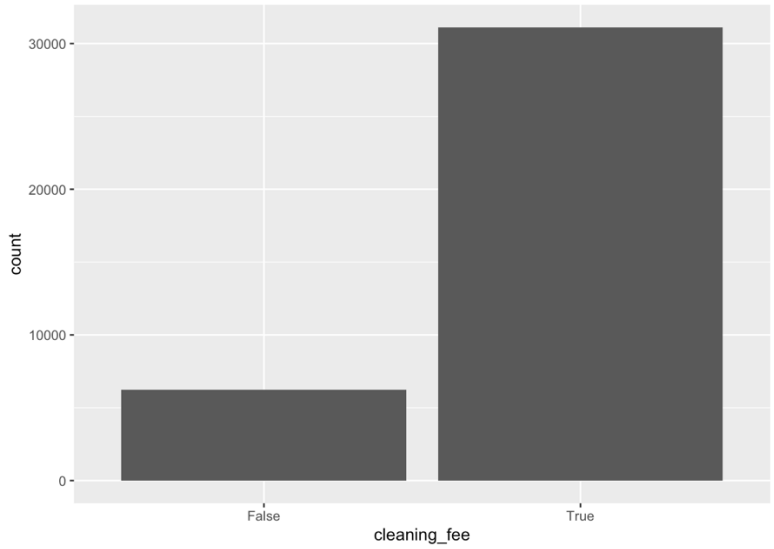
숙소의 주거 형태를 나타낸 변수이다. 총 3가지, entire home/apt와 private room, shared room으로 나뉘는 Nominal Qualitative Variable이다. shared room의 개수가 나머지 두 범주보다 굉장히 적은 모습을 보여준다.

1-1-2-4. bed_type



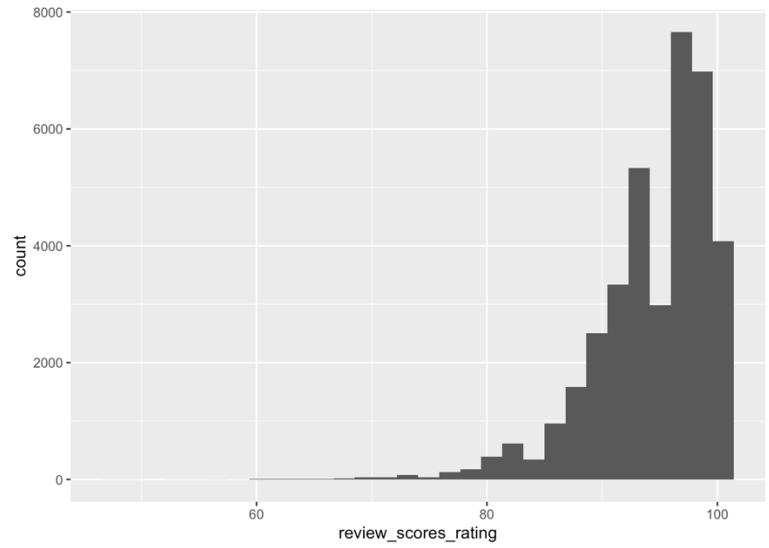
침대의 종류를 나타내는 변수이다. 극히 일부를 제외하고는 전부 Real Bed 타입이기에, 이 변수는 향후의 분석에서 제외하였다.

1-1-2-5. cleaning_fee



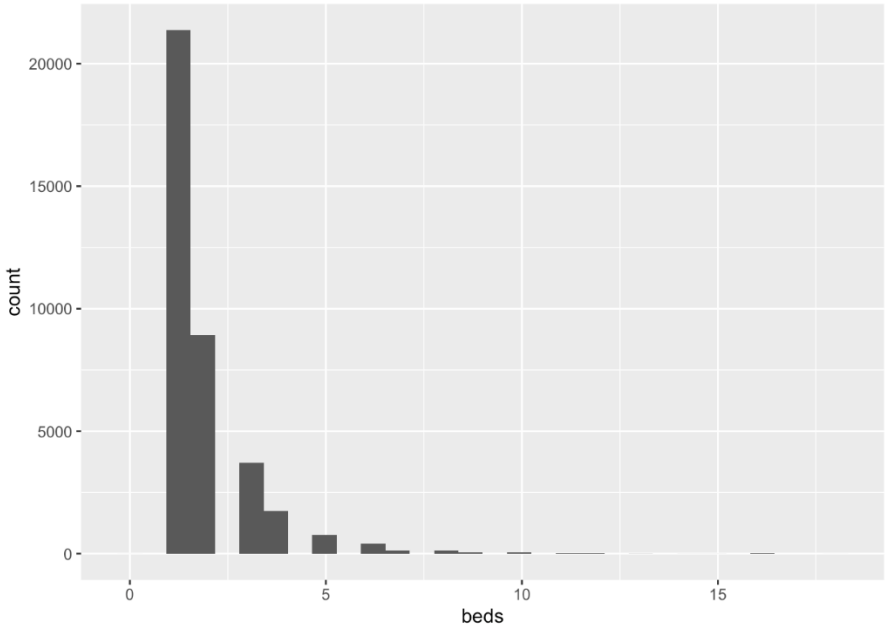
청소비가 추가적으로 있는지를 나타내는 cleaning_fee 변수다. true/false형 변수로 이 역시 Nominal Qualitative Variable이다.

1-1-2-6. review_scores_rating



숙소의 평점을 나타내는 Discrete Quantitative Variable이다. 95~100점 사이의 평점에 거의 대부분의 데이터가 존재하는 불균형 현상이 존재한다. 다만, 이 변수와 관련깊은 변수가 number_of_reviews, 즉 리뷰의 개수이다. 리뷰의 개수가 1~2개 정도로 굉장히 적은 숙소의 평점은 신뢰할 수 있는 수치가 아니라고 생각하여, 전체 데이터에서 리뷰 개수가 5개보다 많은 데이터만 분석을 시도하였다.

1-1-2-7. beds

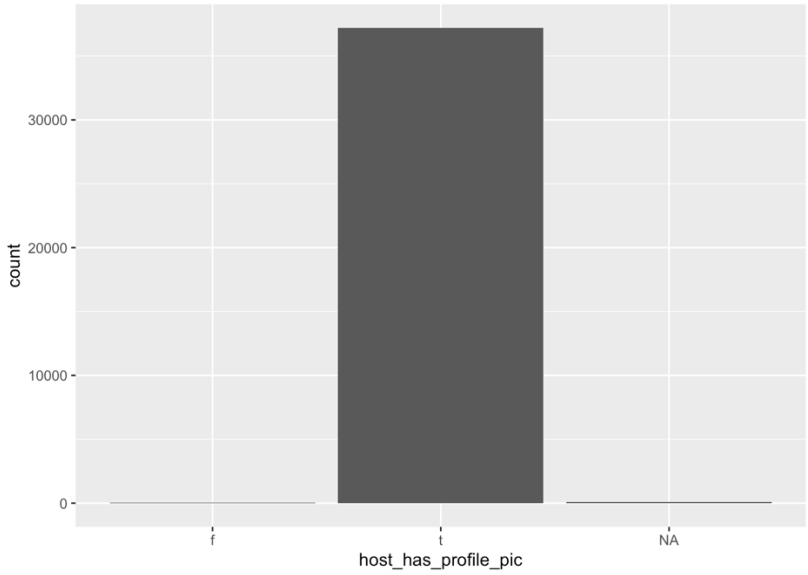


침대의 개수를 나타내는 변수로, Discrete Quantitative Variable이다. 1~2개의 침대를 가진 숙소가 거의 대부분이고, 그 이상의 숙소는 굉장히 적은 것을 볼 수 있다.

1-2. Qualitative Variables

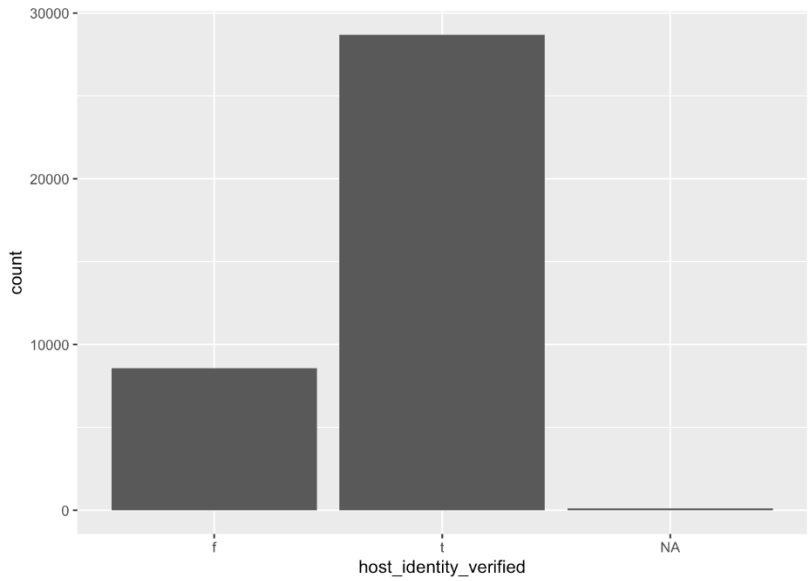
1-2-1. Nominal Variables

1-2-1-1. host_has_profile_pic



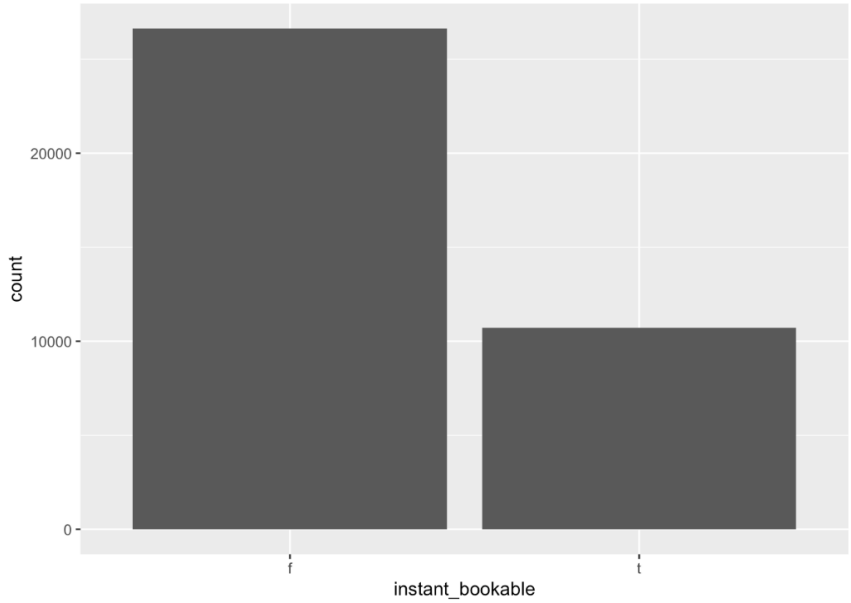
호스트가 프로필 사진이 있는지 여부에 대한 변수이다. 거의 모든 호스트들이 프로필 사진을 가지고 있기에, 의미가 없는 변수라 생각되어 분석에 사용하지 않았다.

1-2-1-2. host_identity_verified



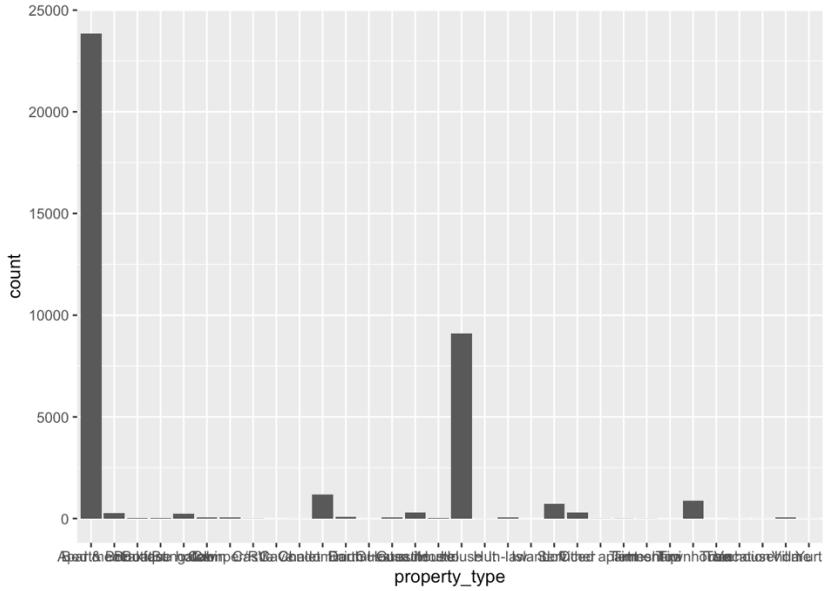
호스트의 신원이 밝혀져 있는지에 대한 변수이다. NA 값이 일부 있는데, 이 행들은 분석에서 제외하였다.

1-2-1-3. instant_bookable



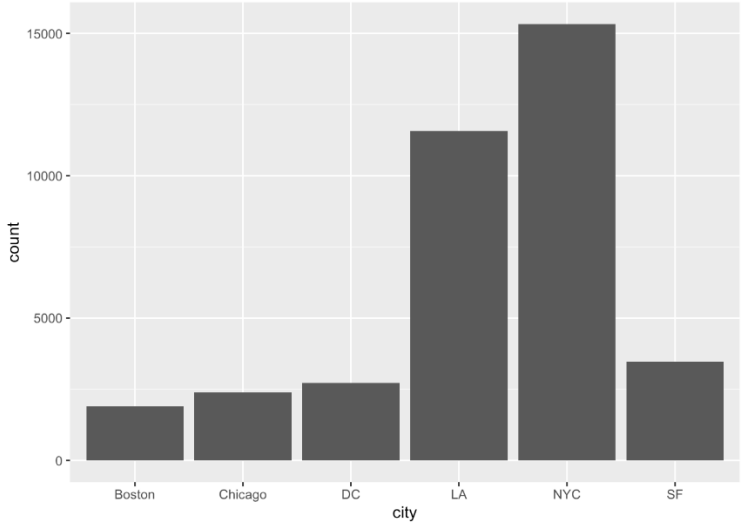
Airbnb에서 '즉시 예약' 기능을 제공하는지 여부에 대한 Logical Variable이다. '즉시 예약'이란 호스트의 조건을 만족하는 게스트는 호스트에게 따로 예약 승인 요청을 보낼 필요 없이 바로 예약을 할 수 있는 것이다.

1-2-1-4. property_type



제공되는 숙소 건물의 형태인 property_type variable이다. Nominal Qualitative Variable 중 하나다. 다양한 형태의 property_type이 있는 것을 확인 할 수 있고 가장 많은 것은 많은 것부터 순서대로 나열해보면 Apartment, House, Condominium 순이다.

