Midterm Project

Start Assignment

* **Due** Mar 23 by 11:59pm

* **Points** 100

* **Submitting** a text entry box, a website url, or a file upload

**Purpose**

The midterm project will give you experience with the following skills:

1. Using caret to fit machine learning models and select an optimal model.
2. Using double cross-validation to assess the predictive ability of your modeling process.
3. Interpreting and communicating your results for a general audience, including appropriate graphs.

The midterm project will help you practice and receive feedback as preparation for the final project.

**References**

Material / Methods covered in:

* Lessons 1 and 3–7 for supervised-learning methods.
* Lessons 2 and 8 for model assessment and selection.

**Data Sets**

Choose 1 of the data sets listed in [Data Set Descriptions](https://uweau.instructure.com/courses/566248/files/55290748?wrap=1)[Download Data Set Descriptions](https://uweau.instructure.com/courses/566248/files/55290748/download?download_frd=1)to analyze:

* [04cars.csv](https://uweau.instructure.com/courses/566248/files/55290765?wrap=1)[Download 04cars.csv](https://uweau.instructure.com/courses/566248/files/55290765/download?download_frd=1)
* [Athletes.csv](https://uweau.instructure.com/courses/566248/files/55290766?wrap=1)[Download Athletes.csv](https://uweau.instructure.com/courses/566248/files/55290766/download?download_frd=1)
* [Ozone\_LA.csv](https://uweau.instructure.com/courses/566248/files/55290767?wrap=1)[Download Ozone\_LA.csv](https://uweau.instructure.com/courses/566248/files/55290767/download?download_frd=1)
* [diabetes.csv](https://uweau.instructure.com/courses/566248/files/55290758?wrap=1)[Download diabetes.csv](https://uweau.instructure.com/courses/566248/files/55290758/download?download_frd=1)

**Task**

The midterm project involves four major phases: data preparation, fitting the models, interpreting the best model, and communicating the analysis and results. These phases can be broken down into the smaller steps listed below. We recommend reading through all the steps to understand “the big picture” before you begin working.

**Data preparation**

1. Choose 1 of the data sets listed in [Data Set Descriptions](https://uweau.instructure.com/courses/566248/files/55290748?wrap=1)[Download Data Set Descriptions](https://uweau.instructure.com/courses/566248/files/55290748/download?download_frd=1)to analyze. Although many of the methods we’ve covered in this class would be appropriate for these data sets, we have selected two appropriate methods for each data set. These are the methods you will compare. (This helps keep things simple for the midterm project, because you don’t need to spend time strategizing about which methods are likely to give the best results for a given data set.)
2. Think about who your target audience is. Why would someone be interested in predicting the response variable?
3. Read the data into R and do an exploratory data analysis and data cleaning. Things to consider:
   1. If there is missing data, how will you deal with it?
   2. Will you modify any of the predictors (such as log-transforming a quantitative variable or combining some categories of a categorical variable)?
   3. What range of values might be a good choice for any tuning parameters of your machine learning methods?

**Fitting the models**

1. Use caret to conduct a single layer of 10-fold cross validation on both of the modeling types indicated for the data set you chose. Tune at least one parameter for at least one of the modeling types. (Tuning multiple parameters may help improve your model’s performance.)
2. Write a for loop to conduct an outer layer of 5-fold cross-validation, containing both of the modeling types. Modify your code from the single layer of CV to use traindata.out rather than the full data set. Assess the performance of your model selection process.
3. Based on your results from steps 4 and/or 5, select one “best” model (including its tuning parameter values). Fit the model on the entire data set. (One way to do this is to extract the $finalModel component of the caret object from step 4, when used to perform a single layer of CV on the entire data set.)  This is the final model you will be interpreting in steps 7-8.

**Interpreting the best model**

1. Identify which 2 predictors are the most important in the model. Describe their relationship with the response.  (For example, are the predictors positively associated with the response? Negatively associated? Or something else?) Choose at least one of the following questions to answer:
   1. Does this (the most important predictors and their relationship with the response) make sense, based on your prior knowledge about this subject, or does it surprise you? Explain.
   2. Find a published news article or academic paper analyzing a response variable related to the one you analyzed. Does it agree with your analysis about the direction of the relationship between the predictors and the response? If not, what are some differences between your analysis and the article that could explain the difference in results?  (For example, a different population being studied, or an additional covariate.) Why might these change the results?
      * For example: “One possible explanation is that Smith et al.’s study focused on low-income patients in rural areas, for whom it was very difficult to visit the doctor. For these patients, one additional visit to the doctor represents a large percentage increase in medical supervision, resulting in a health improvement. In contrast, my data set includes patients who visited the doctor up to 20 times per year. Such large numbers of visits indicate health problems requiring close supervision. This resulted in a negative association between number of visits and overall health."
   3. What, if anything, can your target audience do to improve the response variable?  Is it plausible that the relationship between the predictors and the response is causal?  (There might be an action step for your target audience even if the relationship is not causal.  For example, even if buying shoes doesn’t cause an increased probability of buying socks (but they’re both associated with income and a general need for footwear), a company might still want to target sock ads to customers who have purchased shoes.)  If your two most important predictors are not actionable, choose another predictor that is actionable.  Describe its relationship with the response, and explain what your target audience could do about it.
2. Consider the accuracy/MSE of your model selection process (from step 5).  In your opinion, is your best model sufficiently accurate to be used for new predictions of individual data points?  (If you are doing a regression problem, it may be helpful to compute the MAE.)  If not, what specific thing(s) would you try in the future to improve your model?  (A different machine learning technique?  A different set of tuning parameters?  Additional predictor variables--if so, which variables would you want to gather data on?)  It’s fine to suggest gathering more data from a specific subpopulation (for example, “My model does poorly at classifying tennis players, so I would want to collect data from more tennis players.”), but simply suggesting “more data” without specifics is not a useful recommendation.

