Peter Han Individual Report Edit

Mallory Flynn

March 6, 2020

Overall Suggestions and Impressions

- The report is concise and mostly well organized, with clear breakdown of report sections using subtitles. I would suggest another revision with the intent goal of rearranging ideas into their appropriate subsections (some examples of this are elaborated below).
- You clearly understand the client's goals. This report presents an appropriate first attempt at analyzing the relationship between variables of interest to the client.
- Watch use of tense in each section, and try to stay consistent. Also try to reword sentences so that you don't need to refer to "us" or "we". For example, "We decided to calculate the sum. . . " could become "The sum was calculated. . . "
- Revise again for mistakes in spelling and grammar, typos, as well as proper spacing after a period.
- Code section needs major revision to make it easier to follow. Include code comments, avoid multiple renaming of variables, and block libraries and seed-setting into the start of the code, or at least the start of a section.

See the following sections for a more detailed edit of the sections of the report.

Summary

- Revise again for spelling and grammar correction, paying careful attention to words that should be plural.
- Begin the summary with a concise description and motivation of the problem.
- Summary clearly identifies the client's main goals, but does not even address the linear models that were used for this, or their results. This was a large component of the report and should be summarized here also.

Intro

- The introduction jumps right into the particulars of what the client wants and the exact data analysis used. The intro should instead mainly be motivating the problem and providing an overview and background of the topic (for example, explaining property tax generally and why one would care about being able to predict this). Why is this significant, and why do we care? Then it should ease into your particular (broader) objectives for this report.

- Your intro finishes with an overview of your analysis methods. Instead, a more broad description of what's to come should finish your introduction, which goes over what's to come in the following body paragraphs of the report.
- Also be careful not to switch between which tense you are using. The intro starts in the present tense ("we did this", "we saw this") and ends in the future tense ("we will do this"), but all analysis included has already been finished. When revising the intro, try to rewrite using the present tense consistently throughout the intro.

Statistical Analysis

Data Description

- Again, be careful with remaining consistent in one tense, especially in one sentence. For example "we used past property assessment data, which is provided by our client" uses two different tenses. Stick to past tense for describing the data analysis which was carried out, and present tense to report the findings.
- Maybe consider putting variable names in quotation (for example, "TotalAssessedValue"). To add to this point, you have made it clear which variables you used by listing them above, but it might be made clearer by also providing the column name in this list, since you are going to refer to it by this in the report. For example, "- Assessment Values (”TotalAssessedValue“) for 2016-2019" as the bullet point for assessment.

Methods

- It is also possible to label figures in your Rmd report by adding to the code at the beginning of the chunk, and then referring to specific figures just by using `\ref{figurelabel}` in your usual body sentences. This would help the readability of this section of the report, especially the *Exploratory Data Analysis* section, where you could refer the reader to specific figures since there are quite a few included.
- You mention some results you saw for exploratory data analysis. Since you have also included a *Results* section that is broken down into parts, I would not include results in the methods section.
- Are you evaluating goodness of fit solely on R^2 values? Is there other ways you could do this? Is there a downside or limitation to only looking at R^2 values when comparing models? You might also mention that R^2 is computed on training data.
- For *Prediction power*, I think a little more could be said about how you chose training and test data (is this done by choosing random observations?), if you did several rounds of cross validation, etc. It is relatively easy to understand for the client, and helps either support your methods or highlight limitations.
- For the *Linear Model* section, you mention Vancouver, Richmond, Surrey, and Burnaby were outliers. How was this determined? Maybe expand on this and move to the results section.
- "The reduced model's performance in terms of goodness of fit will decrease with fewer features, however, the performance of prediction power might be improved.": This sentence might be modified a little and better suited to the paragraph above, "Measure of goodness of fit and prediction power", where it would complement with the explanation here.
- What was the motivation for attempting Ridge, Lasso and Elastic Net? These methods are typically motivated by certain features of the data. I would defend the consideration of these frameworks by providing further clarification as to why you might want to consider these methods.

- While doing the above, you could explain Ridge, Lasso and Elastic Net in a little more "client friendly" way (that is, discussing that they use a penalty may be correct but not particularly meaningful to the client unless it is clear what this means and why one would want to try this). Maybe compare to the ordinary linear model, if the client has a solid understanding of that.
- The *Neural Network* paragraph is very clear. Since this is not a part of this report's analysis or findings, you could consider including this information in the conclusion as a future direction of this study.