1-2-1-5 city



숙소가 위치한 도시의 이름이다. 총 6개의 도시에서의 숙소 데이터이다.

1-2-1-6. amenities

amenities
{""Wireless Internet"",""Air conditioning"",Kitchen,Heating,""Family/kid friendly"",Essentia...
{""Wireless Internet"",""Air conditioning"",Kitchen,Heating,""Family/kid friendly"",Washer,...

기존 raw data의 amenities는 각 항목이 다음과 같이 긴 문자열로 처리되어 있었다. 분석하기에 적절한 형태가 아니기 때문에, 이를 파싱하여 각 amenity를 분리하였고, 총 137개의 amenity를 추려내었다.

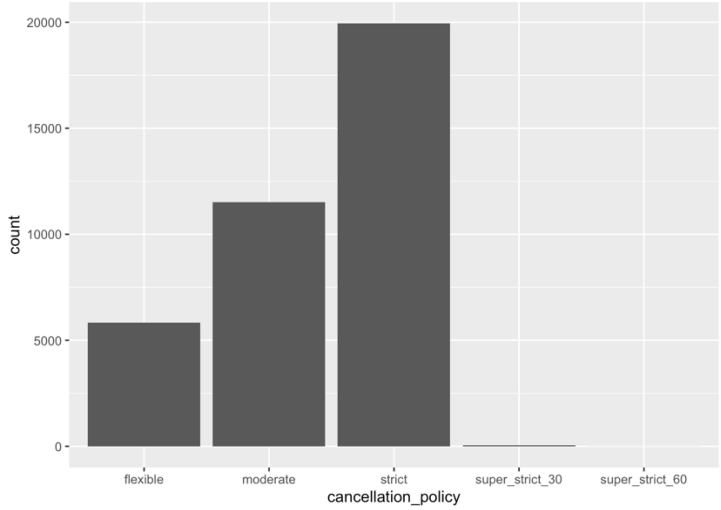
그 다음, 아래 그림과 같이 각 amenity별로 이 amenity를 가지고 있는 숙소는 TRUE, 가지고 있지 않은 숙소는 FALSE를 배정하였다.

Wireless.Internet.	Air.conditioning.
TRUE	TRUE
TRUE	TRUE
TRUE	TRUE
TRUE	FALSE
TRUE	TRUE
TRUE	FALSE
TRUE	TRUE

amenity의 종류가 많기 때문에 이를 포함하는 여부에 따라 데이터의 크기가 크게 차이났다. 그래서 amenity를 포함한 데이터와 포함하지 않은 데이터 두 가지에 대해서 각각 분석을 진행해 보기로 하였다.

1-2-2. Ordinal Variables

1-2-2-1. cancellation_policy



취소 정책을 나타내는 cancellation_policy다. 이는 각 Qualitative Variable이긴 하나 각 항목들이 수준에 따라 순서를 정할 수 있는 Ordinal Qualitative Variable이다. super_strict_30과 super_strict_60은 매우 적은 수준이므로, strict에 병합하였다.

2. 분석 목표

이 프로젝트는 숙박 공유 플랫폼 서비스인 Airbnb 에서의 숙박 업소 가격을 예측하는 모델을 만들고, 어떤 변수들이 숙소 가격에 영향을 미치는지 알아보기 위한 프로젝트이다.

숙박 업소에는 굉장히 다양한 옵션들이 존재한다. 이를테면 숙박 업소의 형태(주택, 아파트, 펜션 등)와 같이 굉장히 중요한 것부터, 전자레인지나 커피포트와 같이 사소한 사항까지 존재한다. 이렇게 다양한 옵션들은 분명 숙박 가격에 영향을 미칠 것이다. 숙소의 형태, 위치와 같이 쉽게 예상할 수 있는 것들부터, 예상하지 못했지만 숙소의 가격에 영향을 미치는 숨겨진 변수들을 찾아보고, 또한 그러한 변수들이 숙소의 가격에 얼마만큼의 영향을 미치는지 측정해 본다.

또한, 숙박 가격을 예측하는 일반적인 Regression 모델 뿐만 아니라, 숙박 가격을 중앙값 이상과 이하로 나누어, 가격이 높은 숙소와 낮은 숙소 두 가지로 Binary Classification 하는 모델도 만들어 보고, Regression 과 Classification 에서 공통적으로 얻어낼 수 있는 결과를 도출한다.

뿐만 아니라, 이를 위해 단순한 선형회귀분석부터 앙상블 방법까지 다양한 모델들을 적용해 볼 수 있는데, 이 모델들의 장단점을 알아보고 각 모델에서의 결과와 선택된 주요 변수들을 해석할 것이다.

3. 분석 계획

앞에서 전처리한 데이터를 이용하여, 다음과 같은 분석 계획을 마련하였다. 우선 Response Variable인 'log_price'를 원데이터에서 주어진 대로 Continuous Scale로 분석하는 계획과(3-1), 중앙값보다 크면 1, 중앙값보다 작으면 0으로 인코딩하여 Binary Scale로 분석하는 계획을 마련하였다(3-2). Continuous Scale을 이용할 경우 'log_price'의 분포가 정규분포와 형태가 유사하다는 점에 착안하여 정규분포에 기반한 모형들을 시도할 수 있다. 또한 Binary Scale을 이용할 경우, 'log_price'의 크고 작음을 단순화하여 분류 모델을 이용할 수 있다는 장점이 있다.

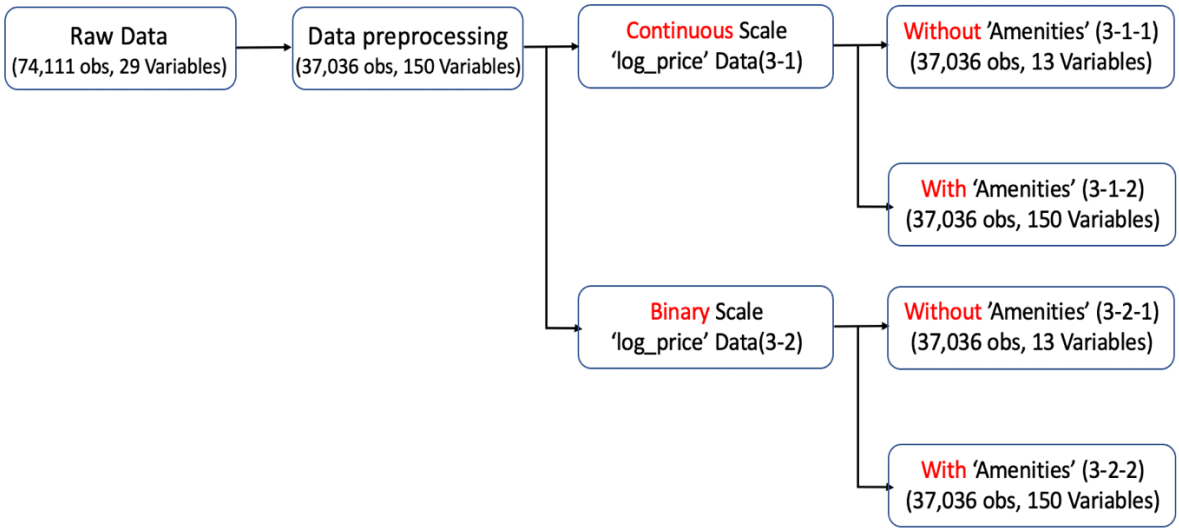
Response Variable을 Continuous Scale로 두는 경우에는 Multiple Linear Regression, LASSO, 그리고 XGBoost와 같은 모형을 적합하였다. 앞에서 확인한 바와 같이 'log_price'의 분포가 정규분포와 유사하다는 점에서 Multiple Linear Regression을 적합할 수 있다. 또한 일반적인 Multiple Linear Regression에 L1-norm에 대한 Penalty를 주는 LASSO 모형을 이용하여 유의한 변수를 선택함과 동시에 회귀모형을 적합하였다. 이 경우 lambda는 10 fold Cross Validation을 이용하여 구하였다. 마지막으로 이용한 XGBoost는, 기본적으로 Gradient Boosting Method(이하 GBM)를 기반으로 한 Ensemble Method이다. 이는 일반적인 GBM과는 다르게 Regularization 항이 있고 그 밖에도 Greedy Algorithm등을 사용한 자동 가지치기가 가능해, 과적합 이슈에서 보다 자유롭다. 또한 다양한 Hyper Parameter를 조정할 수 있어 문제 상황에 맞게 최적화가 가능하다. 무엇보다 분산 처리를 사용하고, 코드 내부는 실행 속도가 빠른 C로 구현되어 있기에 결과를 출력하기까지의 시간이 GBM보다 빠르다는 점에서 XGBoost를 이용하였다. 이를 이용하기 위하여 R에서 xgboost 라이브러리를 사용하였으며, Tree의 Max Depth는 결과가 가장 우수하였던 6으로 고정하였고, Train 횟수는 너무 많으면 과적합이 일어나는 것으로 보이기에, 과적합이 일어나기 전인 50으로 설정하

였다. XGBoost 같은 경우 입력값으로 Numerical Value만 허용된다. 따라서 Categorical Variables는 전부 Dummy Variables로 One-hot Encoding을 하는 것이 필요하다. 다만, 이 경우 Cancellation_policy같은 Ordinal Variables는 Order가 사라진다는 문제점이 있지만, Ordinal Variables를 임의의 숫자에 Mapping하는 것은 숫자를 지정함에 있어서 임의 판단이 많이 들어가므로, Dummy Variable로 바꾸는 것이 최선의 방법이라고 생각하였다.

Response Variable을 Binary Scale로 두는 경우에는, Logistic Regression, (Logistic) LASSO, 그리고 XGBoost와 같은 모형을 적합하였다. (Logistic) LASSO의 경우 Logistic Regression에 L1-norm에 대한 Penalty를 주는 모형이다.

한편, 'log_price'를 Continuous Scale로 이용하는 경우, Binary Scale로 이용하는 경우 각각을 다시 'amenities' 변수를 제외하고 모형을 적합하는 경우와, 포함하고 모형을 적합하는 경우 두 가지로 나누어서 분석하였다. 'amenities'에는 앞에서 확인한 것과 같이 각 숙소에 어떠한 물품이 존재하는지 안 하는지의 여부에 대한 정보가 존재하는데, 이들 변수는 137개로 다른 설명변수들의 개수 총합(12개)에 비하여 매우 많은 편이다. 따라서 'amenities'를 포함하지 않는 경우를 Setting 1이라 하고(Without 'amenities' : 3-1-1, 3-2-1), 'amenities'를 포함하는 경우를 Setting 2라 할 때(With 'Amenities' : 3-1-2, 3-2-2) 두 가지 Setting에서의 Evaluation 결과가 큰 차이를 보이지 않을 경우 'amenities'를 포함하지 않고도 효율적인 모형을 만들 수 있을 것으로 기대할 수 있다.

위의 과정을 도식화 한 결과는 아래와 같다.



각 모형을 적합한 이후, 'log_price'와 관련이 있는 유의한 변수들을 다음과 같은 기준으로 선정하였다. Multiple Linear Regression과 Logistic Linear Regression에서는 회귀 계수에 대한 p-value가 0.05 이하인 변수들을 유의한 것으로 판단하였다. LASSO와 (Logistic) LASSO에서는 회귀 계수가 0이 아닌, 즉 LASSO 적합을 통하여 선택된 변수들을 유의한 것으로 판단하였다. 마지막으로 XGBoost에서는 각 변수의 Importance를 계산하여 Importance가 높은 변수들을 'log_price'와 관련이 있는 변수들로 선정하였다. XGBoost에서 Importance의 경우 Gain이라는 개념과 연관 지을

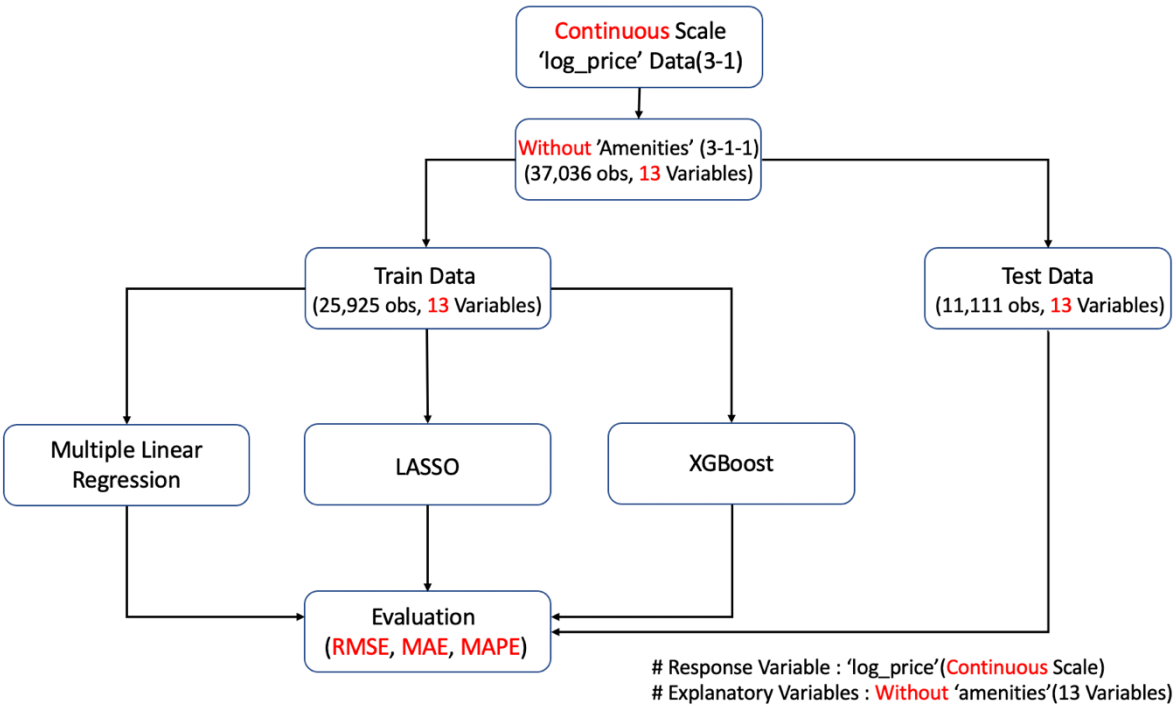
수 있다. 이는 Tree에 새로운 Feature를 이용하여 Split을 추가하였을 때 결과가 얼마나 더 좋아지는지에 대한 척도이다. 모든 변수의 Gain의 합은 1이며, 그렇기 때문에 변수들 간의 상대적인 Importance를 구할 수 있다.

Train Data에 각 모형을 적합한 후, 적합된 모형을 바탕으로 Test Data를 이용하여 예측한 후 다음과 같은 Metric들을 이용하여 비교하였다. 우선 Continuous Scale 'log_price'의 경우, Metric으로 RMSE(Root Mean Squared Error), MAE(Mean Absolute Error), MAPE(Mean Absolute Percentage Error)를 이용하였으며, Binary Scale 'log_price'의 경우 Confusion Matrix로부터 Accuracy, Sensitivity, Specificity와 같은 Metric을 이용하여 Evaluation 결과를 비교하였다.

