Psych 752 (Introduction to Applied Machine Learning) Final Project

General Instructions

Your task in this final project is to answer two questions: a prediction question (part 1), and an explanation question (part 2). For each, use the following scenario to guide your work: you are communicating with me, your collaborator, about your projects. Your mode of communication is rendered Quarto documents (HTML). I have a general statistics and machine learning background but know nothing about your specific projects. These files should allow me, as your collaborator, to understand:

- The methodological & analytic decisions you made
- How you came to make these decisions (i.e., your decision-making process)
- The results
- The conclusions you are drawing from the results with respect to your questions

I will review the documents you send me independently. This means that you need to explain your process and conclusions in the document well enough that I'll be able to understand them without you there to walk me through them. Remember that for many aspects of machine learning, there is not one right answer. This means that the most important thing is for me to see your decision-making process so that I know what factors you considered and why you made the choices that you did. It is also important that you demonstrate an awareness of the costs/benefits of the decisions you made relative to other options.

You should include a written description of your decision-making process *for each code chunk* in your Quarto documents. These descriptions may be briefer or more comprehensive depending on the scope of the code chunk. By decision-making process, I mean:

- What you did,
- Why you did it or chose to do it that way (vs. other options), and
- What you learned to inform next steps (if there is a result/output from that code chunk)

In addition, you will conclude your Quarto documents with a summary section where you will provide final, summary observations for parts 1 & 2 (see details below).

Part 1: Prediction

Overview

Your task is to build the best performing prediction model you can. You will be predicting a binary outcome from any or all of your available predictor variables. You may consider as many or few model configurations as you wish, and you may vary as many or as few characteristics (e.g., statistical algorithms, features/feature sets, hyperparameters, etc.) of these configurations as you wish. Remember to explain all your choices and decision-making processes to me, your collaborator.

Data

You will build a prediction model using the airline passenger satisfaction dataset, which provides a binary rating of whether a customer was satisfied with their airline experience or not as well as ~20 predictor variables. These predictor variables include customer characteristics (e.g., age), customer ratings of certain components of their airline experience (e.g., inflight WiFi), and flight characteristics (e.g., departure time delay).

Instructions & Questions

You should submit at least one rendered Quarto document (you may submit more than one depending on your workflow). Put your student ID number as your name in the document(s) and include the last four digits in your file name(s). Include as many code chunks in your document as you need. At a minimum, you should include:

- Code for *cleaning & modeling EDA* necessary to clean your data, understand your data, and build a (functional) prediction model.
- Code for *fitting and evaluating prediction models*.
- Your *decision-making process* for each code chunk (see general instructions above).

At the end of (one of) your Quarto document(s), provide a "summary section" that includes explanations regarding your decision-making process within each of the following domains. You do not need to reproduce any results, code, output, or visualizations that help support your answers, though you should refer to them clearly (e.g., referring to a named/numbered section within the file) as necessary. Here are some questions to guide your summary, though answering each question individually is not required – just explain what makes sense with your workflow and choices.

- **Spending your data:** How will you spend (i.e., divide/split) your data with respect to model fitting, model selection, and model evaluation? What factors did you weigh in your decision? What are the costs and benefits of the decision you made?
- **Model configurations:** Which model configuration characteristics will you vary? How will you know you've considered a wide enough range of characteristics to feel confident that you've built a good prediction model?
- **Model selection:** How will you evaluate model performance in model selection? What options did you consider, and why did you pick the metric that you did?
- **Model characterization:** How will you characterize your model's estimated performance in new data? What metrics do you need to look at to characterize performance comprehensively? What do these characterizations *mean* in "real world" terms? What are the costs of different types of errors that your model makes?

Part 2: Explanation

You have two options for Part 2: to use a dataset we provide you (Option 1), or to use your own dataset (Option 2). Details for each option appear below.

