Final Project

Start Assignment

* **Due** May 2 by 11:59pm

* **Points** 100

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

**Purpose**

The final project is your opportunity to apply what you have learned in this course to answer a question that interests you, by analyzing a real-world data set of your choice.

This final project is designed to guide you to:

1. Apply the methods you have learned and practiced in the course via a more free-form application.
2. Appropriately evaluate the methods used to analyze the data.
3. Appropriately interpret the results.
4. Demonstrate creativity and sophistication in your analysis and coding.
5. Communicate your methods, results, and interpretations to a general audience.

**References**

| *References* | |
| --- | --- |
| **For information about:** | **Review:** |
| Supervised-learning material and methods | Lessons 1, 3-7, 9-10 |
| Unsupervised learning material and methods | Lessons 11-12 |
| Model selection and assessment | Lessons 2, 8 |
| Unsupervised grouping assessment | Lessons 11-12 |

**Task**

The final project involves four major phases:  data preparation, fitting the models, interpreting the models, 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. Read the feedback provided on your final project proposal.
2. Decide whether you will use R or another programming language. Be aware that your instructors will be able to provide greater assistance with troubleshooting if you choose R than if you choose a different language.
3. Read the data into R or another program 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?

**4. Fitting the models**

* If you are using **two supervised learning techniques:**[Steps](https://uweau.instructure.com/courses/566248/files/55290727?wrap=1" \t "_blank)[Download Steps](https://uweau.instructure.com/courses/566248/files/55290727/download?download_frd=1)
* If you are using **one supervised and one unsupervised learning technique:**[Steps](https://uweau.instructure.com/courses/566248/files/55290772?wrap=1" \t "_blank)[Download Steps](https://uweau.instructure.com/courses/566248/files/55290772/download?download_frd=1)
* If you are using **two unsupervised learning techniques:**[Steps](https://uweau.instructure.com/courses/566248/files/55290779?wrap=1" \t "_blank)[Download Steps](https://uweau.instructure.com/courses/566248/files/55290779/download?download_frd=1)

**5-7. Interpreting the models**

* [Interpreting the best model: Supervised learning](https://uweau.instructure.com/courses/566248/files/55290777?wrap=1)[Download Interpreting the best model: Supervised learning](https://uweau.instructure.com/courses/566248/files/55290777/download?download_frd=1)
* [Interpreting the model: Association Rules](https://uweau.instructure.com/courses/566248/files/55290776?wrap=1)[Download Interpreting the model: Association Rules](https://uweau.instructure.com/courses/566248/files/55290776/download?download_frd=1)
* [Interpreting the model: Clustering](https://uweau.instructure.com/courses/566248/files/55290778?wrap=1)[Download Interpreting the model: Clustering](https://uweau.instructure.com/courses/566248/files/55290778/download?download_frd=1)
* [Interpreting the model: PCA](https://uweau.instructure.com/courses/566248/files/55290733?wrap=1)[Download Interpreting the model: PCA](https://uweau.instructure.com/courses/566248/files/55290733/download?download_frd=1)

**Communicating the analysis and results**

1. Use R to create clear, readable graphs that communicate at least one of the following:
   * Supervised learning:
     + 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 steps 5-6
   * Association rules:
     + The relationship between the antecedent and the consequent in one of the rules you selected in part 6
       - For example, this could be a conditional bar graph showing the probability of the consequent among data points with and without the antecedent.
     + The trends in the items included in the subsets of rules you selected in part 7.
       - For example, this could be a bar graph of the number of rules involving different items, with colors to indicate each subset of rules.
   * Clustering:
     + The relationship between which cluster each data point is in and 1-2 other variables
   * PCA:
     + How the variables contribute to one or two principal components
       - For example, this could be a biplot or a bar graph of the loadings.
     + How the data are arranged in the space of principal components
       - For example, this could be a biplot.

Optionally, you may also use R to create clear, readable graphs that communicate one or more of the following:

* Why you chose the data cleaning that you did
* Which predictors are most important in the best model
* Why you chose the unsupervised learning parameters that you did

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 1-6 graphs to include in your executive summary.

You are encouraged to view the Gallery of Graphs 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 500-1700 words summarizing your analysis and interpretation. Your summary should include the following:
   * An introduction that catches the reader’s attention, states the question you investigated, and explains why the target audience should be interested in the answer.
   * 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 discussion of your results and interpretations from steps 5-7.
   * The 1-6 graphs (and 0-2 tables) from step 8.  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. Note: You may use R or another programming language of your choice. Be aware that your instructor will be able to provide greater assistance with troubleshooting if you choose R than if you choose a different language.
   * This should be a clean, commented, final version of the code.
   * If you are using R, please submit either a .R file or a .Rmd file.
   * If you are using Python, please submit either a .py or .ipynb file.
   * If you are using another programming language, please consult with your instructor to identify an appropriate file format.
   * Code maximum is 800 lines (note that a complete, efficient analysis can reasonably be done within 250-500 lines of code, including commenting).
   * Do not put your executive summary and code into a zip file.