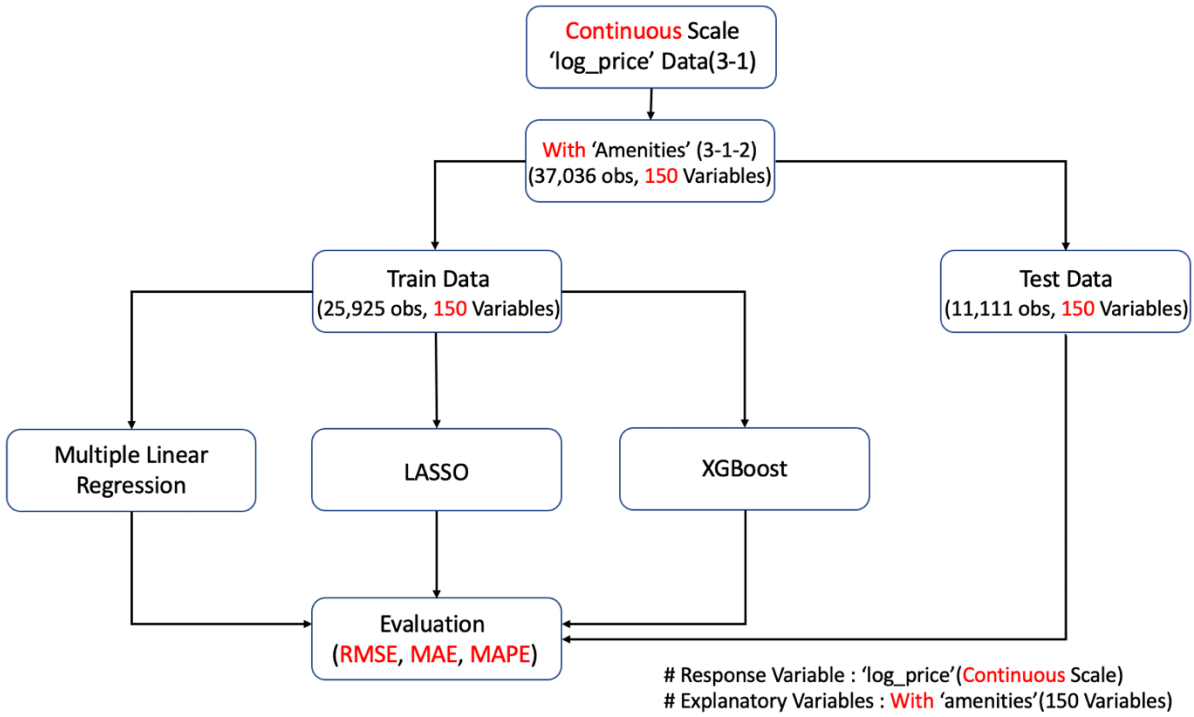
위의 계획들을 종합하여 도식화한 결과는 아래의 3-1과 3-2와 같다.

3-1. Continuous Scale 'log_price'에 관한 분석 계획

3-1-1. Setting 1: Without 'amenities' Variables(13 Variables)

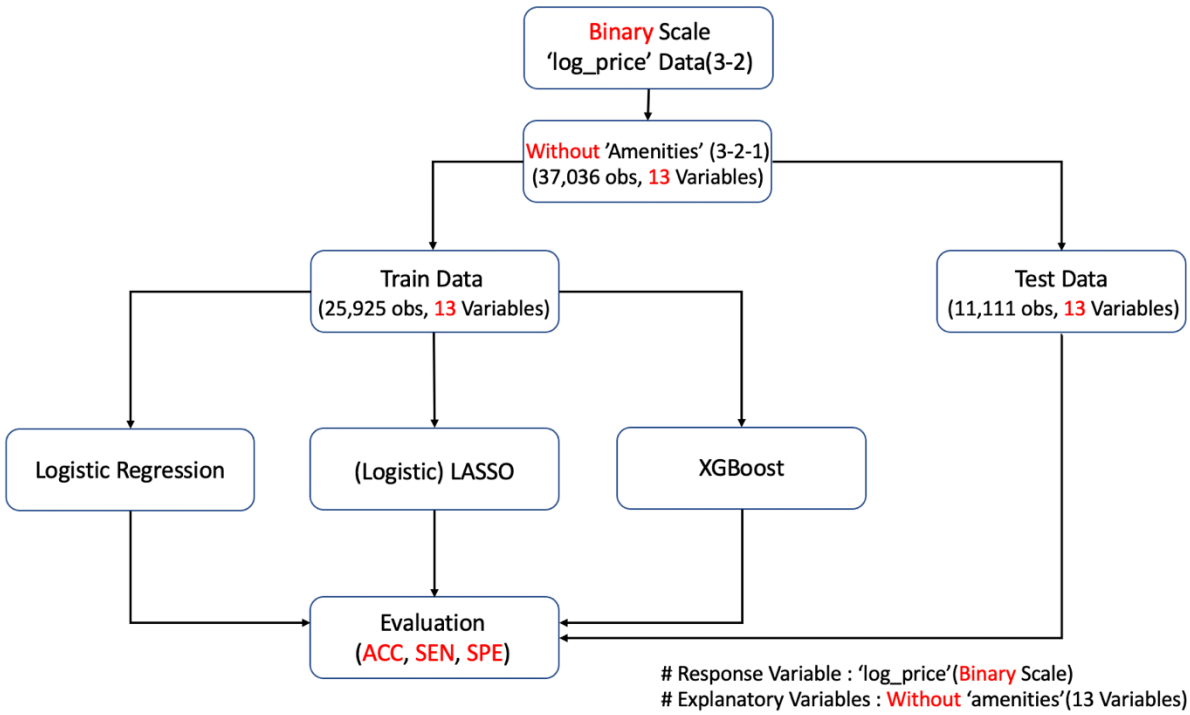


3-1-2. Setting 2: With 'amenities' Variables(150 Variables)

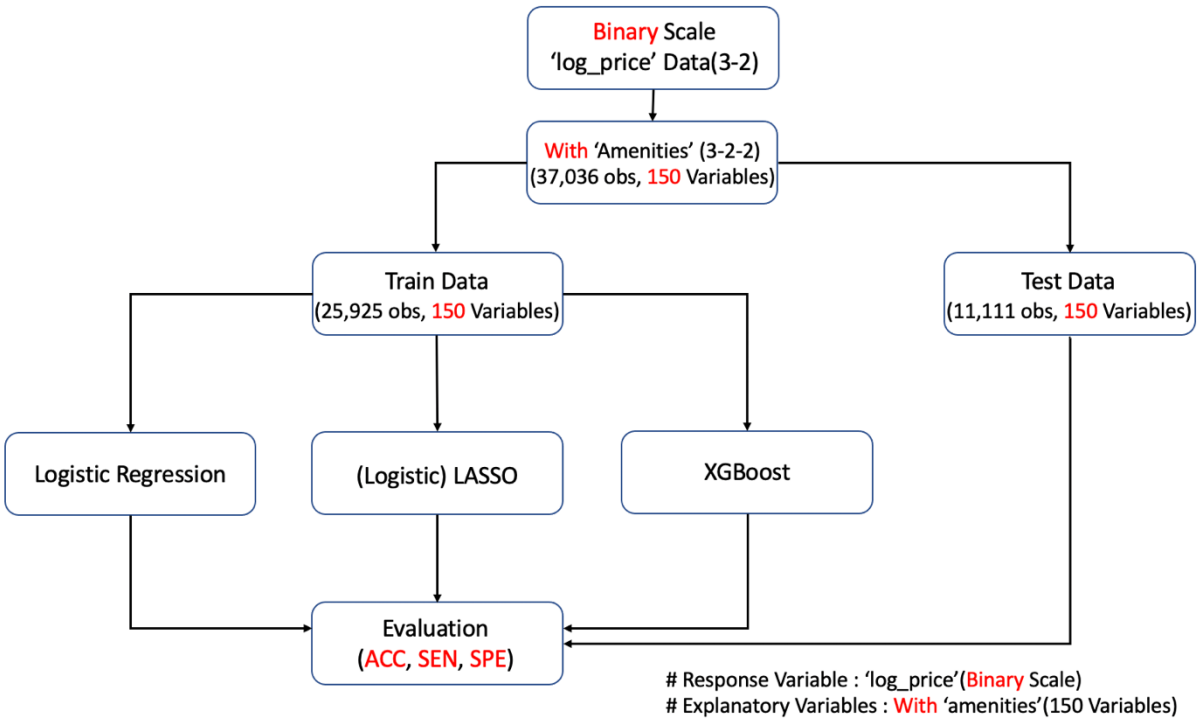


3-2. Binary Scale 'log_price'에 관한 분석 계획

3-2-1. Setting 1: Without 'amenities' Variables(13 Variables)



3-2-2. Setting 2: With 'amenities' Variables(150 Variables)



4. 분석 결과

4-1. Continuous Scale 'log_price'에 관한 분석

Multiple Linear Regression, LASSO, 그리고 XGBoost을 Train Data를 이용하여 적합한 모델을 바탕으로, Test Data에 예측하여 얻은 Evaluation 결과는 다음과 같다.

Setting	Statistical Method	RMSE	MAE	MAPE
Setting 1 w/o amenities (12 Variables)	Multiple Linear Regression	0.3955	0.3055	0.0653
	LASSO	0.3991	0.3081	0.0659
	XGBoost	0.3823	0.2944	0.0628
Setting 2 w/ amenities (149 Variables)	Multiple Linear Regression	0.3775	0.2902	0.0621
	LASSO	0.3829	0.2944	0.0631
	XGBoost	0.3619	0.2760	0.0589

각 Setting과 모형에 따른 세부적인 분석 결과는 아래에서 소개한다. 이 때 앞으로 소개될 결과들에서 각 범주형 변수에 대한 Reference Category는 아래의 표와 같다.

	Reference Category
property_type	property_typeApartment
room_type	room_typeentire home/apt
cancellation_policy	cancellation_policyflexible
city	cityBoston

4-1-1. Setting 1 : Without 'amenities' (13 Variables)

4-1-1-1. Multiple Linear Regression

전처리한 데이터에서 'amenities'를 제외한 나머지 변수들(property_type, room_type, accommodates, bathrooms, cancellation_policy, cleaning_fee, city, host_identify_verified, instant_bookable, review_scores_rating, bathrooms, beds)만을 이용하여 Multiple Linear Regression을 적합한 결과는 다음과 같다.

	Predictors	Estimate	Std. Error	Pr(>|t|)
(Intercept)	**(Intercept)**	**3.179**	0.050	**0.000**
property_type	property_typeBed & Breakfast	**0.086**	0.029	**0.003**
	property_typeBoat	0.075	0.109	0.491
	property_typeBoutique hotel	-0.015	0.086	0.866
	property_typeBungalow	-0.040	0.031	0.202
	property_typeCabin	-0.092	0.067	0.169
	property_typeCamper/RV	**-0.279**	0.066	**0.000**
	property_typeCastle	0.161	0.161	0.317
	property_typeCave	0.270	0.278	0.331
	property_typeChalet	0.054	0.393	0.892
	property_typeCondominium	**0.076**	0.014	**0.000**
	property_typeDorm	**-0.325**	0.059	**0.000**
	property_typeEarth House	0.412	0.393	0.295
	property_typeGuest suite	-0.094	0.058	0.107
	property_typeGuesthouse	**-0.113**	0.028	**0.000**
	property_typeHostel	**-0.592**	0.082	**0.000**
	property_typeHouse	**-0.064**	0.007	**0.000**
	property_typeHut	-0.370	0.197	0.060

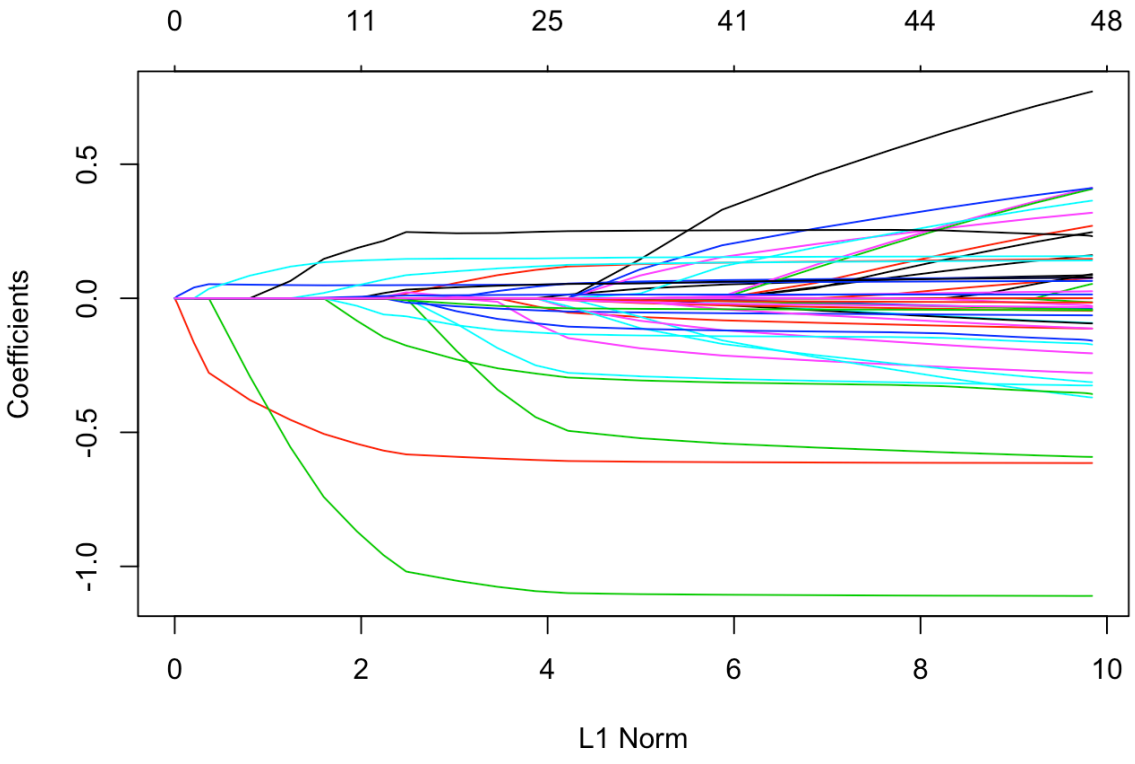
	property_typeIn-law	**-0.205**	0.068	**0.002**
	property_typeIsland	**0.771**	0.393	**0.050**
	property_typeLoft	**0.147**	0.018	**0.000**
	property_typeOther	-0.042	0.029	0.149
	property_typeServiced apartment	0.001	0.176	0.998
	property_typeTent	**-0.313**	0.131	**0.017**
	property_typeTimeshare	**0.319**	0.161	**0.047**
	property_typeTipi	0.247	0.278	0.375
	property_typeTownhouse	**-0.047**	0.017	**0.005**
	property_typeTrain	0.407	0.393	0.300
	property_typeTreehouse	**0.411**	0.197	**0.037**
	property_typeVacation home	0.364	0.227	0.109
	property_typeVilla	**-0.112**	0.057	**0.049**
	property_typeYurt	0.090	0.278	0.746
room_type	**room_typePrivate room**	**-0.614**	0.006	**0.000**
	room_typeShared room	**-1.110**	0.018	**0.000**
accommodates	**accommodates**	**0.065**	0.002	**0.000**
bathrooms	**bathrooms**	**0.141**	0.006	**0.000**
cancellation_policy	**cancellation_policymoderate**	**0.026**	0.008	**0.001**
	cancellation_policystrict	**0.078**	0.007	**0.000**
cleaning_fee	cleaning_feeTrue	0.000	0.007	0.955
city	**cityChicago**	**-0.357**	0.015	**0.000**
	cityDC	**-0.159**	0.014	**0.000**
	cityLA	**-0.173**	0.012	**0.000**
	cityNYC	**-0.029**	0.012	**0.014**
	citySF	**0.232**	0.014	**0.000**
host_identify_ verifed	**host_identity_verifiedt**	**0.016**	0.006	**0.007**
instant_bookable	**instant_bookablet**	**-0.045**	0.006	**0.000**
review_scores_ rating	**review_scores_rating**	**0.014**	0.001	**0.000**
bedrooms	**bedrooms**	**0.157**	0.005	**0.000**
beds	**beds**	**-0.033**	0.004	**0.000**

