STAT 422/722 Spring 2016 PROJECT

Professor Adam Kapelner

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

(this document last updated Sunday 15th January, 2017 at 1:51pm)

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 Zillow. The limitation on the data population for what *you will be asked to predict* will be "Queens, NY" as location and home types "Apartments" and/or "Condos / co-ops" up to a maximum listing price of $1M. Thus, historical data will be a subset of the this search although you may want to alter the query to include higher prices.

You will be responsible for both (a) a writeup and (b) predictions for future data.

2 The Writeup

2.1 Gathering Data

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
- 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, then I will divvy up responsibilities among the class.
Then, you will list the predictors you used and provide

- how many you have
- their names
- their data types
- a description of the information captured
- a report on missingness and the missingness mechanism
- 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 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?
- 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 500 apartments currently on the market (found in the listing URLs Google sheet).

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 in the listing URLs sheet. 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. Approximately 8 apartments are sold per day. The listing URL sheet has 500 apartments representing about 1/6 of the market. Thus we can expect 1.3 apartments to sell per day and thus we will have about 18 data points to evaluate our 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.