

For our second project this semester, you will undertake an individual data analysis on a specified problem of interest, and present your findings in a report to your client. We will have some discussion in class about writing up a data analysis and good practices. You are the statistician in charge, and you need to justify your decisions. You'll have a lot of decisions to make.

The methods that this project aims to practice are methods from Chapter 8 - GLMs and trees. However, you can apply other prediction methods if you like, in addition to these.

1 The Data Analysis Task

1.1 The Data

For our analysis, we will use the King County, Washington (State) house sales data set, which I am re-hosting from Kaggle (see .Qmd template for data read-in). The Kaggle reference is: <https://www.kaggle.com/swathiachath/kc-housesales-data/version/1>. You can use the Kaggle page to read about the data set and see the original analysis question. We are tackling a different question. A data dictionary taken from Kaggle is provided for your use (separate pdf file).

1.2 What's the task?

A real estate developer is interested in understanding the features that are predictive of homes selling for more than half a million dollars in this area of Washington, and has turned to you, a statistical consultant for help. The developer wants a model that can be applied to make predictions in this setting (predict if a home sells for more than half a million dollars or not) and wants to understand how the variables in the model are impacting the prediction.

To practice new techniques from class, in your analysis, you are required to use both an appropriate generalized linear model and decision tree to address the developer's questions of interest, including a model comparison, and present your entire analysis, from EDA to results, in a report.

For purposes of this submission, the audience for your report is the real estate developer, who is not a statistical expert, so you'll probably need to include some explanations that you'd leave out of other assignments. For example, the developer isn't going to know what a GLM is, or what type you are using, or why you'd use it. You should think about where that information should go, and do your best to explain what you need in order to report your findings. Similarly, assume that you found this data to assist the developer and they need basic details about it to understand the variables. They didn't hand it over to you.

2 Project Timeline

Here is the timeline for this assignment:

1. In class, we'll introduce the project on Monday, Oct. 6th, and talk about writing up a data analysis on Monday and Wednesday.
2. You'll need to work on the analysis, do your EDA, fit your models, determine your final model(s) of each type, and do the comparison. None of this is submitted separately.
3. Ideally, you want draft models (and maybe even a write-up) by class on Monday, Oct. 20th, because that day, the plan is to share model ideas/challenges/writing concerns in small groups in class. And then to brainstorm/help each other with said challenges.
4. You have to write-up your process/results. The write-up is due Friday, Oct. 24th by midnight. (Hmk 4 will be assigned on Oct. 15th (or so), but isn't due till November 3rd. So this project and the reading should be your focus from the midterm till this project's due date). There is a report template and we'll be brainstorming what goes in each section in discussions in class.
5. After submissions, I'll review your findings, and we'll do an individual discussion where I will ask you questions about your decisions/choices/final models, etc. Having the discussion will enable me to ask any questions I have on the submission, and you'll get to practice defending your data analysis choices. The oral discussions will take place the week of Oct. 27th (with the most slots on the Thursday again). Sign-ups will go up the week before, and there will be more detailed information about what to expect in the conversation, example questions, etc.

Feel free to set any sub-deadlines for yourself that you find useful.

This project is likely to be the one most familiar in style and scope to you of our three projects. We're practicing new techniques (including best practices in modeling) but also spending some extra time thinking about how the write-up should go, so it should result in a stronger submission than say, a data analysis report from a past class.

3 Things to Think About

There will be some suggestions about items to think about per section in the report template. I encourage you to put all your data analysis in one .Qmd, and then use a second one to construct the final report. That way, you can fit many models, but the final report can describe your model fitting process and highlight key models, not show all of them.

Some other items to think about include:

1. Your report should be well-organized and proofread.

2. Your code should be readable and not run off pages. Employ good commenting practices.
3. All code must be included (class requirement).
4. The entire process must be reproducible.
5. Think carefully about the organization of the report. Is it useful for the reader to have pages of output and THEN a description of what to look for in that output? Be sure to look at your pdf before submitting!
6. Should you assume the reader / audience here knows where values in the R output are? (See above about audience.)
7. If you compute a number, and then write a sentence that says, the MSE is small, will the reader associate that number with being the MSE? What would a better sentence look like?
8. Think about our writing discussions. (Does each of your sections have a purpose? Do you have appropriate signposts, etc.)

4 Additional Information

Some other items are important to note for this project (and our future projects).

4.1 Assessment and Organization

As with the first project, to complete assessment of this project, I will examine both your submission in Gradescope and pull your repository to examine your work there. This means your repository needs to contain your work on the project, in an organized fashion. I should be able to find all items associated with this project easily in the repo. This is easy if you just have a folder for the data analysis project work and keep it organized.

Additionally, everyone will have the following assessment components besides the report components:

1. Participation in class discussions (multiple days) will factor into our class participation grade (not the project grade).
2. Oral discussion - I haven't figured out the exact categories yet, but likely some sort of breakdown about Data/EDA, Models, Justification, Communication Clarity.
3. Engaged in Responsible / Reasonable commit behavior (as described previously).

For the report, I have a draft rubric but still need to make some edits. Still the general categories (with some details) for assessment are:

1. Presentation / Organization - overall document presentation
2. Code - coding style, comments, etc.
3. Intro - thoroughness, quality of explanation, appropriate information included
4. EDA - thoroughness, quality of exploration, justification of decisions
5. Modeling - process and description of process, best practices
6. GLM - appropriate GLM used, evidence of model selection, description of results
7. Tree - appropriate tree used, evidence of model selection, description of results
8. Model Comparison - adequate comparison, final model stated clearly
9. Conclusion - address question of interest, standalone summary of analysis

4.2 Generative AI Guidance

There is no requirement to engage with generative AI for this or any course activity.

If you choose to engage with generative AI for any part of this project, remember to follow the class framework. Generative AI may be used only as a thought partner, and only in places that don't jeopardize your learning/our course learning outcomes.

You should immediately export, save, and commit any chats with generative AI about the project as you have them to keep a complete record to submit (Sharing a link has worked as well).

Adding the information to the integrity page as you go is also strongly recommended, rather than trying to do it at the end. You could keep a running pdf addendum of all the chats for easy reference too. Programs that merge pdfs (like Adobe Acrobat) can be useful for that.

Examples of what generative AI might be useful for in the context of the project and that are in line with the class framework are:

1. Asking AI for help coming up with a timeline that works for the project.
2. Asking AI for help with a code error, making sure you can understand how it fixed it and could re-explain it later.
3. Asking AI to help fix an issue with a plot you are having, as long as you can explain how to fix it yourself later.
4. Asking AI to tighten up writing in a particular paragraph or suggest improvements in a section for clarity. (Remember to review it to be sure the AI didn't add anything it shouldn't have!)

Note how these are uses that you have available human support for, without concerns for academic integrity. Me, your peers, and the SLC (Strategic Learning Center) are resources for learning how to make timelines. Google (or any other search engine), me, and your peers are good resources for code issues or formatting challenges. Outlining and writing skills have support in the Writing Center, or you could ask a peer to review a submission.

Contrast that with the uses below, where academic integrity concerns arise. These show issues where the AI use is jeopardizing class learning outcomes / your learning experience. Obviously, the uses below (and others like them) are not allowed.

Examples of how generative AI might be used in the context of the project but that are not in line with the class framework (and therefore, not allowed):

1. Asking AI to do EDA for you.
2. Asking AI to do the data wrangling for you for the project.
3. Asking AI to fit models for you.
4. Asking AI to draft the report for you.

Neither of these lists are exhaustive. Please refer back to the course syllabus or talk to me if you want to engage with Generative AI tools and have questions about appropriate use.