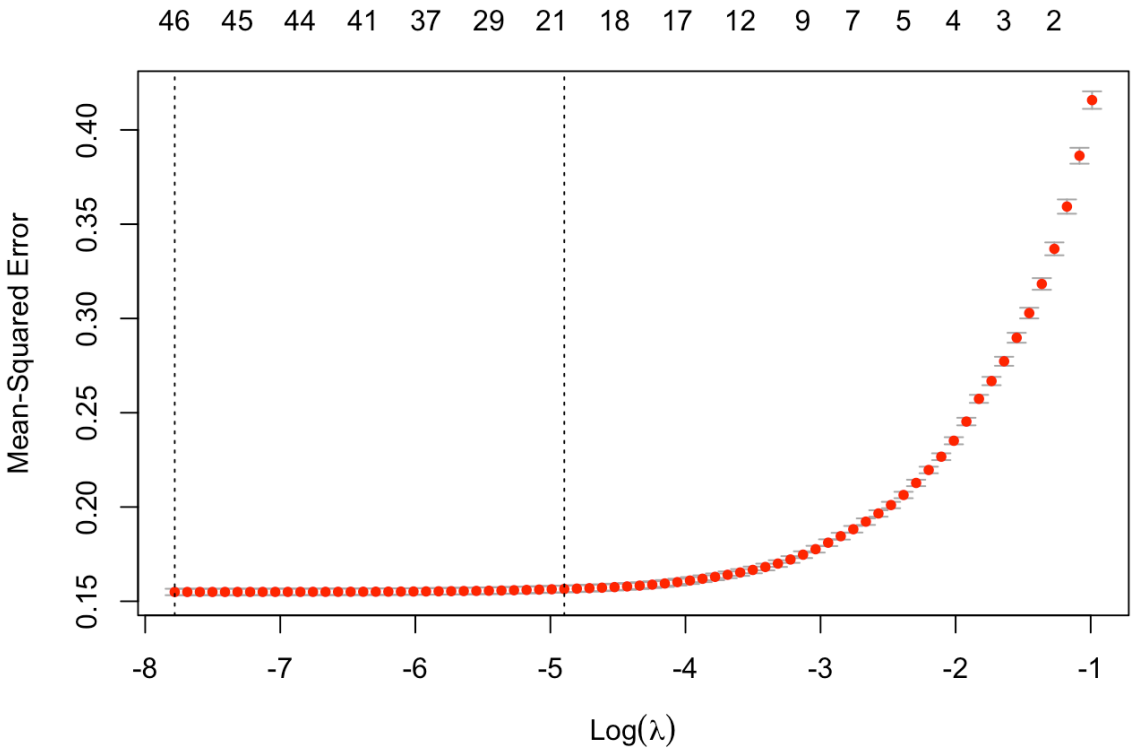
위의 결과로부터 'room_type', 'accommodates', 'bathrooms', 'cancellation_policy', 'city', 'host_identify_verified', 'instant_bookable', 'review_scores_rating', 'bathrooms', 'beds'와 같은 변수들이 모두 유의한 것으로 확인되었다. 즉, 'room_type'이 entire home/apt인 경우에 비하여 private room이거나 shared room인 경우 더 가격이 비싸지는 것으로 보이며, 'accommodates'(숙박 인원)이 많을수록, 'bathrooms', 'bedrooms'도 많을수록 가격이 올라가는 것으로 보인다. 특이하게도 'beds'는 가격과 작은 음의 계수를 가지고 있는데, 이는 'beds' 변수는 'bedrooms' 등의 변수들과 상관관계가 클 것이고, 따라서 'bedrooms'의 계수에 이미 'beds'의 정보가 반영돼서 그런 것으로 판단한다.

숙박 업소가 위치한 도시를 나타내는 'city' 변수가 전부 유의한 것도 상식적인 결과물이다. 샌프란시스코의 숙박 비용이 가장 높고, 시카고의 숙박 비용이 가장 낮은 것으로 보인다. 'cancellation_policy'와 같은 경우는 strict한 업체가 moderate한 업체보다 가격이 높은 것으로 보인다. 이는 숙박 업소의 예약과 취소에 대하여 엄격한 규정을 두는 곳일수록 가격이 높다는 것으로 해석할 수 있다.

'instant_bookable'도 마찬가지인데, 'instant_bookable'이 true인 경우, 즉 즉석으로 예약 가능한 경우의 가격이 좀 더 낮다. 이도 마찬가지로 상관성을 생각하면 가격이 높은 숙소들이 즉석 예약이 안 될 가능성이 높다고 생각하면 자연스럽다.

4-1-1-2. LASSO





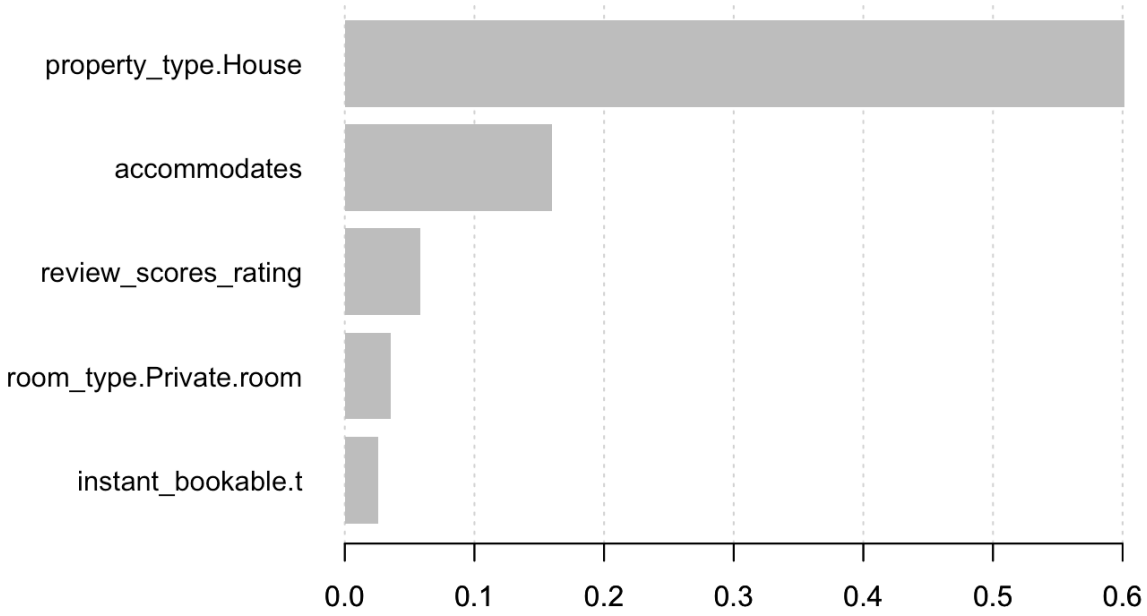
	Predictors	Estimate
(Intercept)	**(Intercept)**	**3.220**
property_type	**property_typeBed & Breakfast**	**0.091**
	property_typeBoat	**0.127**
	property_typeBoutique hotel	**0.005**
	property_typeBungalow	**-0.022**
	property_typeCabin	**-0.119**
	property_typeCamper/RV	**-0.316**
	property_typeCasa particular	0.000
	property_typeCastle	**0.266**
	property_typeCave	**0.218**
	property_typeChalet	**-0.010**
	property_typeCondominium	**0.082**
	property_typeDorm	**-0.339**
	property_typeEarth House	**0.003**
	property_typeGuest suite	**-0.125**
	property_typeGuesthouse	**-0.109**
	property_typeHostel	**-0.550**

	property_typeHouse	**-0.061**
	property_typeHut	**-0.291**
	property_typeIn-law	**-0.188**
	property_typeIsland	**0.690**
	property_typeLighthouse	0.000
	property_typeLoft	**0.161**
	property_typeOther	**-0.015**
	property_typeParking Space	0.000
	property_typeServiced apartment	**-0.002**
	property_typeTent	**-0.315**
	property_typeTimeshare	**0.282**
	property_typeTipi	**0.192**
	property_typeTownhouse	**-0.049**
	property_typeTrain	**0.333**
	property_typeTreehouse	**0.372**
	property_typeVacation home	**0.239**
	property_typeVilla	**-0.071**
	property_typeYurt	0.000
room_type	**room_typePrivate room**	**-0.611**
	room_typeShared room	**-1.103**
accommodates	**accommodates**	**0.066**
bathrooms	**bathrooms**	**0.143**
cancellation_policy	**cancellation_policymoderate**	**0.022**
	cancellation_policystrict	**0.075**
cleaning_fee	**cleaning_feeTrue**	**0.000**
city	**cityChicago**	**-0.346**
	cityDC	**-0.146**
	cityLA	**-0.162**
	cityNYC	**-0.018**
	citySF	**0.238**
host_identity_verified	host_identity_verifiedt	0.000
instant_bookable	**instant_bookablet**	**-0.046**
review_scores_rating	**review_scores_rating**	**0.014**
bedrooms	**bedrooms**	**0.154**
beds	**beds**	**-0.032**

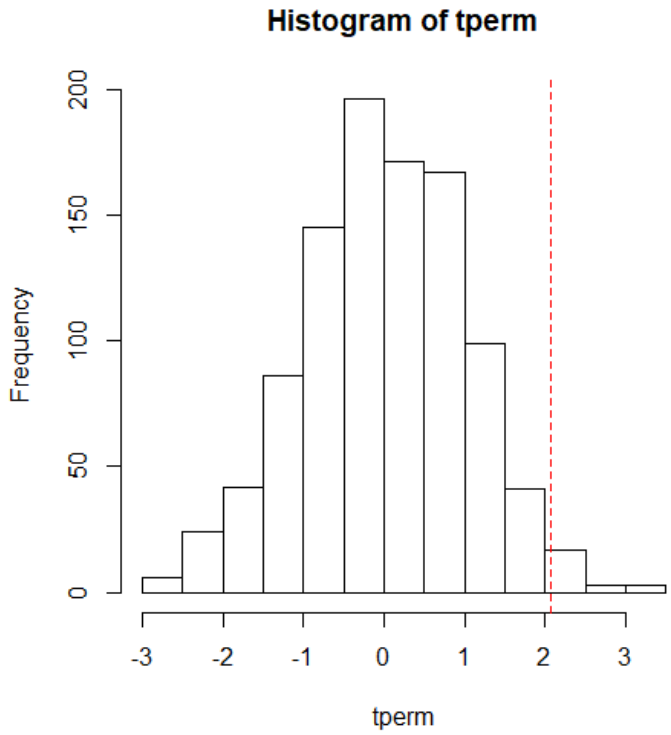
앞의 4-1-1-1의 Multiple Linear Regression과 상당히 비슷한 결과를 도출하였다. 선택된 변수들이 비슷하며, 계수들은 약간씩 차이가 나지만 전체적인 방향성은 같은 것을 확인할 수 있다.

이는 위에서 제시한 λ-plot을 볼 때 λ의 값이 0에 가깝게 선택되고, 그래서 Multiple Linear Regression과 큰 차이가 없어져 그런 것으로 판단된다. 실제로 L1-norm 이 조금만 커져도 Variable Selection의 효과가 거의 사라지는 것을 확인할 수 있다.

4-1-1-3. XGBoost



'property_type'이 house인지의 여부가 가격 여부에 가장 압도적인 영향(대략 60%)을 끼치는 것으로 드러났다. 실제로, 아래와 같이 'property_type'가 house가 아닌 데이터들의 'log_price'와 house인 데이터의 'log_price'를 이용하여 Permutation Test를 시행하였을 때도 house인지의 여부가 큰 영향을 끼치는 것으로 보인다. 아래에서 빨간 선이 property type이 house인 데이터들이다.



마찬가지로 'accommodates'에서도 Permutation Test를 시행하였을 때 각각의 'accommodates' 수에 대해 귀무가설의 p-value가 2*10^-16 이하로 굉장히 작게 나와, 각 'accommodates' 수에 따라 'log_price'의 차이가 있다는 뚜렷한 증거가 있다.

4-1-2. Setting 2 : With 'amenities' (150 Variables)

4-1-2-1. Multiple Linear Regression

변수의 개수가 너무 많은 관계로 아래 표에서는 p-value가 0.05 이하인 계수들만 기록하였다.