**Communicating the analysis and results**

1. Use R to create clear, readable graphs that communicate either or both of the following:
   * The relative performance of different models/combinations of tuning parameters
   * The relationship between the response (or predicted response) and one or more of the predictors you examined in step 7   
     Optionally, you may also use R to create clear, readable graphs that communicate either or both of the following:
   * Why you chose the data cleaning that you did
   * Which predictors are most important in the best model

You are encouraged to experiment with different ways of graphing your data.  However, to demonstrate your critical thinking and good judgement, select the most useful 1-5 graphs to include in your executive summary.

You are encouraged to view the [Gallery of Graphs Links to an external site.](https://media.uwex.edu/content/ds/ds740_r21/ds740_lesson8_presentation3.sbproj/)to give you ideas about graphs that might be useful for your data and/or models.

Optionally, you may include 0-2 tables that demonstrate useful information about your data and/or models.  For purposes of this project, tables are **not** a substitute for graphs.

Your report should **not** include raw R output (for example, from *summary( )*).  We do not recommend including a table that is just a list of variables.  If you include a table of variables, the table should also include additional information to add value (such as a summary of how each variable was cleaned).

1. Prepare an executive summary of 500-1200 words summarizing your analysis and interpretation.  Your summary should include the following:
   * An introduction that catches the reader’s attention and explains why the target audience should be interested in predicting the response variable.
   * An explanation of what data cleaning you did and why.
   * An explanation of what values you used for the tuning parameters of the machine learning methods.
   * A summary of which model you selected as the best, and a discussion of your interpretation of the model (from step 7).
   * A numerical assessment of the accuracy of the model-fitting process from the double CV, and a discussion of your interpretation of the accuracy (from step 8)
   * The 1-5 graphs (and 0-2 tables) from step 9.  Each graph and table should connect to the text of your executive summary.  For example, “Logistic regression had the highest ROC curve (Figure 1), indicating that it was the most accurate model.”
   * Clear, professional writing with transition sentences where relevant.  You are encouraged to make use of the UW [Online Writing Lab Links to an external site.](https://owl.wisconsin.edu/)for help with editing your executive summary.
2. Submit
   * Your executive summary as a .docx or .pdf file. Figures may be embedded with the text or included on an additional page. References, if any, may be included at the end of the executive summary.
   * A script containing the code you used to analyze the data. This should be a clean, commented, final version of the code, in either a **.R file or a .Rmd file**.   Code maximum is 800 lines (note that a complete, efficient analysis can reasonably be done within 200-450 lines of code, including commenting).
   * Do not put your executive summary and code into a zip file.

**Midterm Project Rubric**

| Midterm Project Rubric | | |
| --- | --- | --- |
| **Criteria** | **Ratings** | **Pts** |
| This criterion is linked to a Learning OutcomeTechnical accuracy | |  |  |  | | --- | --- | --- | | **35 to >33.0 pts**  **Excellent**  Double CV is implemented correctly and Data cleaning is appropriate and well-explained and Code is clear, correct, and well-commented (34-35pts) | **33 to >27.0 pts**  **Adequate**  Small bugs in double CV code or Connection between code and analysis described in executive summary is not clear (28-33pts) | **27 to >0 pts**  **Weak**  Only 1 layer of CV or Major errors in code or Very inappropriate analysis choices (0-27pts) | | 35 pts |
| This criterion is linked to a Learning OutcomeInterpretation | |  |  |  | | --- | --- | --- | | **30 to >28.0 pts**  **Excellent**  A thorough, thoughtful answer to one of questions a-c (in step 7) and A thorough, thoughtful discussion of the model accuracy (in step 8) (29-30 pts) | **28 to >22.0 pts**  **Adequate**  Answers to steps 7 and 8 are acceptable but one or both could be more thorough (24-28 pts) | **22 to >0 pts**  **Weak**  One of steps 7 or 8 omitted or Interpretation contains objectively false statements (0-23pts) | | 30 pts |
| This criterion is linked to a Learning OutcomeCommunication | |  |  |  | | --- | --- | --- | | **30 to >28.0 pts**  **Excellent**  Clear justification of why the analysis is important and 1-5 graphs, all of which are clear, appropriate, and well-connected to the analysis/interpretation and Writing is professional and engaging (29-30pts) | **28 to >23.0 pts**  **Adequate**  Weak or implicit justification of why the analysis is important or Some graphs are hard to read, or the connection between the graph and the analysis/interpretation is unclear or Writing is mostly professional but could be more engaging (24-28pts) | **23 to >0 pts**  **Weak**  No graphs or One or more graphs is wholly inappropriate for the data or Problems with organization or grammar interfere with reader’s ability to understand the analysis (0-23pts) | | 30 pts |
| This criterion is linked to a Learning OutcomeExceed Expectations  The last 5 points will be added to reports that exceed expectations in one or more of the following ways: | |  |  | | --- | --- | | **5 pts**  **Excellent**  - The analysis is exceptionally thorough. This could involve an exceptionally thorough job of data cleaning, or an exceptionally thoughtful exploration of the tuning parameter space. - The writing is exceptionally eloquent, or the layout of the paper (including graphs and tables) is exceptionally well designed. - The interpretation and discussion of results is exceptionally thorough. This could involve answering multiple questions about the important predictors in a thoughtful, detailed way. | **0 pts**  **None** | | 5 pts |
| Total Points: 100 | | |