Results

Exploratory Analysis

- Plots should have a caption, where Figure number is specified (this can be done automatically, making it easy to refer to figure numbers within the body of the report, and the numbers update automatically if new figures are inserted or deleted - see description above)
- "Dollars ($)" might be more informative as an axis title than "amount"
- For Vancouver, Harry mentioned the "school tax" which began just recently, resulting in a different calculation for amount of tax being paid by properties worth over $2 million (this is all tax class code 1 properties in our dataset). Was this incorporated to the appropriate years to calculate total tax income collected in Vancouver for the relevant years?

Correlation Analysis

- Plots a very condensed on the left horizontal axis. Instead of plotting raw totals assessment dollar amounts, consider rescaling. Since different municipalities have drastically different numbers of properties, this may be one reason amounts are so much higher for some areas. Maybe rescale by number of properties and see how that plots instead?
- Also, consider transforming your data, for example by taking logarithms. So much clustering could be spread out by an appropriate transformation, making the correlations clearer. (Could try Box-Cox to see which transformation might be appropriate)
- You visually identified potential outliers - is there another way to determine if they are outliers?

ANOVA Analysis

- I like the tables you included - this makes results clear and easy to compare. However, explain what "MSE" and "PMSE" stand for in the body paragraphs or in a table caption - this may be a common abbreviation to you but it is not likely in the everyday lingo of the client. Also, your client might wonder why there are NA's in the table for some of your models. Maybe a brief mention of this
- Do you expect municipality to be significant? Some plots I have seen from your previous analysis showed tax rates are all very similar, following similar trends for the municipalities you considered. Including that plot might help explain why this came up as insignificant. It might be worth considering that this may not be the case once the model is expanded to greater BC.
- Also, "R_squared" should be R^2 or "R-squared" in both body paragraphs and table column headers.
- MSE and PMSE are just quoted as values. How does the client interpret whether these values are good or bad?

Conclusion

- "The client suggested that we can transform numerical factors such as assessment total into percentage change. This data transformation does not perform as well as we expected, which only yields an R^2 of around 0.2 across all of our fitted models. We believe the method that the client has suggested might leave out some important information about the housing market in each municipality." This is new information, that is maybe better to include briefly in the methods and results sections rather than the conclusions.
- If you are going to also consider a neural net as a future piece of this analysis, include this here as another future direction.

References and Appendix

- I like that you included the link to the GitHub repo.
- Use a page break to put references and appendix each on a new page of their own.
- Consider including references to support ridge regression, elastic net, and LASSO. There are lots of sources you could point the client to that explain these methods clearly, if he were to want to do additional reading.
- Make sure to adjust your YAML to specify that you don't want code to run out of the page margins. I think Gaby suggested a fix for this on Slack.
- If the client would like to follow through your code as you directed him to in the body of the report, my impression is that he will need significantly more guidance in the form of comments throughout, as to what you are doing and why.
- Is this the code for the rate change of mill rate? Be more clear on which model you are fitting.
- Include the code relevant to summarizing and aggregating the data from the client, which produces in the dataframe used for analysis.
- Also for readability, consider putting all library's used and setting seed at the beginning of the code appendix, rather than throughout the code. At the very least, they should be at the beginning of each section of code used for various models.
- Using comments, maybe highlight what code produced figures you included in the report.
- Following through the code as in the appendix, it looks like you are training your lasso model on the whole data set, not a training set. Is this the case? How might this affect robustness to new data?
- Change relevant variables to the appropriate type (for example, factors) before analysis, rather than specifying `factor(variable)` every time it is used.
- Since you are only creating one training set and one test set, this may as well be done once at the beginning of the code, rather than repeatedly throughout with each model. That will eliminate the repetitive naming of objects 'x' and 'y' which are constantly being rewritten from one part to the next.
- Since the reduced model is a nested version of the full model, you could try a stepwise calculation of AIC values to see if this supports the reduced model you chose.
- Missing data and imputation should be discussed be discussed in the *Data Description* section, rather than at the end of the code appendix. Include a commentary on how and why you imputed values the way you did.