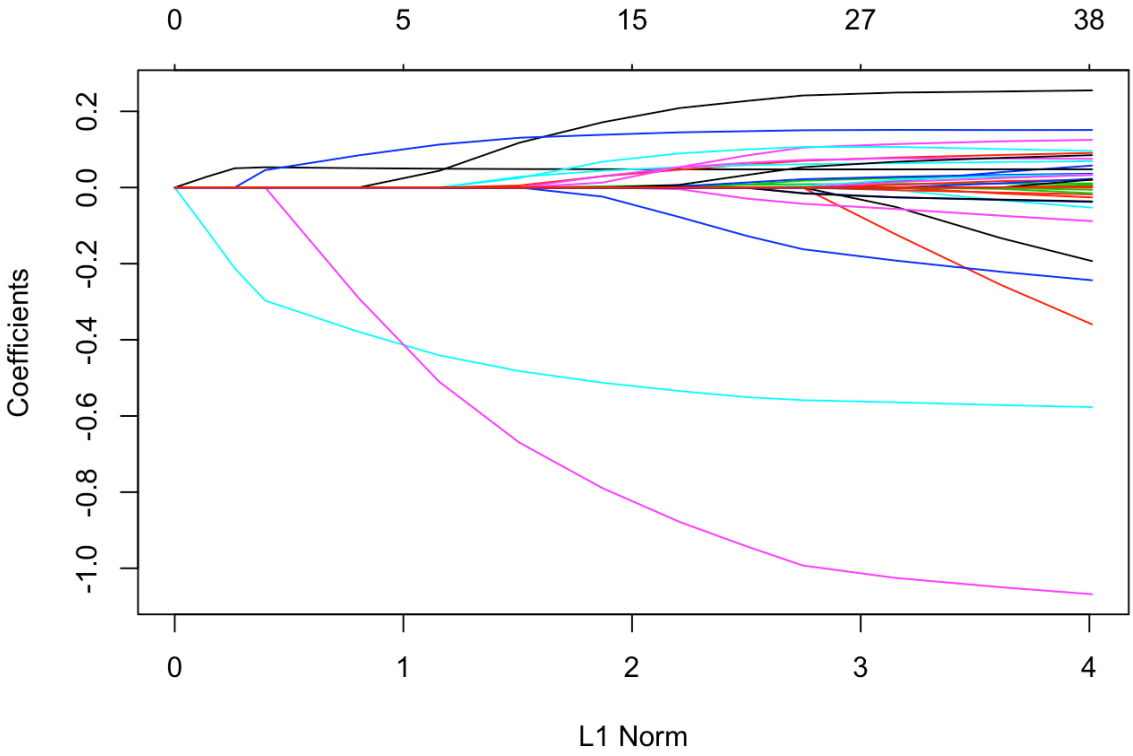
	Predictors	Estimate	Std. Error	Pr(>|t|)
(Intercept)	**(Intercept)**	**3.270**	0.052	**0.000**
property_type	**property_typeBed & Breakfast**	**0.088**	0.028	**0.002**
	property_typeCamper/RV	**-0.127**	0.063	**0.046**
	property_typeCondominium	**0.036**	0.014	**0.010**
	property_typeDorm	**-0.301**	0.057	**0.000**
	property_typeHostel	**-0.607**	0.079	**0.000**
	property_typeHouse	**-0.016**	0.007	**0.017**
	property_typeIsland	**0.778**	0.376	**0.038**
	property_typeLoft	**0.115**	0.018	**0.000**
	property_typeTent	**-0.263**	0.126	**0.036**
	property_typeTreehouse	**0.556**	0.188	**0.003**

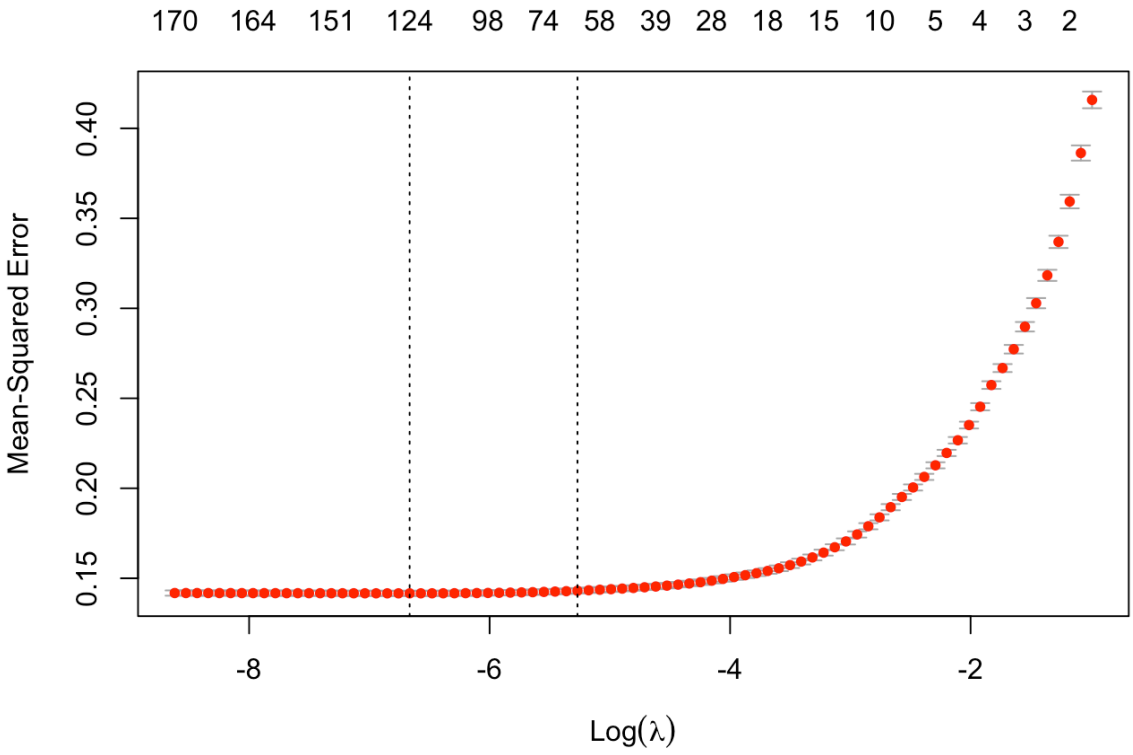
room_type	room_typePrivate room	-0.585	0.007	0.000
	room_typeShared room	-1.090	0.017	0.000
accommodates	accommodates	0.059	0.002	0.000
bathrooms	bathrooms	0.110	0.006	0.000
cancellation_policy	cancellation_policymoderate	0.023	0.007	0.002
	cancellation_policystrict	0.060	0.007	0.000
city	cityChicago	-0.356	0.015	0.000
	cityDC	-0.161	0.014	0.000
	cityLA	-0.178	0.013	0.000
	cityNYC	0.041	0.012	0.001
	citySF	0.217	0.014	0.000
instant_bookable	instant_bookablet	-0.028	0.005	0.000
review_scores_rating	review_scores_rating	0.013	0.000	0.000
bedrooms	bedrooms	0.158	0.005	0.000
beds	beds	-0.030	0.003	0.000
amenities	Heating.TRUE	0.028	0.012	0.018
	Family.kid.friendly.TRUE	0.016	0.005	0.003
	Hair.dryer.TRUE	0.048	0.007	0.000
	Iron.TRUE	-0.021	0.007	0.004
	Shampoo.TRUE	0.035	0.006	0.000
	Hangers.TRUE	-0.019	0.008	0.018
	TV.TRUE	0.048	0.006	0.000
	Cable.TV.TRUE	0.076	0.005	0.000
	Breakfast.TRUE	-0.024	0.008	0.003
	Buzzer.wireless.intercom.TRUE	0.071	0.006	0.000
	Indoor.fireplace.TRUE	0.103	0.008	0.000
	First.aid.kit.TRUE	-0.014	0.006	0.017
	Pool.TRUE	0.043	0.012	0.000
	Free.parking.on.premises.TRUE	-0.056	0.006	0.000
	Gym.TRUE	0.039	0.011	0.001
	Hot.tub.TRUE	0.022	0.010	0.028
	Doorman.TRUE	0.134	0.013	0.000

	Cat.s..TRUE	**-0.027**	0.013	**0.032**
	Lock.on.bedroom.door.TRUE	**-0.043**	0.006	**0.000**
	Private.entrance.TRUE	**-0.053**	0.008	**0.000**
	Suitable.for.events.TRUE	**0.062**	0.011	**0.000**
	Elevator.TRUE	**0.080**	0.009	**0.000**
	Microwave.TRUE	**-0.113**	0.023	**0.000**
	Oven.TRUE	**0.079**	0.035	**0.023**
	Smoking.allowed.TRUE	**-0.032**	0.012	**0.007**
	Outlet.covers.TRUE	**-0.093**	0.032	**0.004**
	Dishwasher.TRUE	**0.115**	0.019	**0.000**
	Step.free.access..1TRUE	**-0.128**	0.056	**0.023**
	Babysitter.recommendations.TRUE	**0.061**	0.024	**0.012**
	Free.parking.on.street.TRUE	**-0.148**	0.060	**0.014**
	Private.living.room.TRUE	**0.063**	0.016	**0.000**
	BBQ.grill.TRUE	**0.077**	0.029	**0.007**
	Host.greets.you.TRUE	**-0.034**	0.018	**0.055**
	Doorman..1TRUE	**0.104**	0.039	**0.008**
	Stair.gates.TRUE	**0.053**	0.036	**0.140**
	Window.guards.TRUE	**-0.071**	0.024	**0.003**
	Crib.TRUE	**0.133**	0.032	**0.000**
	Doorman.Entry.TRUE	**-0.079**	0.032	**0.014**
	Baby.bath.TRUE	**-0.121**	0.037	**0.001**
	Beachfront.TRUE	**-0.272**	0.100	**0.007**
	Lake.access.TRUE	**-0.493**	0.173	**0.004**

기존의 변수들은 앞의 4-1-1-1과 상당히 비슷한 것을 확인할 수 있다. 그 밖에도 'amenities'에서도 다수의 유의한 변수들이 발견되었다. 대부분 amenity들은 상식적으로 가격과 양의 상관관계를 가지고 있는데, 간혹 음의 상관관계를 가지고 있는 amenity 변수들도 있다. 이는 4-1-1-1에서도 언급했듯 인과관계가 아닌 상관관계에서 오는 효과일 수도 있고, 각 amenity 변수 간의 상관관계 때문에 한 변수의 기여도가 다른 변수에서 이미 반영되었기 때문에 나타나는 효과일 수도 있을 것으로 판단한다.

4-1-2-2. LASSO


0
5
15
27
38
Coefficients
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
0
1
2
3
4
L1 Norm


170
164
151
124
98
74
58
39
28
18
15
10
5
4
3
2
Mean-Squared Error
0.40
0.35
0.30
0.25
0.20
0.15
-8
-6
-4
-2
Log(λ)

마찬가지로 변수의 개수가 매우 많은 관계로, 회귀계수가 0이 아닌 변수들만 기록하였다.

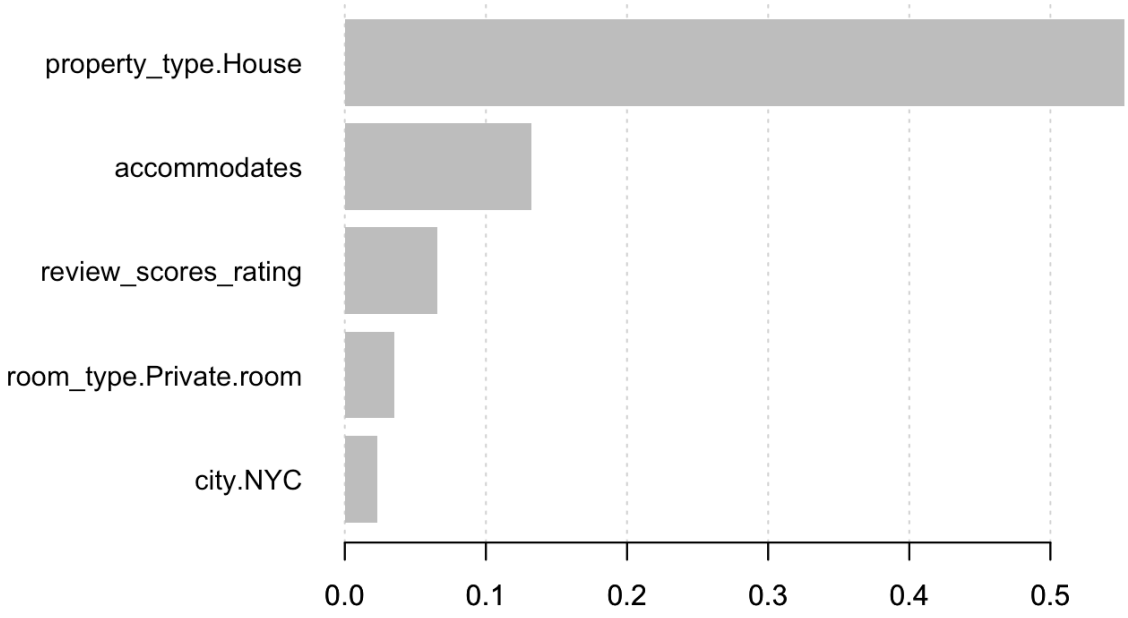
	Predictors	Estimate
(Intercept)	(Intercept)	**3.454**
property_type	**property_typeCondominium**	**0.004**
	property_typeDorm	**-0.205**
	property_typeHostel	**-0.333**
	property_typeHouse	**-0.006**
	property_typeLoft	**0.076**
room_type	**room_typePrivate room**	**-0.572**
	room_typeShared room	**-1.061**
accommodates	**accommodates**	**0.049**
bathrooms	**bathrooms**	**0.094**
cancellation_policy	**cancellation_policystrict**	**0.033**
city	**cityChicago**	**-0.243**
	cityDC	**-0.052**
	cityLA	**-0.087**
	citySF	**0.254**
instant_bookable	**instant_bookablet**	**-0.018**
review_scores_rating	**review_scores_rating**	**0.011**
bedrooms	**bedrooms**	**0.148**
amenities	**Hair.dryer.TRUE**	**0.002**
	Dryer.TRUE	**0.025**
	Shampoo.TRUE	**0.005**
	TV.TRUE	**0.035**
	Cable.TV.TRUE	**0.068**
	Buzzer.wireless.intercom.TRUE	**0.080**
	Indoor.fireplace.TRUE	**0.090**
	Pool.TRUE	**0.009**
	Free.parking.on.premises.TRUE	**-0.039**
	Gym.TRUE	**0.029**
	Doorman.TRUE	**0.131**
	Lock.on.bedroom.door.TRUE	**-0.041**
	Private.entrance.TRUE	**-0.023**

	Self.Check.In.TRUE	**-0.016**
	Suitable.for.events.TRUE	**0.029**
	Elevator.TRUE	**0.096**

주요 13개의 변수에 대한 흐름은 Multiple Linear Regression과 비슷하다. 하지만 'amenities' 부분에서는 Variable Selection이 강하게 이루어진 것을 볼 수 있다. 기존의 Multiple Linear Regression에서 41개의 'amenities' 관련 변수를 택했던 것에 비하여, LASSO에서는 16개의 변수밖에 선택하지 않았다.

마찬가지로 λ-plot을 보면 λ가 0에 가까운 값이 선택되는데, 그럼에도 불구하고 Variable Selection의 효과가 존재하는 것으로 판단된다.

4-1-2-3. XGBoost



Setting 2와 같이 'amenities'의 변수들을 추가하였을 때에도 5번째로 Importance가 높은 것으로 나온 city.NYC를 제외하고는 순위에 변동이 없었다. amenity의 항목들은 가격에 큰 영향을 미치지 않는 것으로 추정된다.

