4 Airbnb quality screening

Airbnb is a platform that allows users (*hosts*) to rent out their properties to other users (*guests*). Hosts can place multiple *listings* on Airbnb. Guests book a listing for some duration of time through Airbnb; the guest pays a price for the booking, and Airbnb takes a small percentage.

A problem that Airbnb faces is maintaining high-quality listings. Guests who stay at high-quality listings are happier with the platform and continue to use it in the future, bringing in more revenue to Airbnb. Conversely, guests who stay at low-quality listings can have very poor experiences, and may abandon the platform.

In this problem, we will consider how Airbnb can use its data to identify or flag low-quality listings. The file `listings-small.csv` contains basic information about a set of listings in Los Angeles. The dependent variable is `low_quality` (is the listing of unacceptably low quality to Airbnb or not, determined by detailed inspection of the listing and interview of the host and previous guests). The table below summarizes the variables available to Airbnb:

Variable	Description
`accommodates`	Max. number of guests that can be accommodated by the booking
`bathrooms`	Number of bathrooms
`bed_type`	Type of bed in the listing
`bedrooms`	Number of bedrooms
`beds`	Number of beds
`calculated_host_listings_count`	How many listings does the listing's host have?
`cancellation_policy`	The policy for canceling bookings (flexible, moderate, or strict)
`extra_people`	The charge per extra person (beyond the number accommodated)
`guests_included`	Number of guests included in the booking (by default)
`host_identity_verified`	Is the identity of the host verified?
`host_is_superhost`	Is the host a "super-host"?
`host_response_rate`	What fraction of inquiries does the host respond to?
`host_response_time`	How quickly does the host respond to inquiries?
`id`	The ID of the listing
`instant_bookable`	Can the listing be booked instantly?
`low_quality`	Is the listing a low quality listing?
`maximum_nights`	The maximum number of nights for any stay
`minimum_nights`	The minimum number of nights required per stay
`number_of_reviews`	Number of reviews available for this listing
`price`	Price per night for the listing
`property_type`	Type of property (Apartment, Condominium, House, Loft)
`require_guest_phone_verification`	Does the host require the guest to verify the booking by phone?
`require_guest_profile_picture`	Does the host require the guest to provide a photo?
`review_comments`	Aggregated reviews for the listing (reviews are separated by "||")
`reviews_per_month`	The average number of reviews the property receives per month
`room_type`	Type of room (Entire home/apt, Private room, Shared room)
`summary`	Host's summary of the listing

Part 1: An initial model

Set your seed to 88 and split the data 60-40 into a training set and a testing set; ensure that the relative proportion of listings with `low_quality` = 1 and `low_quality` = 0 is preserved in the two sets. Answer the following questions:

a) Using the variables `bed_type`, `room_type`, `property_type` and `host_response_time`, construct a logistic regression model. What is the test set accuracy of your model? (Use a threshold of 0.5.)

b) Suppose that Airbnb would like to ensure that for any low quality listing, there should be an approximately 80% chance that the model correctly flags it as a low quality listing. If we used our model to flag low quality listings in accordance with this requirement, what is the probability that the model would incorrectly flag a "good" listing? Give the lowest such value that would meet the 80% requirement; an approximate answer is OK. Explain your answer.

c) Airbnb currently uses a different method for flagging low quality listings. This method achieves a sensitivity of 30% and specificity of 80%. Does our model improve on this existing model? Explain your answer.

Part 2: A better model

The model we developed in Part 1 has relatively low predictive power. Airbnb would like to find a better model.

For this question, analyze the data to obtain a better model. Use the same training set to build your model, and the test set to evaluate them. You may use any of the data in the data set to build your model. If you use a function that requires the random number generator, set your seed to 88 immediately beforehand.

Present your analysis. Your answer should address the following questions:

a) What type of model is your final model?

b) Which variables/features did you use to build your model?

c) What does your model suggest about which variables/features are predictive of a listing's quality?

d) What process did you use to arrive to your final model?

e) Why should Airbnb use your model? (What is the predictive performance of your final model? Does your model have other advantages, besides its predictive performance?)

f) What weaknesses, if any, does your model have?