

Final Project Guidelines and Rubric

Overview

You must complete a **data analysis research project** as your final project for this course. Your research project will focus on a real-world topic of your choice, as approved by me. You will pick a topic from the list provided or with my approval, and create a data analysis plan and machine learning model (e.g., a decision tree model) based on a real-world scenario. This assessment will provide you with the opportunity to employ highly valued decision support skills and concepts for data within a real-world context. You can also reference the Final Project Notes document.

The project is divided into **four milestones**, which will be submitted at various points throughout the course to scaffold learning and ensure quality final submissions. These milestones will be submitted in separate modules. **The final submission will occur in the final module.**

This project will address the following course outcomes:

- Appraise data in context according to industry-standard methods and techniques for its utility in supporting decision making
- Determine suitable data manipulation and modeling methods for decision support
- Articulate data frameworks for organizational decision support by applying data manipulation, modeling, and management concepts
- Evaluate the ethical issues surrounding organizational use of decision-oriented data based on industry standards and one's personal ethical criteria
- Create and assess the agility of solutions through application of data-mining procedures for decision support in various industries

Prompt

Your data analysis model and report should answer the following prompt: How does your model and evaluation resolve uncertainty in making a decision? In order to produce your analytic report, you will need to choose and investigate a data set using the data analysis and visualization techniques you learned in class. Then you will formulate a research question, write an analytic plan, and implement it. Your report should not solely consist of descriptions of what you did. It should also contain detailed explorations into the meaning behind your model and the implications of its results. You will also be testing your model's fitness and evaluating its strengths and weaknesses.

The project in a nutshell:

1. Choose a data set (get ideas from the source list in the spreadsheet Final Project Topics and Sources.xls)
2. Formulate your data analysis research question
3. Write an analytic plan and choose a Machine Learning Algorithm/Model
4. Perform the implementation of the algorithm/model (e.g., top-down or bottom-up modeling in decision trees)
5. Perform model diagnostics and Evaluate

These activities are broken up into milestones so that the work is spread throughout the term and you can get early assistance with any obstacles.

A data analysis report is similar to any other analytic report. These reports introduce a problem, state a line of inquiry, explain a model that the author developed, discuss results and limitations, and then make conclusions and recommendations. Some decision models seek the best expected value among a discrete set of choices. Other decision analyses might seek the threshold values at which the model changes from one recommendation to another, describe the implications, and leave it to the reader to decide what to do. Still other decision models might look for the likeliest path to explain patterns that are already present in a data set. In all cases, they have something in common: They are trying to help resolve uncertainty. Your job is to bring clarity to the decision being made.

Data analysis seeks less to produce a definitive result, and more to accurately explain the combinations of possibilities that can lead decision makers to clearer choices. This is the modeling aspect. If you model the weather but never take into account barometric pressure, your model would fail if trying to determine the worst hurricane trajectories. These are the kinds of things you will be looking at in your decision models: searching for ways to explain the conditions that produce outcomes and to evaluate the strengths and weaknesses of the models you produce.

The three main ideas that your report should encompass are your ability to formulate a data analysis research question based on an appropriate data set, develop your model, and finally evaluate the model's utilities, results, strengths, and weaknesses. In short, if your report fully encompasses these three concepts, you will produce an authentic document that would stand on its own in a professional setting.

Data sources to choose from: The included spreadsheet lists data sets used in previous classes. Students found these data sets, prepared them for modeling on their own, and wrote excellent papers on the topics. Remember that your data set needs to be appropriate for modeling a discrete set of choices. Either those choices are built into the model as categorical variables, or you will need to do some legwork by converting continuous variables into rational categorical groups. This activity would be part of the data preparation and documented in your data appraisal section.

Your final project must include the following sections:

- **Title Page**
- **Abstract:** 300 words or less
- **Table of Contents**
- **Introduction, with research question:** Up to two pages
- **Data Appraisal:** Up to two pages
- **Techniques (a.k.a. Methods):** Up to two pages
- **Evaluation:** Two to four pages
- **Model,** including optimizations
 - Up to one page for graphic(s)
 - Up to two pages for model explanation
- **Results:** Up to two pages

- **Limitations:** Up to two pages
- **Conclusion:** Up to two pages

Sources: Note that the core elements add up to about 15–20 pages, double-spaced. **The overall target for the core elements is still 15–20 pages**, so that you have room to adjust each section according to the needs of the project. Everything you need to say in the report should fit within 15–20 double-spaced, 12-point font pages with one-inch margins.