4-2. Binary Scale log_price에 관한 분석

Logistic Regression, (Logistic) LASSO, 그리고 XGBoost을 Train Data를 이용하여 적합한 모델을 바탕으로, Test Data에 예측하여 얻은 Evaluation 결과는 다음과 같다.

Setting	Statistical Method	Accuracy	Sensitivity	Specificity
Setting 1 w/o amenities (12 Variables)	Logistic Regression	0.8212	0.7963	0.8464
	(Logistic) LASSO	0.8195	0.7738	0.8660
	XGBoost	0.8210	0.7954	0.8469
Setting 2 w/ amenities (149 Variables)	Logistic Regression	0.8272	0.8086	0.8460
	(Logistic) LASSO	0.8241	0.7867	0.8622
	XGBoost	0.8330	0.8126	0.8538

각 모형에 따른 세부적인 분석 결과는 아래와 같다.

4-2-1. Setting 1 : Without 'amenities' (13 Variables)

4-2-1-1. Logistic Regression

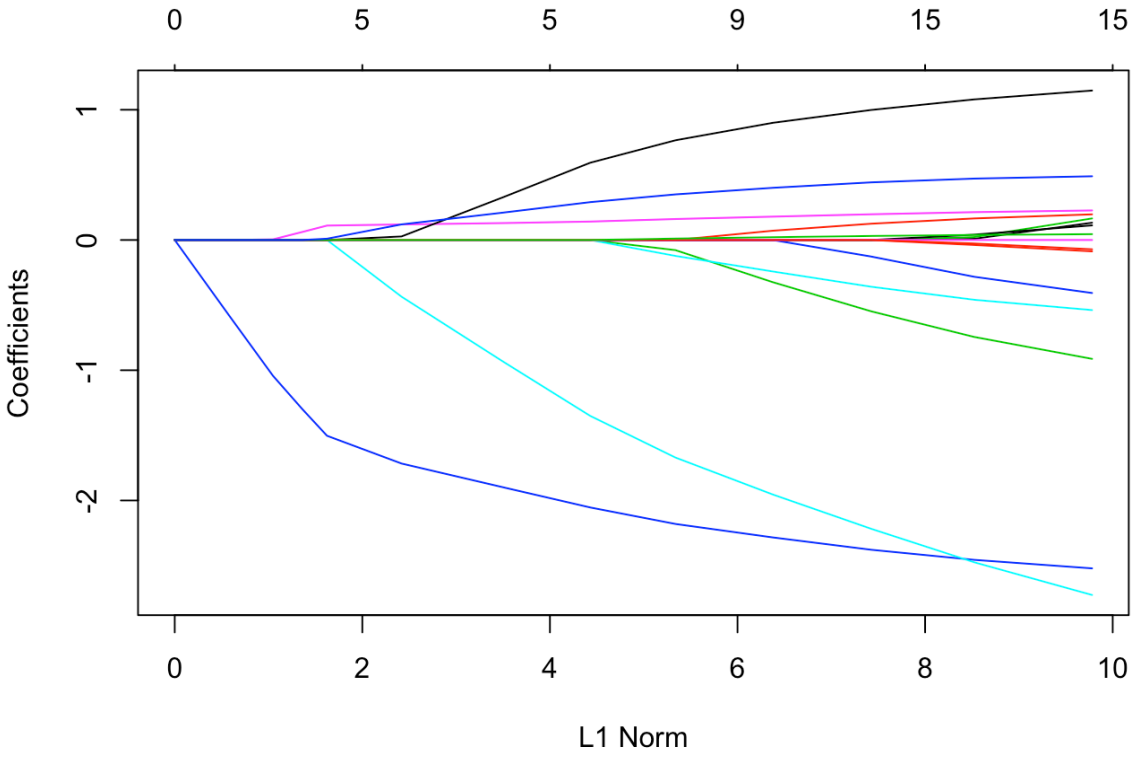
	Predictors	Estimate	Std. Error	Pr(>|z|)
(Intercept)	**(Intercept)**	**-7.030**	0.372	**0.000**
property_type	**property_typeBed & Breakfast**	**0.531**	0.207	**0.010**
	property_typeBoat	0.290	0.686	0.672
	property_typeBoutique hotel	-0.358	0.632	0.572
	property_typeBungalow	-0.257	0.178	0.148
	property_typeCabin	-0.700	0.388	0.072
	property_typeCamper/RV	**-1.090**	0.414	**0.008**
	property_typeCastle	2.387	1.270	0.060
	property_typeCave	14.783	594.557	0.980
	property_typeChalet	11.460	882.743	0.990
	property_typeCondominium	**0.526**	0.110	**0.000**
	property_typeDorm	-12.576	111.596	0.910
	property_typeEarth House	12.394	882.743	0.989
	property_typeGuest suite	**-1.116**	0.365	**0.002**
	property_typeGuesthouse	**-0.426**	0.160	**0.008**

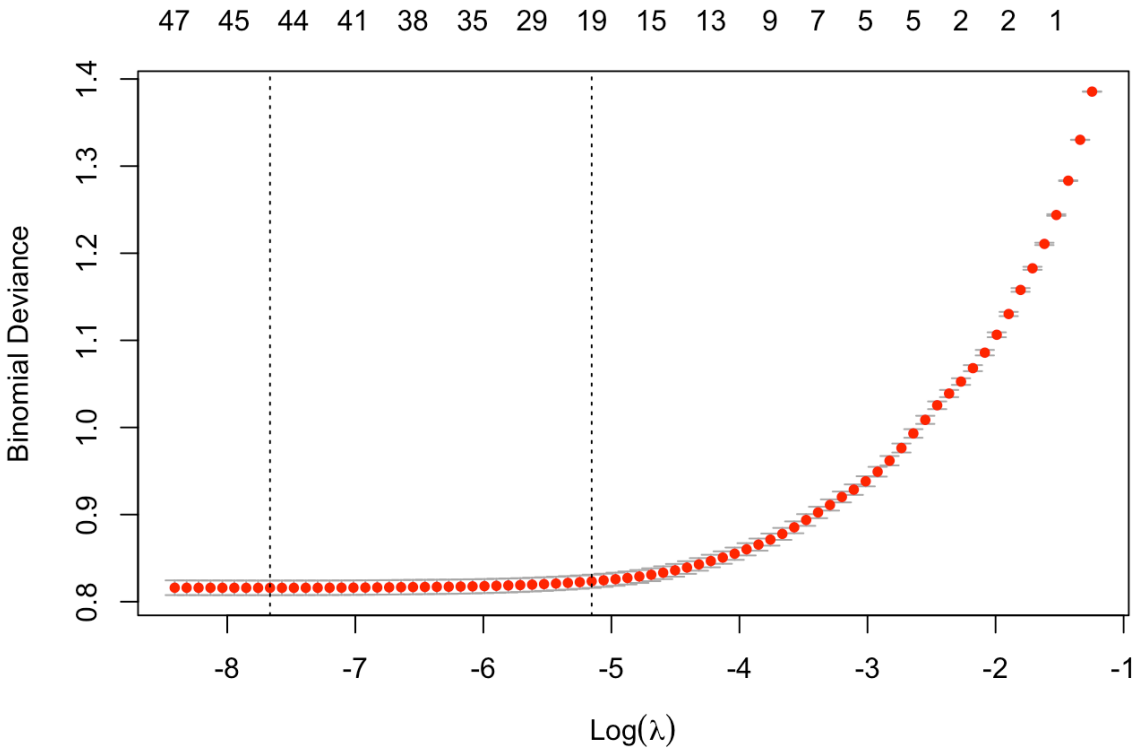
	property_typeHostel	-1.049	0.886	0.237
	property_typeHouse	**-0.322**	0.049	**0.000**
	property_typeHut	-14.246	389.103	0.971
	property_typeIn-law	-0.679	0.476	0.154
	property_typeIsland	16.872	882.743	0.985
	property_typeLoft	**0.681**	0.133	**0.000**
	property_typeOther	-0.259	0.195	0.185
	property_typeServiced apartment	-0.133	1.168	0.910
	property_typeTent	-0.024	1.219	0.984
	property_typeTimeshare	0.987	1.601	0.538
	property_typeTipi	0.741	2.116	0.726
	property_typeTownhouse	**-0.334**	0.134	**0.013**
	property_typeTrain	14.140	882.743	0.987
	property_typeTreehouse	-1.457	3.273	0.656
	property_typeVacation home	14.415	422.194	0.973
	property_typeVilla	-0.074	0.541	0.891
	property_typeYurt	-13.544	546.565	0.980
room_type	**room_typePrivate room**	**-2.811**	0.046	**0.000**
	room_typeShared room	**-3.904**	0.219	**0.000**
accommodates	**accommodates**	**0.323**	0.021	**0.000**
bathrooms	**bathrooms**	**0.435**	0.050	**0.000**
cancellation_policy	**cancellation_policymoderate**	**0.145**	0.054	**0.008**
	cancellation_policystrict	**0.408**	0.052	**0.000**
cleaning_fee	cleaning_feeTrue	0.037	0.051	0.462
city	**cityChicago**	**-1.844**	0.110	**0.000**
	cityDC	**-1.138**	0.103	**0.000**
	cityLA	**-1.086**	0.089	**0.000**
	cityNYC	**-0.318**	0.086	**0.000**
	citySF	**1.169**	0.102	**0.000**
host_identity_verified	host_identity_verifiedt	0.025	0.042	0.553
instant_bookable	**instant_bookablet**	**-0.229**	0.039	**0.000**
review_scores_rating	**review_scores_rating**	**0.068**	0.004	**0.000**

bedrooms	**bedrooms**	**0.574**	0.034	**0.000**
beds	**beds**	**-0.091**	0.031	**0.003**

계수들의 부호 및 크기와 같은 전반적인 경향성 자체는 Continuous Scale에서의 계수들과 같다. 다만 'host_identity_verified'와 같이 Continuous Scale에서는 유의한 변수였지만 Binary Scale에서는 유의하지 않다고 판단된 변수들도 존재한다. 이는 아마도 일정 가격대 이상(비싼 숙소)에서의 세부 가격을 결정하던 변수이거나, 일정 가격대 이하(싼 숙소)에서의 세부 가격을 결정하던 변수였을 가능성이 있다.

4-2-1-2. (Logistic) LASSO





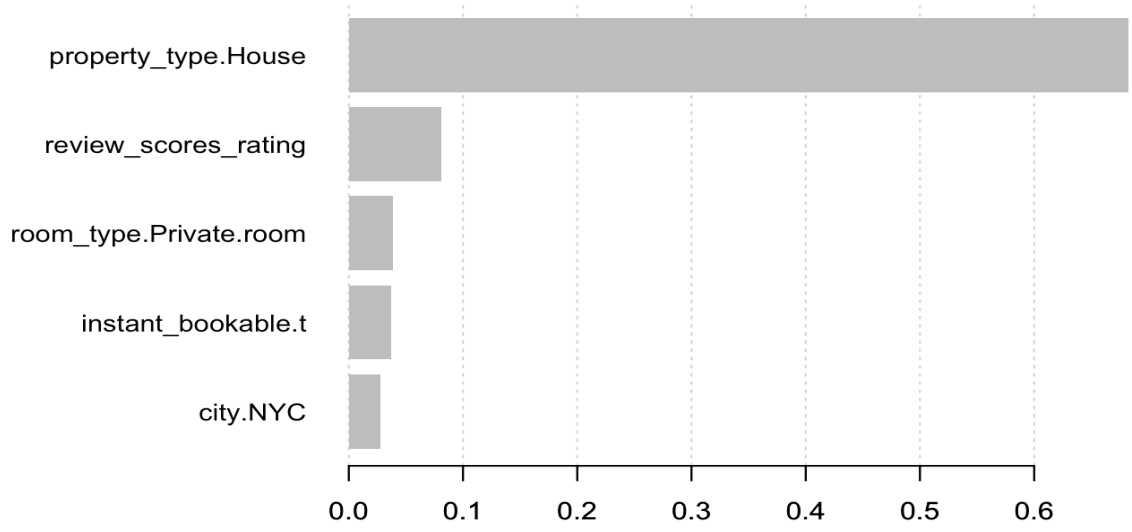
	Predictors	Estimate
(Intercept)	**(Intercept)**	**-0.023**
property_type	property_typeBed & Breakfast	0.000
	property_typeBoat	0.000
	property_typeBoutique hotel	0.000
	property_typeBungalow	0.000
	property_typeCabin	0.000
	property_typeCamper/RV	0.000
	property_typeCasa particular	0.000
	property_typeCastle	0.000
	property_typeCave	0.000
	property_typeChalet	0.000
	property_typeCondominium	**0.029**
	property_typeDorm	0.000
	property_typeEarth House	0.000
	property_typeGuest suite	0.000

	property_typeGuesthouse	**-0.016**
	property_typeHostel	0.000
	property_typeHouse	**-0.010**
	property_typeHut	0.000
	property_typeIn-law	0.000
	property_typeIsland	0.000
	property_typeLighthouse	0.000
	property_typeLoft	**0.029**
	property_typeOther	0.000
	property_typeParking Space	0.000
	property_typeServiced apartment	0.000
	property_typeTent	0.000
	property_typeTimeshare	0.000
	property_typeTipi	0.000
	property_typeTownhouse	0.000
	property_typeTrain	0.000
	property_typeTreehouse	0.000
	property_typeVacation home	0.000
	property_typeVilla	0.000
	property_typeYurt	0.000
room_type	**room_typePrivate room**	**-0.515**
	room_typeShared room	**-0.529**
accommodates	**accommodates**	**0.024**
bathrooms	bathrooms	0.000
cancellation_policy	cancellation_policymoderate	0.000
	cancellation_policystrict	**0.030**
cleaning_fee	cleaning_feeTrue	0.000
city	**cityChicago**	**-0.126**
	cityDC	**-0.051**
	cityLA	**-0.078**
	cityNYC	0.000
	citySF	**0.159**
host_identity_verfied	host_identity_verifiedt	0.000

instant_bookable	**instant_bookablet**	**-0.007**
review_scores_rating	**review_scores_rating**	**0.006**
bedrooms	**bedrooms**	**0.073**
beds	beds	0.000

앞서 확인하였던 다른 LASSO 모형에 비하여 Variable Selection이 강하게 이루어진 모습을 보인다. 이 경우 단순 Logistic Regression보다 예측 결과가 낮았는데, 너무 과도한 Variable Selection으로 인하여 단순 Logistic regression에 비해 충분한 정보를 얻지 못해서 그런 것으로 판단하였다. 이 경우는 λ-plot에서도 상당히 적은 변수들만 존재하는 것을 확인할 수 있다.

