Math 342W Final Project 2021

Professor Adam Kapelner

Writeup due Tuesday, May 25, 11:59PM by email

(this document last updated Thursday 27th May, 2021 at 4:33pm)

You will be writing a report about a prediction model for apartment selling prices in Queens, NY using the dataset found on github, `housing_data_2016_2017.csv` where the outcome to be predicted is the column named `sale_price`. This dataset is 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 sold between February, 2016 and February, 2017 and limited to the zip codes found in Table . The dataset was harvested with Amazon's MTurk and it is a raw download from their system.

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		

Table 1: The zip codes for the houses in the dataset. They call come from mainland Queens. We are leaving out the Rockaways, a peninsula near JFK airport that is geographically distinct from the rest of the neighborhoods.

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 for apartments are quite lame but indeed have improved over the years. Their reported results for the NYC area as of 4/24/20 are below:

	Predictions within x% of sale price		
Median Error	5%	10%	20%
2.3%	80.2%	94.1%	98.4%

Unfortunately they do not report for Queens (nor for apartments only) so this will have to be our

benchmark. I imagine the collective brainpower of all of you plus the elementary concepts and tools from this class can produce more accurate predictions of apartment prices.[1]

Deliverables

You will commit all files used for your project to your github repository under the subfolder "final_project". You will also write a formal report to be emailed to me by the due date as a PDF. The formal report should look like below. Each section should address concepts given below. You can work together but please list your collaborators and you must do your own, individual writeup. No copying from others. No paraphrasing from others.

[TITLE]

Final project for Math 342W Data Science at Queens College
[date due]

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

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 and your previous writing assignment in describing the problem. Cite sources in APA style, i.e Johnson et al. (1999) within text and (Johnson et al., 1999) within a parenthetical.

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. How representative do you think

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

it is of the population of interest (you define the populatio of interest)? If you supplemented the dataset from other sources, write about it here too. Are there outliers? Are there any dangers of extrapolation?

2.2. Featurization

How many and what measurements did you take on the observations? Which were provided to you in the raw data and which did you featurize yourself? 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. Errors and Missingness

Did you find obvious errors (not missingness) in the dataset? How did you handle these errors? Summarize the missingness across the features in Section 2.2. How did you handle missingness in your data? 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. But you also would like to explore a little bit.

3.1 Regression Tree Modeling

Fit one regression tree. Visualize the top layers. Comment on the top 10 features that are seemingly most important for predicting sale price. Include the visualization as a figure. If you cannot get `YARF` to work, use the canonical CRAN package `rpart`.

3.2 Linear Modeling

Fit a vanilla OLS linear model. Comment on its in-sample error statistics and interpret them. For the most important features found in the regression tree, interpret the coefficients in this model. Will a linear model be good for prediction? Include the OLS output as a table.

3.3 Random Forest Modeling

Why should this be your choice of prediction model? Explain the theory as best as you could. Is it parametric / non-parametric? What did you gain by choosing this model? Lose? Was modeling an iterative process in some way? Do you think you underfit? Do you think you overfit? How were you able to know? Which variables do you believe have an effect

on sale price that is truly causal, why and would you be able to prove it? Use the package `mlr` (or `mlr3`) to find the best tuning parameters for the RF model. Use these parameters to build the production model. If you cannot get `YARF` to work, use the canonical CRAN package `randomForest`.

4. Performance Results for your Random Forest Model

Report your oob goodness-of-fit metrics: R^2, $RMSE$ (no need for MAE unless you want to report it) and interpret them. 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? In addition to using oob validation, do a hold-out test set validation as well and report the error metrics. Summarize these results in a figure or table. Did your random forest model beat the linear model? Why / why not?

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?

Acknowledgments

If relevant, list people or organizations (not me, not your collaborators nor the TA) who have helped with this project in some way and state how they helped. Give credit where credit is due.

References

If you cited any articles, books, blogs, etc. Use APA format for biliographic entries.

pagebreak

Code Appendix

Print all your code that was used to do this project here.

You may do more than what is above too. Don't hesitate to include additional figures, tables, illustrations if you believe it helps relate what you have done.