STAT 422/722 Spring 2017 PROJECT

Professor Adam Kapelner

Due February 23, 5PM at JMHH 4th floor (in the dropoff box)

(this document last updated Sunday 12th February, 2017 at 10:43am)

1 Introduction

In short, you will be predicting apartment selling prices in Queens, NY. You will be responsible for:

- gathering historical data including
  - deciding which features (predictors) will be of use to you,
  - cleaning up data errors (if they exist),
- deciding which model and which model-fitting technique to use
- handling missing data (if they exist)
- making predictions on apartments currently listed for sale

We will be using the *raw data representation* found at MLSI. The limitation on the data population for what *you will be asked to predict* will be "Queens, NY" as location and home types "Condo / homeowner assoc." and "Co-op" up to a maximum sale price of $1M.

You will be responsible for both (a) writing a report about your data science endeavors and (b) predictions for future data where you will be in competition with one another.

1.1 Motivation

I picked this project because I know you can all do better than `zillow.com` who make their own secret-sauce predictions that they whimsically call "zestimates". However, in Queens, zestimates are quite lame (e.g. this one). I imagine the collective brainpower of all of you plus the elementary concepts and tools from this class can produce better estimates.[1]

[1] At the very least, I imagine you can score a pretty good job interview at Zillow if your predictive performance is any good.

2 The Writeup

2.1 Gathering Data

Your data will come from the following zip codes in mainland Queens.[2]

Northeast Queens	11361	11362,	11363	11364					
North Queens	11354	11355	11356	11357	11358	11359	11360		
Central Queens	11365	11366	11367						
Jamaica	11412	11423	11432	11433	11434	11435	11436		
Northwest Queens	11101	11102	11103	11104	11105	11106			
West Central Queens	11374	11375	11379	11385					
Southeast Queens	11004	11005	11411	11413	11422	11426	11427	11428	11429
Southwest Queens	11414	11415	11416	11417	11418	11419	11420	11421	
West Queens	11368	11369	11370	11372	11373	11377	11378		

You can then enter all these zipcodes into an MLSI search, plus the Co-op and Condo and $\leq$ \$1M restriction, or you can use my search in my account. Login to MLSI by using the login `kapelner@wharton.upenn.edu` and password `stat422` and then go to this link to load the saved search. As of the time of this writing, there are $\approx$1,200 sold properties. And under "Status", if you uncheck "Sold" and check "Active", you will see $\approx$1,000 currently on the market. Of these, you will predict some of them (we will get to this part later).

As you can see, I'm not providing you with a CSV, JMP or RData file — this is part of your job (and likely the most important part). This is too much work for one person to do. I suggest a number of things:

- team up early on
- create shared google sheets where the information gets iteratively populated
- use `MTurk.com` ... very easy way to crowdsource mini-jobs such as extracting data

In your writeup, you will write about your observations by

- indicating how many you have,
- describing how you sampled them (and if this is not a simple random sample explain in detail why not) and
- explaining the degree to which you feel this sample is representative of the population.

If I see no work done on this front by week two of the course, then I will divvy up responsibilities among the class.

Then, you will list the predictors you used and provide

[2]For those of you who know Queens, we are leaving out the Rockaways, a peninsula near JFK airport that is geographically distinct from the rest of the neighborhoods.

- how many you have,
- their names,
- their data types,
- a description of the information captured,
- a report on missingness and the missingness mechanism and
- a description about how the measurements were taken.

There are no collaboration limits on this part of the project. But you will be responsible for submitting an electronic copy of your data frame to canvas at the time that the project is due. (More details on upload specifics coming soon).

Building a dataset from scratch is a big job, but highly educational. You will see just how valuable creativity in this domain is and you will never look at data the same way again.

2.2 Building a Model

You should make use of any of the tools we covered in this class. Please, no methods or algorithms from outside the class.

In your writeup, you must explain the modeling choice you made and describe how you iteratively came to the model you will use for "production" (that's lingo for the one you use to make your "real-world predictions"). That means you will likely report model fit metrics such as likelihoods, AIC's, etc for a few different models. For the final model, you must report your in-sample and out-of-sample:

- R^2
- MSE
- RMSE
- MAE

There are a few other things that need to be reported on:

- I will ask you to rank your most important predictors and explain how sensitive your model performance is if your predictor information were to vanish.
- For the top three variables, how does your model's conditional mean change as the variable changes?
- Which variables do you believe have an effect on sale price that is truly causal and why? Would you be able to prove it?
- You will then be asked to comment on areas of covariate space that are in danger of extrapolation in your model.

- I will ask you to comment on if there is any overfitting in your data.
- Then, you will need to explain clearly how you handle missing data. If you are truly reporting out-of-sample performance metrics, you will have no problem when you predict real, future data that contains missing data.

You will do this part of the project yourself with no collaboration.

3 The Prediction Competition

You will then have one day to predict the selling prices of a set a of 947 apartments currently on the market found in the prediction CSV file (visit the link and right click and "save as...").

This file was MTurked and thus contains a lot of the MTurk metadata (which you will ignore). It has also been cleaned by me in a few hours so the cleaning is not perfect. This is a example of what you get in the real world on the higher end of quality.

I've included some features in this prediction set. You are under no obligation to use them whatsoever. You can capture your own features by using the link provided in the `URL` column.

You will upload a file named `<Your Penn ID>.csv` to canvas by Friday, Feb 24 5PM. (More details on upload specifics coming soon). The number of lines in your CSV will be the same number of lines as the prediction CSV file less one (for the header line). Each line will consist of a single prediction value (in dollars without the dollar symbol or commas). For instance:

```
145645
862684
452890
.
.
.
235977
```

Figure 1: An example file `<Your Penn ID>.csv`

I will then wait until the grading deadline May 12 or until 100 apartments are sold. I estimate 2-3 apartments sell per day of the 947 on this sheet thus we will have a nice set to evaluate *your* future predictions against.

You will be graded on your out-of-sample R^2 value. How to allocate the points I have not determined yet.

If you finish in the top five among the two sections of Stat 422/722, I (or the TA) will try to reproduce your results via your submitted data frame and your crystal clear writeup explaining your modeling technique. Reproducible work / research is a becoming more and more demanded these days. I will also want to ensure a fair playing field.