4-2-1-3. XGBoost



Continuous Scale을 이용하지 않고 Binary Scale을 이용하여 예측을 하는 경우, 'accommodates'가 상위 2번째에서 사라지고, 5위권 밖이었던 'review_scores_rating'이 2위를 차지하였다. 가격이 비싼 것과 싼 것(중간값 기준으로)을 예측하는 기준에는 'review_scores_rating'가 중요한 역할을 차지하는 것으로 보인다.

이에 대하여 더 구체적으로 확인하기 위하여 'review_scores_rating'에 Permutation Test를 시행하였을 때 각각의 'review_scores_rating' 에 대해 귀무가설('review_scores_rating'과 'log_price'간의 유의미한 관계가 없다)의 p-value가 2*10^-16 이하의 값으로 굉장히 작게 나온다. 이로부터 'review_scores_rating' 값에 따른 'log_price'의 차이가 있다는 뚜렷한 증거가 있다고 판단할 수 있다.

4-2-2. Setting 2 : With 'amenities' (150 Variables)

4-2-2-1. Logistic Regression

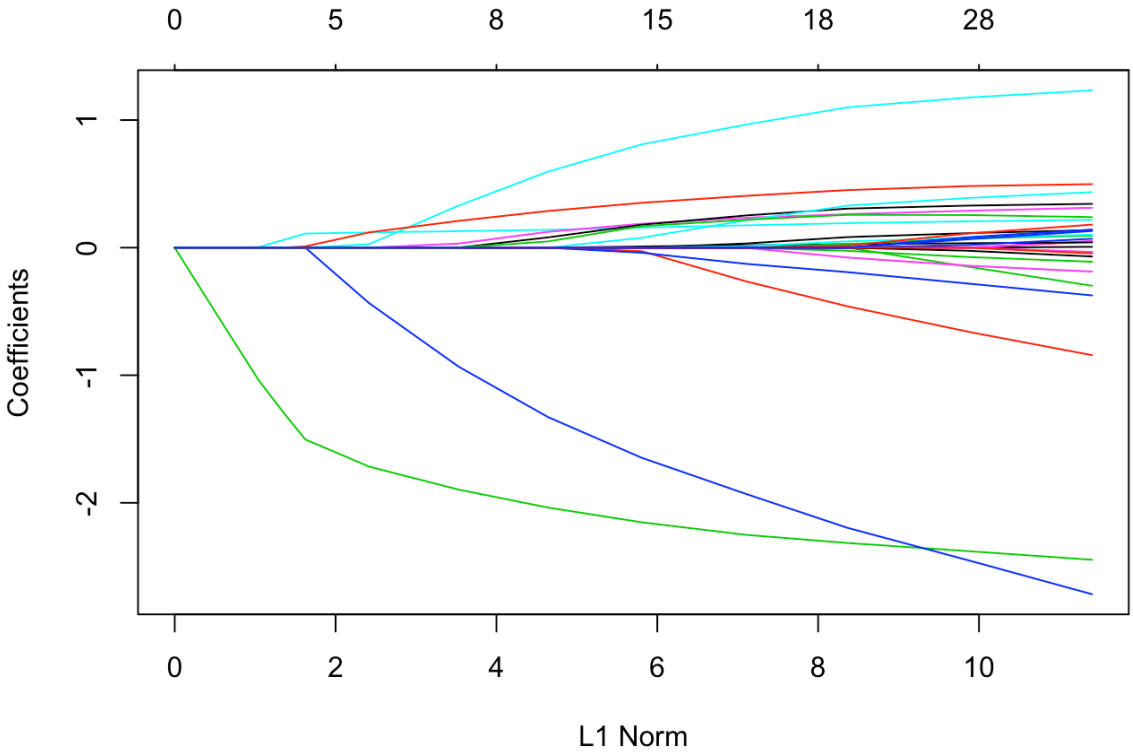
변수의 개수가 매우 많은 관계로 p-value가 0.05 이하인 변수들만 기록하였다.

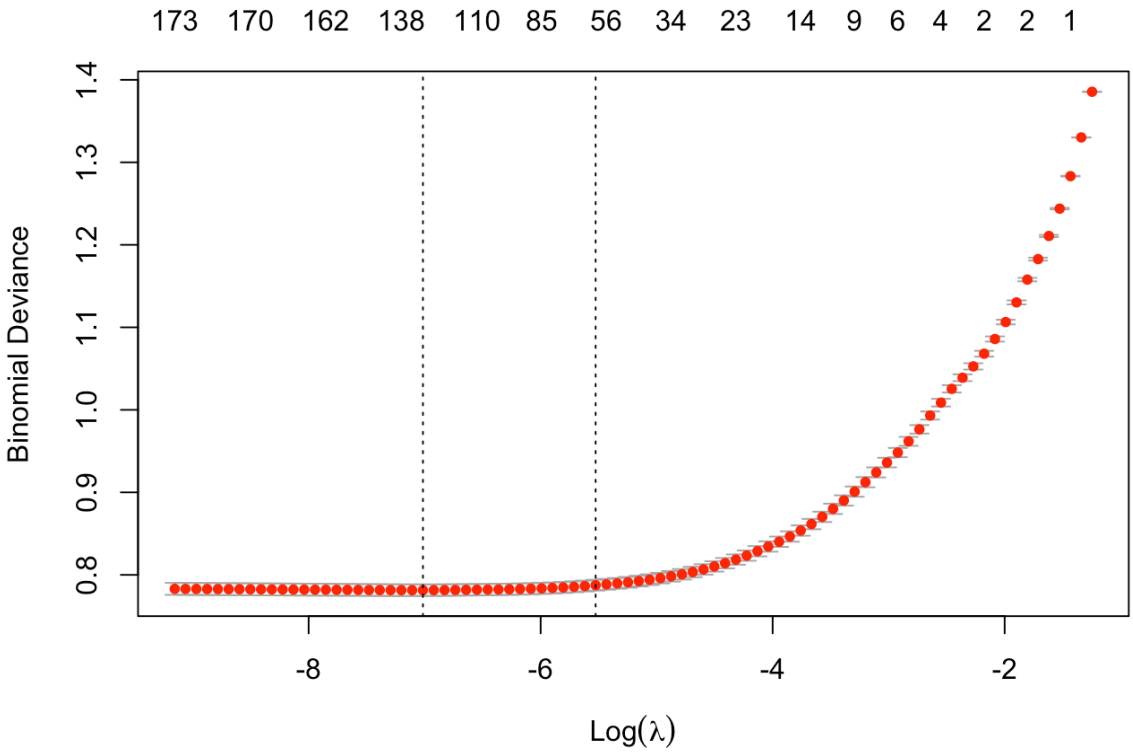
	Predictors	Estimate	Std. Error	Pr(>|z|)
(Intercept)	**(Intercept)**	**-7.199**	0.417	**0.000**
property_type	**property_typeBed & Breakfast**	**0.714**	0.217	**0.001**
	property_typeCondominium	**0.374**	0.114	**0.001**
	property_typeLoft	**0.605**	0.141	**0.000**
room_type	**room_typePrivate room**	**-2.777**	0.051	**0.000**
	room_typeShared room	**-3.943**	0.222	**0.000**
accommodates	**accommodates**	**0.301**	0.022	**0.000**
bathrooms	**bathrooms**	**0.343**	0.054	**0.000**
cancellation_policy	**cancellation_policymoderate**	**0.121**	0.056	**0.031**
	cancellation_policystrict	**0.308**	0.054	**0.000**
city	**cityChicago**	**-1.979**	0.118	**0.000**
	cityDC	**-1.239**	0.112	**0.000**
	cityLA	**-1.150**	0.100	**0.000**
	cityNYC	**-0.349**	0.094	**0.000**
	citySF	**1.148**	0.114	**0.000**
instant_bookable	**instant_bookablet**	**-0.166**	0.042	**0.000**
review_scores_rating	**review_scores_rating**	**0.066**	0.004	**0.000**
bedrooms	**bedrooms**	**0.579**	0.036	**0.000**
beds	**beds**	**-0.081**	0.032	**0.010**
amenities	**Kitchen.TRUE**	**0.174**	0.071	**0.015**
	Family.kid.friendly.TRUE	**0.115**	0.040	**0.004**
	Hair.dryer.TRUE	**0.228**	0.056	**0.000**
	Smoke.detector.TRUE	**-0.137**	0.068	**0.044**
	Shampoo.TRUE	**0.204**	0.049	**0.000**
	Hangers.TRUE	**-0.214**	0.061	**0.000**
	TV.TRUE	**0.193**	0.047	**0.000**
	Cable.TV.TRUE	**0.408**	0.041	**0.000**
	Breakfast.TRUE	**-0.170**	0.061	**0.005**

	Buzzer.wireless.intercom.TRUE	**0.376**	0.047	**0.000**
	Indoor.fireplace.TRUE	**0.408**	0.064	**0.000**
	Elevator.in.building.TRUE	**0.230**	0.085	**0.007**
	Free.parking.on.premises.TRUE	**-0.268**	0.046	**0.000**
	Gym.TRUE	**0.192**	0.086	**0.026**
	Wheelchair.accessible.TRUE	**-0.165**	0.076	**0.030**
	Doorman.TRUE	**0.588**	0.098	**0.000**
	Cat.s..TRUE	**-0.208**	0.103	**0.044**
	Lock.on.bedroom.door.TRUE	**-0.310**	0.049	**0.000**
	Pets.live.on.this.property.TRUE	**-0.215**	0.092	**0.019**
	Private.entrance.TRUE	**-0.269**	0.060	**0.000**
	Dishes.and.silverware.TRUE	**0.547**	0.274	**0.046**
	Elevator.TRUE	**0.228**	0.068	**0.001**
	Microwave.TRUE	**-0.348**	0.177	**0.050**
	Cooking.basics.TRUE	**-0.632**	0.250	**0.012**
	Smoking.allowed.TRUE	**-0.339**	0.095	**0.000**
	Outlet.covers.TRUE	**-0.592**	0.270	**0.028**
	Dishwasher.TRUE	**0.701**	0.154	**0.000**
	Step.free.access..1TRUE	**-0.966**	0.467	**0.039**
	Private.living.room.TRUE	**0.391**	0.121	**0.001**
	Private.bathroom.TRUE	**0.853**	0.407	**0.036**
	Doorman.Entry.TRUE	**-0.572**	0.236	**0.015**

Setting 2에서 'amenities'의 변수들을 추가한 결과, 추가하지 않았을 때에 비하여 기존의 변수들에 대한 결과도 상당히 달라졌다. 전반적인 경향성 자체는 같으나, 'property_type' 등에서 일부의 변수들만 유의한 것으로 선택한 모습을 볼 수 있다. 이는 'amenities'들과 'property_type'에 높은 상관관계가 존재하여 기존의 'property_type'으로 설명할 수 있는 부분을 'amenities'와 관련된 변수들로 설명할 수 있기 때문으로 판단된다.

4-2-2-2. (Logistic) LASSO





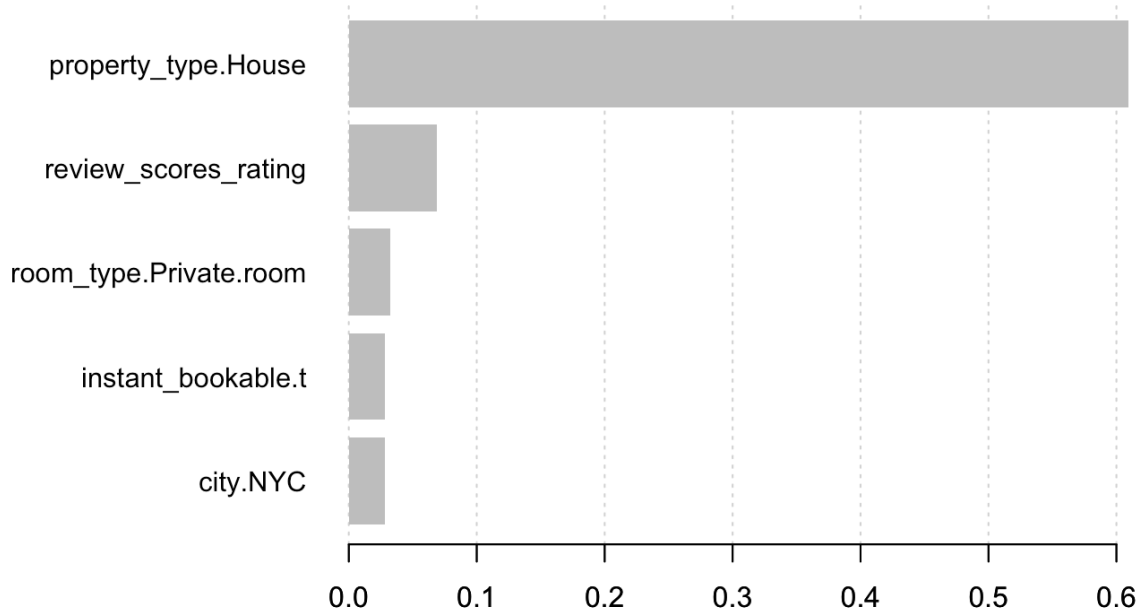
마찬가지로 변수의 개수가 매우 많은 관계로 회귀 계수가 0이 아닌 변수들만 기록하였다.

	Predictors	Estimate
(Intercept)	**(Intercept)**	**-0.015**
property_type	**property_typeCondominium**	**0.007**
	property_typeLoft	**0.013**
room_type	**room_typePrivate room**	**-0.494**
	room_typeShared room	**-0.517**
accommodates	**accommodates**	**0.022**
cancellation_policy	**cancellation_policystrict**	**0.021**
city	**cityChicago**	**-0.117**
	cityDC	**-0.039**
	cityLA	**-0.061**
	citySF	**0.168**
instant_bookable	**instant_bookablet**	**-0.001**
review_scores_rating	**review_scores_rating**	**0.005**
bedrooms	**bedrooms**	**0.072**
amenities	**Kitchen.TRUE**	**0.017**
	Family.kid.friendly.TRUE	**0.001**
	Hair.dryer.TRUE	**0.001**
	Dryer.TRUE	**0.013**
	TV.TRUE	**0.018**
	Cable.TV.TRUE	**0.049**
	Buzzer.wireless.intercom.TRUE	**0.050**
	Indoor.fireplace.TRUE	**0.029**
	Free.parking.on.premises.TRUE	**-0.013**
	Gym.TRUE	**0.027**
	Doorman.TRUE	**0.072**
	Lock.on.bedroom.door.TRUE	**-0.025**
	Pets.live.on.this.property.TRUE	**-0.009**
	Private.entrance.TRUE	**-0.003**
	Elevator.TRUE	**0.035**

	Dishwasher.TRUE	**0.013**

'amenities'를 제외한 12개의 설명변수들에 대한 계수들의 경향성(부호나 상대적 크기)은 앞에서 확인하였던 Logistic Regression과 상당히 비슷하다. 다만 'amenities' 부분에서는 31개의 변수가 16개의 변수로 줄어, Variable Selection이 이루어진 것으로 보인다.

