INFO370 Problem Set 6: Linear Regression (100pt)

Deadline: Thu, Dec 5th 5pm

Instructions

Make a good model!

Your task is to use the AirBnB data to predict the listing price.

1. In this PS you have to submit two files:
   (a) your code (notebook, .rmd, or a separate code file, whatever you use)
   (b) your final report (as html or pdf).
2. Please write clearly! Answer each question in a way that if the code chunks are removed from your document, the result is still readable!
3. Please keep data file in the same folder as your code, and read these w/o any path like `"data.csv"` (or `"./data.csv"`). This makes the checking your code much easier!

1 Data description (15pt)

Your first task is to load, describe, and clean the data. Here you also do all the feature engineering you may want to do. This is about AirBnB data, downloaded from insideairbnb.com.

1. Load the data *airbnb-seattle-listings-train.csv*. Broadly describe the variables you see, their encoding, and discuss if these may be valuable in determining the price. For instance, you may want to thell that *house_ rules* is text, and you may want to check if smoking allowed/not allowed is related to the price.

2. Consider how will you handle missing data. For instance, 95% of the "square feet" observations are missing, 17% of "security deposti" observations are missing. You loose too many observations if you just ignore those.

3. Consider which variables you are going to use below. For all of these, create a summary table that contains relevant summary information. In particular pay attention to the missing values. Note that missings may not just be coded as such, they may also be empty strings and values like "N/A". You may return to this point repeatedly as you develop your model.

2 Model (60pt)

Your next task is to predict the Seattle AirBnB listing prices, the variable *price*. You develop as good a model as you can. But let's stay with OLS. No trees, no kNN, no convolutional neural networks. But you can engineer all the features you want, you can put log on the left hand side, and you can bring in more data (say, neighborhood walk score) from other sources.

Start slow with your model and only include one or two variables you consider most important/easiest to handle. Thereafter add more and more variables. Your task is to get the adjusted R^2 as good as you get.

1. Either split your data into training and validation sets, or just use cross validation below.

2. Develop the models. Report all the variables and how do you clean/encode those. While the exact details are visible in the code, explain the broad choices in text.

3. Report the final number of observations, the estimated coefficient values, adjusted R^2, and RMSE on validataion data (or k-fold CV) for three models:

   (a) a simple one that only contains a few most important variables/best predictors. What do you think are 2-3 best predictors in the data?

   (b) the full model: everything you consider useful.

   (c) something in between.

   Please do not report all the coefficients of the large models, only a small subset (10 or so) of the most important/interesting ones.

4. Interpret the coefficients of the reported models. Again, only interpret the most interesting/important ones, not all of those! Do the coefficient values differ between the models? Can you explain why?

5. Use your models to predict the price. Report RMSE in the table above.

   Note: ensure you fit the models on your training data and compute RMSE on the validation data. Here we care about overfitting. It is less important when you just interpret coefficients, as you do above.

3 Think (15pt)

The final task is to think what does your prediction mean. Write a discussion section where you address the following questions:

1. does your model do a good job in predicting the prices?

2. how will your model be useful to

   (a) AirBnB hosts

   (b) AirBnB customers

3. Did you include any other price-related variables, such as "weekly price" or "security deposit" in your model? Do you think it is a good idea to use these attributes while trying to predict price?

4. Do you think this model can be used by Airbnb itself or the government?

5. Do you see any ethical issues with this work?

4 Additional task (10pt)

This problemset includes an additional task: test your model. The testing dataset will be released morning Dec 5th.

1. load the testing data *arbnb-seattle-listings-test.csv*. This has exactly the same structure and variables as the original dataset.

2. compute RMSE on the testing dataset. This is the ultimate goodness measure of your model. Present it prominently in your report.

3. Do not tinker with the model any more. This was your final test.

   Note: you may still have to fix coding errors if there is something wrong so you cannot compute RMSE on the test data.

5 Extra credit

The 5 best teams in terms of RMSE on testing set will receive 4pt extra credit.