ECON 8610 Homework 1

Ross Bowen

March 2, 2021

The objective of the data exercise is to predict the price of AirBNB listings in New York.

Cleaning Data

Some data was jettisoned (in excel)

Neighborhood with Borough – would take many dummy variables

Last review, reviews per month – does not plausibly affect price

Name – a wordy description

Name of owner

Availability – if anything this could reduce price, as may indicate lack of availability

Latitude and longitude – save for later. Could infer price from other nearby rentals

The rest of the data was kept

Borough – Manhattan and Brooklyn are more desirable

Room type – entire apartment most desirable, shared room least

Minimum nights

Number of reviews – might indicate frequency of renting, which could be positive

Host Count – number of rentals host controls – indicates dedication

Borough and Room type were turned into dummy variables (5 boroughs and 3 types)

There are 11 predictors against the target, “price”.

Data was shuffled prior to regression – the rooms seemed loosely organized by borough.

**K Fold Linear Regression**

Using 4-fold regression, the model was run four times, each time with ¼ of the data set aside as “test”. After the regression is run on 75% of the data, that model is used to predict the actual set aside data, as if it were not known. Each run produces a set of coefficients and an R squared value.

Here are the coefficients

MnNt NumRev HostCnt Bronx Brooklyn Manhat Queens StIsl EntApt PrivRm ShRoom

0.13 -0.205 0.066 -24.685 -4.911 51.939 -17.528 -4.815 80.563 -29.613 -50.95

0.121 -0.209 0.056 -24.191 -3.601 51.846 -14.455 -9.598 81.52 -29.642 -51.878

0.236 -0.207 0.057 -24.764 0.213 55.985 -13.5 -17.934 82.978 -28.076 -54.902

0.102 -0.205 0.07 -23.619 -3.753 52.325 -15.074 -9.877 81.771 -28.279 -53.492

The coefficients are fairly stable. As expected, rentals in Manhattan command a premium price, with Brooklyn the next most desirable borough. Minimum nights has an immaterial coefficient. Number of reviews and Host Count are also both small.

Renting an entire home or apartment commands a premium of about $80. A shared room is penalized about $50 and a private room is penalized about $29.

The R squared for each of the 4 folds are shown here:

.0836

.0797

.0958

.0767

This is an average of .0840. If we could find the actual ratings associated with each property, I believe our model would become more accurate.

**Lasso Regression**

Lasso regression is a penalized regression. Rather than just merely calculating least squared deviations, there is an additional penalty which causes some variables to report a “0” coefficient.

4 fold regression was run using lasso. Coefficients are reported below.

MnNt NmRev HostCnt Bronx Brooklyn Manhat Queens StIsl EntApt PrivRm ShRoom

0.135 -0.168 0.011 -7.591 0.000 57.309 -4.226 0.000 109.078 0.000 -15.238

0.168 -0.171 0.022 -11.159 0.000 54.546 -6.509 0.000 106.427 0.000 -14.129

0.000 -0.172 0.025 -9.989 0.000 54.331 -7.723 0.000 110.702 0.000 -11.16

0.075 -0.166 0.028 -8.936 0.000 53.174 -10.709 0.000 107.908 0.000 -11.415

Brooklyn, Staten Island, and Private Room are no longer included in the final regression.

The R squared for each of the 4 folds are shown here:

.0833

.0871

.0842

.0768

This is an average of .0829. This R squared is slightly lower than for linear regression.

**Conclusion**

I recommend adopting the lasso model. Although its R squared is slightly lower, there is no guarantee that will continue when new random numbers are used.

The advantage to Lasso is that there are fewer variables, and this will give us some protection from overfitting.