4-2-2-3. XGBoost



'amenities'의 변수들을 추가하였을 때도 순위에 변동이 없었다. 'amenities'의 항목들은 Binary로 Encoding된 가격을 예측하는데 큰 영향을 못 미치는 것이라 생각된다.

이외에도 Random Forest도 시도해보았으나, XGBoost가 1분 내외로 결과물인 나오는 것에 반하여 Random Forest는 4~5시간이 지나도록 결과가 출력되지 않아 비효율적이고, 이미 XGBoost가 Random Forest와 동일한 Ensemble Method이기 때문에 생략하였다.

5. 토의

5-1. 'log_price'와 유의한 관계가 있는 변수 탐색

4장에서 Continuous Scale, Binary Scale 'log_price' 각각에 대하여 Setting 1(Without 'amenities' : 13 Variables), Setting 2(With 'amenities : 150 Variables)와 같은 조건에서 Multiple Linear Regression/Logistic Regression 및 LASSO를 이용하여 모형을 적합하였다. 이제 이 결과들을 종합하여, 어떤 조건에서든 log_price와 항상 유의한 관계를 가지는 변수들을 탐색할 수 있다. 우선 동일한 'log_price'의 Scale, Setting에서, 모든 모델에서 항상 유의하게 나온 변수들을 정리한 결과는 아래의 1) ~ 4)와 같다.

1) Continuous Scale 'log_price', Setting 1(Without 'amenities' : 13 Variables)

	Predictors	MLR	LASSO
property_type	property_typeCamper/RV	-0.279	-0.316
	property_typeCondominium	0.076	0.082
	property_typeDorm	-0.325	-0.339
	property_typeGuesthouse	-0.113	-0.109
	property_typeHostel	-0.592	-0.55
	property_typeHouse	-0.064	-0.061
	property_typeIn-law	-0.205	-0.188
	property_typeIsland	0.771	0.69
	property_typeTent	-0.313	-0.315
	property_typeTimeshare	0.319	0.282
	property_typeTownhouse	-0.047	-0.049
	property_typeTreehouse	0.411	0.372
	property_typeVilla	-0.112	-0.071
room_type	room_typePrivate room	-0.614	-0.611
	room_typeShared room	-1.11	-1.103
accommodates	accommodates	0.065	0.066
bathrooms	bathrooms	0.141	0.143
cancellation_polkcy	cancellation_policymoderate	0.026	0.022
	cancellation_policystrict	0.078	0.075
city	cityChicago	-0.357	-0.346
	cityDC	-0.159	-0.146
	cityLA	-0.173	-0.162
	cityNYC	-0.029	-0.018
	citySF	0.232	0.238

instant_bookable	**instant_bookablet**	**-0.045**	**-0.046**
review_scores_rating	**review_scores_rating**	**0.014**	**0.014**
bedrooms	**bedrooms**	**0.157**	**0.154**
beds	**beds**	**-0.033**	**-0.032**

2) Continuous Scale 'log_price', Setting 2(With 'amenities' : 150 Variables)

	Predictors	MLR	LASSO
property_type	**property_typeCondominium**	**0.036**	**0.004**
	property_typeDorm	**-0.301**	**-0.205**
	property_typeHostel	**-0.607**	**-0.333**
	property_typeHouse	**-0.016**	**-0.006**
	property_typeLoft	**0.115**	**0.076**
room_type	**room_typePrivate room**	**-0.585**	**-0.572**
	room_typeShared room	**-1.09**	**-1.061**
accommodates	**accommodates**	**0.059**	**0.049**
bathrooms	**bathrooms**	**0.11**	**0.094**
cancellation_policy	**cancellation_policystrict**	**0.06**	**0.033**
city	**cityChicago**	**-0.356**	**-0.243**
	cityDC	**-0.161**	**-0.052**
	cityLA	**-0.178**	**-0.087**
	citySF	**0.217**	**0.254**
instant_bookable	**instant_bookablet**	**-0.028**	**-0.018**
review_scores_rating	**review_scores_rating**	**0.013**	**0.011**
bedrooms	**bedrooms**	**0.158**	**0.148**
amenities	**Hair.dryer.TRUE**	**0.048**	**0.002**
	Shampoo.TRUE	**0.035**	**0.005**
	TV.TRUE	**0.0478**	**0.035**
	Cable.TV.TRUE	**0.076**	**0.068**
	Buzzer.wireless.intercom.TRUE	**0.071**	**0.08**
	Indoor.fireplace.TRUE	**0.103**	**0.09**
	Pool.TRUE	**0.043**	**0.009**

	Free.parking.on.premises.TRUE	-0.056	-0.039
	Gym.TRUE	0.0386	0.029
	Doorman.TRUE	0.134	0.131
	Lock.on.bedroom.door.TRUE	-0.043	-0.041
	Private.entrance.TRUE	-0.053	-0.023
	Suitable.for.events.TRUE	0.062	0.029
	Elevator.TRUE	0.08	0.096

3) Binary Scale 'log_price', Setting 1(Without 'amenities' : 13 Variables)

	Predictors	Logistic	LASSO
property_type	property_typeCondominium	0.526	0.029
	property_typeGuesthouse	-0.426	-0.016
	property_typeHouse	-0.322	-0.01
	property_typeLoft	0.681	0.029
room_type	room_typePrivate room	-2.811	-0.515
	room_typeShared room	-3.904	-0.529
accommodates	accommodates	0.323	0.024
cancellation_policy	cancellation_policystrict	0.408	0.03
city	cityChicago	-1.844	-0.126
	cityDC	-1.138	-0.051
	cityLA	-1.086	-0.078
	citySF	1.169	0.159
instant_bookable	instant_bookablet	-0.229	-0.007
review_scores_rating	review_scores_rating	0.068	0.006
bedrooms	bedrooms	0.574	0.073

4) Binary Scale 'log_price', Setting 2(With 'amenities' : 150 Variables)

	Predictors	Logistic	LASSO
property_type	property_typeCondominium	0.373	0.007
	property_typeLoft	0.605	0.013

room_type	room_typePrivate room	-2.777	-0.494
	room_typeShared room	-3.943	-0.517
accommodates	accommodates	0.301	0.022
cancellation_policy	cancellation_policystrict	0.308	0.021
city	cityChicago	-1.979	-0.117
	cityDC	-1.239	-0.039
	cityLA	-1.15	-0.061
	citySF	1.148	0.168
instant_bookable	instant_bookablet	-0.166	-0.003
review_scores_rating	review_scores_rating	0.066	0.035
bedrooms	bedrooms	0.579	0.013
amenities	Kitchen.TRUE	0.174	0.017
	Family.kid.friendly.TRUE	0.115	0.001
	Hair.dryer.TRUE	0.228	0.001
	TV.TRUE	0.193	0.018
	Cable.TV.TRUE	0.408	0.049
	Buzzer.wireless.intercom.TRUE	0.376	0.05
	Indoor.fireplace.TRUE	0.408	0.029
	Free.parking.on.premises.TRUE	-0.268	-0.013
	Gym.TRUE	0.192	0.027
	Lock.on.bedroom.door.TRUE	-0.31	-0.025
	Pets.live.on.this.property.TRUE	-0.215	-0.009
	Private.entrance.TRUE	-0.269	-0.003
	Elevator.TRUE	0.228	0.035
	Dishwasher.TRUE	0.701	0.013

위의 결과로부터 Continuous Scale 'log_price'에 대해서든, Binary Scale 'log_price'에 대해서든 항상 유의하게 나오는 변수들을 파악할 수 있다. 즉, 'amenities'를 제외한 변수들의 경우, 1) ~ 4)의 모든 Setting에 대하여 항상 유의하게 나오는 변수들을 파악할 수 있고, 'amenities' 관련 변수들의 경우 Setting 2에 해당하는 2)와 4)에서 항상 유의하게 나오는 변수들을 파악할 수 있다. 실제로 이들 변수들은 1) ~ 4) 모두에서 항상 같은 부호의 계수들을 가지므로 log_price와의 경향성 동일한 방향으로 설명한다고 볼 수 있다. 위의 결과로부터 찾아낸 'log_price'와 유의미한 관계가 있는 변수들의 리스트와 그 부호를 아래와 같이 찾을 수 있다.

	Predictors	Sign
property_type	property_typeCondominium	+
room_type	room_typePrivate room	-
	room_typeShared room	-
accommodates	accommodates	+
bathrooms	bathrooms	+
cancellation_policy	cancellation_policystrict	+
city	cityChicago	-
	cityDC	-
	cityLA	-
	citySF	+
instant_bookable	instant_bookablet	-
review_scores_rating	review_scores_rating	+
bedrooms	bedrooms	+

즉, 'property_type'이 Apartment인 경우보다 Condominium일수록, 'accomodates'와 'bathrooms', 'bedrooms'의 개수가 더 많을수록 더 'log_price'가 높았다. 이러한 상식적인 결과 이외에도 예약 취소에 대한 규칙이 엄격하거나, 즉석에서 예약을 하지 못하는 '규정을 엄수하는' 방일수록 'log_price'가 높다는 것을 알 수 있다. 한편으로 Boston, Chicago, DC, LA, SF 지역 간의 차이도 'log_price'에 유의미하게 영향을 미친다는 사실을 알 수 있다. 전반적으로 Boston은 SF보다는 'log_price'가 낮은 편이지만, Chicago, DC, LA보다는 가격이 높은 편이다. 또한 숙박 이용자들의 매긴 평점이 높은 숙박 업소들이 비교적 가격이 비싼 곳이었다는 곳도 확인할 수 있다.

	Predictors	Sign
amenities	TV.TRUE	+
	Cable.TV.TRUE	+
	Buzzer.wireless.intercom.TRUE	+
	Indoor.fireplace.TRUE	+
	Free.parking.on.premises.TRUE	-
	Gym.TRUE	+
	Lock.on.bedroom.door.TRUE	-
	Private.entrance.TRUE	-
	Elevator.TRUE	+

137개의 'amenities' 관련 변수들 중에서 위와 같은 9개의 변수들만이 최종적으로 유의한 변수들로 파악되었다. TV & Cable TV와 같은 전자기기나 Buzzer wireless intercom과 같은 통신 장비들의 유무 여부가 'log_price'에 유의미한 영향을 끼친다는 상식적인 결과를 확인할 수 있지만, 한편으로는 무료 주차 공간이 없거나 침실 잠금 장치가 없는 경우 오히려 'log_price'가 높다는 흥미로운 결과들 역시 확인할 수 있다.

5-2. Setting 1과 Setting 2에 대한 비교

Continuous Scale의 'log_price'를 예측할 때는 3가지 모델 전부(Multiple Linear Regression, LASSO, XGBoost) 'amenities' 변수를 넣었을 때 결과가 약간 더 개선되었다(4-1). 실제로도 Multiple Linear Regression과 LASSO의 실행 결과를 보았을 때, 'amenities'에서 유의한 변수가 다량 존재한다는 것을 확인할 수 있었다. 결론적으로 'amenities'의 각 변수들도 'log_price'와 유의미한 관계가 있는 것으로 확인된다. 다만 'amenities' 각 변수들 간 상관관계와, 'amenities'와 다른 변수들간 상관관계가 존재하기 때문에 'amenities' 변수들을 추가하여 'log_price'를 예측할 때 예상보다 적은 성능 개선을 보여주는 것으로 판단된다.

반면, Binary Scale의 'log_price'를 예측할 때는 'amenities'의 변수를 추가하였을 때 XGBoost를 제외한 두 모델은 오히려 성능이 하락하였다(4-2). Logistic Regression과 (Logistic) LASSO 두 모델에서는 너무 많은 변수들이 Overfitting Issue와 Noise를 증가시켰기 때문으로 판단된다. XGBoost는 Regularization 항이 있고, Decision Tree의 Pruning 또한 모델에 포함되어 있기 때문에 Overfitting과 Noise에 조금 더 Robust하여 결과가 개선된 것이라 생각된다.

Continuous Scale과 Binary Scale 모두 137개의 'amenities' 관련 변수들을 추가했을 때 성능의 증가가 없거나 미미하였다. 성능의 작은 차이가 중요한 상황이 아닐 경우에는, 12개의 설명변수만 가지고도 적은 계산량과 시간으로 효율적인 예측을 할 수 있다는 것을 알 수 있다.

5-3. 각 Model별 차이에 대한 비교

Continuous Scale에서는 근소한 차이로 XGBoost가 제일 좋은 결과를 도출해냈다. 특히 Continuous Scale에서는 RMSE 뿐만이 아니라 MAE와 MAPE 모두 XGBoost가 우수하였다. 다만, XGBoost는 다른 Regression 모델들에 비해 결과를 설명하기 어렵다는 단점이 있다. 그렇기 때문에 성능의 작은 차이가 중요한 상황이 아닐 경우, 단순히 Multiple Linear Regression 모델을 이용하여 설명하는 것도 가능할 것이다.

Continuous Scale에서의 예측을 조금 더 자세히 살펴보면, LASSO가 오히려 Multiple Linear Regression보다 더 좋지 않은 결과를 보여준다. LASSO Regression은 Multiple Linear Regression의 각 계수에 L1-Regularization을 하는 모델로, 대부분 변수들의 회귀 계수들을 0으로 만들려는 성향이 있다. 그래서 Variable Selection 효과까지 있는 모델인데, Continuous Scale의 경우는 각 변수들이 Overfitting 이슈가 크지 않고 예측에 어느 정도 기여하고 있는 것으로 생각된다.

이제 Binary Scale에서의 예측을 살펴보면, Setting 1과 같이 설명변수의 개수가 12개일 때는 Logistic Regression의 정확도가 다른 모델들보다 약간 높다. 'amenities' 변수를 추가하지 않아 변

수의 개수가 적을 때는 단순한 모델들이 더 높은 정확도를 가져올 수도 있기 때문이라 생각된다.

그리고 Continuous Scale의 경우와 마찬가지로 LASSO보다는 Logistic Regression의 정확도가 약간 더 높다. Continuous Scale과 마찬가지로 각 변수들이 예측에 조금씩 기여를 하고 있는 것으로 판단된다.

'amenities' 변수를 추가하였을 때는 XGBoost의 성능이 가장 좋았다. 특기할 점은, 'amenities' 변수가 있는 경우와 없는 경우 모두 Specificity는 LASSO가 가장 좋았다는 점이다. 이는 LASSO의 Variable Selection 효과로 인한 것으로 판단된다.

정리하자면, 전반적인 정확도는 XGBoost 모델이 가장 높았다. 하지만 정확도의 개선보다는 설명하기 쉬운 모델을 원할 때는 단순 회귀 모델들도 'log_price' 예측에 있어서 좋은 성능을 보여주기 때문에, 상황에 따라 적절한 모델을 선택해야 할 것이다.