Option 1 (Provided Data)

Overview

You are interested in the effect of customer age on the percentage that people tip their servers in a restaurant. You value conducting rigorous science and were hoping to preregister your study, but guidance among existing literature is mixed with respect to how to handle key analytic decisions. Specifically, you're not sure about:

- What covariates (if any) you should include alongside your focal predictor and what (if any) transformations are needed for these covariates to enhance your ability to test your focal predictor
- Whether your focal predictor needs to be transformed in some way
- Whether you should allow your focal predictor to interact with customer sex
- What statistical algorithm will best capture the relationship between customer age and tip percentage
- How to handle outliers

You want to be principled in your analyses, and yet you don't want to lock yourself into the wrong analysis because of unclear existing guidance. You've heard about an alternative: using cross-validation to determine the correct model configuration, where "correct" means that it generalizes best in new data, thereby capturing the population data-generating process most closely.

Your task is to use cross-validation to identify the best model configuration for handling *2 or more analytic choices* to test the effect of customer age on tip percentage. Remember to explain all your choices and decision-making processes to me, your collaborator.

Data

You will be working with the tip.csv dataset, which seeks to predict the percentage of their check that a customer tips their server in a restaurant. It includes: an outcome variable, tip percentage; your focal predictor, customer age; and ten other variables (group size, total bill amount, sex of the server, day of the week, time [lunch or dinner], how many alcoholic drinks were ordered, whether the party had any children in it, sex of the customer, whether there was a smoker in the group, and whether any dessert was ordered).

Option 2 (Own Data)

Overview

You are interested in the effect of a focal predictor on an outcome from your own research. You value conducting rigorous science and were hoping to preregister your study, but guidance among existing literature is mixed with respect to how to handle key analytic decisions. Specifically, you're not sure about:

- What covariates (if any) you should include alongside your focal predictor and what (if any) transformations are needed for these covariates to enhance your ability to test your focal predictor
- Whether your focal predictor needs to be transformed in some way
- Whether you should allow your focal predictor to interact with OTHER VARIABLES
- What statistical algorithm will best capture the relationship between the PREDICTOR and outcome
- How to handle outliers

You want to be principled in your analyses, and yet you don't want to lock yourself into the wrong analysis because of lack of existing guidance. You've heard about an alternative: using cross-validation to determine the correct model configuration, where "correct" means that it generalizes best in new data, thereby capturing the population data-generating process most closely.

Your task is to use cross-validation to identify the best model configuration for handling *2 or more analytic choices* to test the effect of your focal predictor on your outcome variable. Remember to explain all your choices and decision-making processes to me, your collaborator.

Data

You may use your own data (e.g., from your own research). You may *not* simply choose your own dataset from somewhere online; this is an opportunity to apply these methods from class directly in your own data. Because I will not be familiar with your data, you should provide some brief descriptions at the outset of your Quarto document, such as:

- What is your outcome variable? What is your focal predictor?
- Are there any key characteristics of your data (e.g., high missingness, longitudinal/nested design, etc.) of which I should be aware? How will these characteristics affect your analytic choices?

Instructions & Questions (Applicable for Options 1 & 2)

You should submit at least one rendered Quarto document (you may submit more than one depending on your workflow). Put your student ID number as your name in the document(s) and include the last four digits in your file name(s). Include as many code chunks as you need. At a minimum, you should include:

- Code for *identifying the best model configuration* with respect to your 2 (or more) analytic decisions.
- Code for *fitting a final model* using that best configuration.
- Your *decision-making process* for each code chunk (see general instructions above).

At the end of (one of) your Quarto document(s), provide a "summary section" that includes explanations regarding your decision-making process within each of the following domains. You do not need to reproduce any results, code, output, or visualizations that help support your answers, though you should refer to them (e.g., referring to a named/numbered section within the file) clearly. Here are some questions to guide your summary, though answering each question individually is not required – explain what makes sense based on your workflow and choices.

- **Candidate model configurations:** Which model configurations did you consider? What characteristics of these configurations did you need to vary to explore how to handle your analytic decisions?
- **Resampling approach:** What resampling approach did you use? What factors did you weigh in your decision? What are the costs and benefits of the approach you chose?
- **Model selection:** What performance metric did you use for model selection? Why? How is this choice of metric related to (or unrelated to) your explanatory task?
- **Conclusions:** Have you answered your explanatory question?
    - If so… What is the answer? Are there other data or statistical tests you would want to support your answer?
    - If not… Why not? What other data or statistical tests do you need to reach a conclusion?
- **Reflection:** What benefits does this approach for selecting a best model configuration offer? Does it come with any limitations?