**Examples**

Below, you can find excellent examples from past DS 740 students who have given permission to share their work.

* [Final\_Project\_example-Airbnb.pdf](https://uweau.instructure.com/courses/566248/files/55290729?wrap=1)[Download Final\_Project\_example-Airbnb.pdf](https://uweau.instructure.com/courses/566248/files/55290729/download?download_frd=1)
  + This project used ANNs, PCA, and clustering. The interpretation of the clusters is especially effective.
* [Final Project Executive Summary\_example.pdf](https://uweau.instructure.com/courses/566248/files/55290773?wrap=1)[Download Final Project Executive Summary\_example.pdf](https://uweau.instructure.com/courses/566248/files/55290773/download?download_frd=1)   
  + This project used logistic regression and ANNs. The ROC curves and interpretation of the Lek profile are especially effective.
* [DS740\_final\_project\_baseball\_example.pdf](https://uweau.instructure.com/courses/566248/files/55290774?wrap=1)[Download DS740\_final\_project\_baseball\_example.pdf](https://uweau.instructure.com/courses/566248/files/55290774/download?download_frd=1)   
  + This project used random forests and ANNs. The nontechnical summary of the methods used, and the discussion of potential ways to improve the models, are especially effective.
* [DS740\_FinalProject\_Exec\_Summary\_example.pdf](https://uweau.instructure.com/courses/566248/files/55290775?wrap=1)[Download DS740\_FinalProject\_Exec\_Summary\_example.pdf](https://uweau.instructure.com/courses/566248/files/55290775/download?download_frd=1)   
  + This project used association rules and random forests.  It does an especially good job of summarizing the association rules and relating them to the results from the random forest and to the action steps.
* [DS 740 Final Project\_example.pdf](https://uweau.instructure.com/courses/566248/files/55290728?wrap=1)[Download DS 740 Final Project\_example.pdf](https://uweau.instructure.com/courses/566248/files/55290728/download?download_frd=1)   
  + This project used PCA, clustering, and association rules.  The interpretation of the clusters is especially effective.

Rubric

**Final Project**

| Final Project | | |
| --- | --- | --- |
| **Criteria** | **Ratings** | **Pts** |
| This criterion is linked to a Learning OutcomeTechnical accuracy | |  |  |  | | --- | --- | --- | | **35 pts**  **Excellent**  Data cleaning is appropriate and well-explained and Code is clear, correct, and well-commented and If performing supervised learning, double validation is implemented correctly (35 points) | **33 pts**  **Adequate**  Connection between code and analysis described in executive summary is not clear or If performing supervised learning, small bugs in double validation code or Cross-validation was computationally feasible, but a validation set was used because it was easier (28-33 pts) | **26 pts**  **Weak**  Major errors in code or Very inappropriate analysis choices or If performing supervised learning, only 1 layer of CV (0-26 pts) | | 35 pts |
| This criterion is linked to a Learning OutcomeInterpretation | |  |  |  | | --- | --- | --- | | **30 pts**  **Excellent**  Thorough, thoughtful answers to steps 5-7 and At least one of steps 5-7 is clearly connected to the original research questions (30 pts) | **28 pts**  **Adequate**  Answers to steps 5-7 are acceptable but one or more could be more thorough or Connection between analysis/interpretation and original research questions is unclear (24-28 pts) | **22 pts**  **Weak**  One of steps 5-7 omitted or Interpretation contains objectively false statements (0-22 pts) | | 30 pts |
| This criterion is linked to a Learning OutcomeCommunication | |  |  |  | | --- | --- | --- | | **30 pts**  **Excellent**  Clear justification of why the analysis is important and 1-6 graphs, all of which are clear, appropriate, and well-connected to the analysis/interpretation and Writing is professional, engaging, and of appropriate depth (30 pts) | **28 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 or concise (24-28 pts) | **22 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 (0-22 pts) | | 30 pts |
| This criterion is linked to a Learning OutcomeExceed expectation | |  |  | | --- | --- | | **5 pts**  **Exceed Expectations**  The last 5 points will be added to reports that exceed expectations in one or more of the following ways: - 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**  **No Marks** | | 5 pts |
| This criterion is linked to a Learning OutcomeCoding Effort  Submitted substantial coding work towards the planned cleaning and analyses. | |  |  | | --- | --- | | **10 pts**  **Full Marks** | **0 pts**  **No Marks** | | 10 pts |
| This criterion is linked to a Learning OutcomeWriting Effort  Submitted coherent and substantial written work towards the interpretation and/or use of the analysis. | |  |  | | --- | --- | | **5 pts**  **Full Marks** | **0 pts**  **No Marks** | | 5 pts |
| Total Points: 115 | | |

[Previous](https://uweau.instructure.com/courses/566248/modules/items/16855586)