STAT 422/722 Spring 2017 PROJECT

Professor Adam Kapelner

Writeup due Friday, March 3, Noon, JMHH 4th floor (in the dropoff box)

(this document last updated Tuesday 21st February, 2017 at 11:44pm)

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 Project Scope

2.1 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

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.

Note: 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).

2.2 Building a Predictive Model for the Future

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

There are no collaboration limits on procuring data and setting it up for a model and doing the actual modeling.

[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.

3 The Formal Writeup

The formal writeup should look like below. Each section should address concepts given below. You will do your own, individual writeup. No copying from others. No paraphrasing from others.

[TITLE]

A project for Stat 422/722 at Wharton Business School
March 3, 2017

By [You]
In collaboration with:
[person 1]
[person 2]
⋮
[person ℓ]

Abstract

A one paragraph summary of the entire writeup that is written to "lure" the reader in.

pagebreak

1. Introduction

Write about the problem here and some context and background. No need to cite papers. Talk about what a predictive model is and what that means here. What is the unit of observation? What is the response? Write about the basics of how you modeled it. You can mention your performance results, but do not go into detail about them (leave it for the discussion section). Use as much vocabulary as you can from the class notes in describing the problem. You do not need to talk about the prediction competition whatsoever in this writeup.

2. The Data

Give a one paragraph introduction to what type of data was used in this project, basically where it came from and the size of the historical data frame.

2.1. Sampling

How did you sample the observations — is it a simple random sample? Was there thought given to its design? How representative do you think it is of the population of interest? Are there any dangers of extrapolation?

2.2. Featurization

How many and what measurements did you take on the observations? Make sure to list them and give a brief explanation as to what they are; describe what these measurements capture about the observation. Give a basic summary of each feature — average, standard deviation, range for those that are continuous data type and percentages of the categories for those that are nominal data type.

2.3. Missingness

Summarize the missingness across the features in Section 2.2. How did you handle missingness in your data? Give your thoughts on pattern mixture models and mising data mechanisms. Talk about how you imputed. Did you include any missingness dummy variables in your expanded feature set? Note: you do not need to explain how you handled missingness in your prediction set.

3. Modeling

You are creating a model to ship to the world to be used for predicting real, new observations. What is your choice of predictive model? Why did you choose this model? Is it parametric / non-parametric? What did you gain by choosing this model? Lose? Note: if you also used derived productors (i.e. interactions, polynomials, etc) describe them (without listing them) and explain why they you chose to include these. Was this an iterative process in some way? Do you think you underfit? Do you think you overfit? How were you able to know? Can you rank your most important predictors in terms of importance (not statistical significance)? Which variables do you believe have an effect on sale price that is truly causal and why and would you be able to prove it?

4. Results

Report your in-sample goodness-of-fit metrics: R^2, $RMSE$ (no need for MAE unless you want to report it) and interpret them. Report your AICc if you can. Report your estimate of generalization error as the same goodness-of-fit metrics: R^2, $RMSE$ and interpret these as well. How do you know this is a valid estimate of how the model will by-and-large perform on future predictions?

5. Discussion

Discuss the project once again. Comment on things that you did informally (assume the reader has been through Sections 2-4). Talk about where you feel you fell short and how you can plug those holes. Talk about future extensions. Do you believe your model is production ready? Can you beat Zillow?

4 The Prediction Competition

This section is to be completed exclusively by you.

You will also predict the selling prices of a set a of 944 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 may want to 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 (and this is on the higher end of real-world 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 Sunday, February 26 at 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 and to the nearest dollar). 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 944 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.