To see some good final projects, consult the exemplars. Not all of them are 100% perfect papers, but they do embody the level of complex thinking that characterizes an interesting project. The idea behind the page limit is to explore the concept of “less is more.” If you add up the text, graphics, sources, and supporting material from all the milestones, you end up with 15 to 20 pages. For the final, that means some compression needs to occur. This means finding the most important information from what you have previously written and leaving room for the new parts that you need to write. Follow the list of required elements for the final to guide how to structure your research paper.

Specifically, the following **critical elements** must be met in your final submission:

- I. **Introduction:** Analyze the purpose, type, intended populations, and uses of the analysis to establish an appropriate context for the data-mining and analysis plan.
- II. **Data Appraisal**
 - A. **Characterize** the data set. For example, what is the purpose such data are generally used for?
 - B. Appraise the data within the **context** of the problem to be solved and industry standards. How will you use the data? For example, expound upon the limitations of the data set in the context of your needs.
 - C. Explain the **utilities** that you will be using and how the data supports that choice.
- III. **Select Appropriate Techniques**
 - A. Determine and explain the appropriate steps for **preparation** of the data sets into a usable form: what steps were taken to make data descriptions clear, how extreme or missing values were addressed, and how data quality was improved.
 - B. Determine the appropriate steps (including: risk assessment, probability calculations, and modeling techniques) for data **manipulation** and in-depth analysis to support organizational decision-making.
 - C. Models and **checkpoints**: How will you optimize the models, what will you test for, and how will you build in checks to determine a successful analysis?
 - D. **Defend** the ethicality and legality of the analytic selections made for use, interpretation, and manipulation of the data based on industry standards for legal compliance, policies, and social responsibility. If there are no potential ethical and legal compliance issues, explain how your prep and use of this data are both ethical and legal.
- IV. **Defend and Evaluate Choices**
 - A. Why are these choices the **best** for the data and problem at hand? What research or industry standards are supportive of your choices of methods? Explain how the methods chosen will support organizational decision-making.

- B. Determine the **agility** of these choices for decision support based on research and relevant examples: how can they be adapted to alternative needs or reapplied to future analysis?
 - C. **Address** ethical and legal **issues** that might arise from the use and interpretation of the data, based on industry standards, policies, and social responsibility. How can you ensure that your selected procedures, use of data, and results will be socially responsible and in line with your own ethical standards?
 - D. **Implement** your plan: Perform data preparation, mining and modeling procedures, and create your decision support solution.
- V. **Machine Learning Model** (e.g., bottom-up or top-down decision trees): Include the detailed process and programming steps necessary to complete the analysis. Be sure to:
- A. Defend the overall **structure** and purpose of the machine learning model in the data analysis.
 - B. Develop process-**documentation** that addresses potential complications. This piece should resemble a recipe/outline that provides enough information for addressing potential implementation issues.
 - C. Evaluate the **results** of your machine learning model. At minimum, attend to the following:
 - 1. Are the results reasonable?
 - 2. How accurate is your model? Are there missing or extraneous elements that could have influenced your results?
 - 3. What common errors are made during creation of the model you chose? How did you ensure that you did not make these errors?
- VI. **Articulation of Response/Final Report**: Utilizes visualization options that effectively address the needs of the audience. Options may include annotated shell tables, visualizations, and a compositional structure.

To guide you in writing your final paper, follow the Final Project Rubric. The rubric is less about format and more about thought. Specifically, you should write sections that detail the limitations and justification for your analysis. You should also take the time to address any ethical or legal issues that connect with your results or decisions being analyzed. You should annotate and caption your graphics. You could include a table that characterizes the data set. You should address what your model does to assist decision makers. You should defend your choices of variables and groupings. Lastly, you should address the agility of your analysis and how it might be applied to future uses.

Milestones

Milestone One: Choose a Data Set and Formulate Data Analysis Research Question

In Milestone 1, you will choose a data set from the curated list of sources (Final Project Topics and Sources.xls), or you may submit your proposal for a different data source than those listed. Then you will write a data analysis research question, which should be two to three pages in length and framed as a discrete set of choices to be analyzed.

Milestone Two: Write Introduction

In this milestone, you will write the introduction to your final paper. Specifically, the introduction should explain and discuss your research, give the appropriate context of the analysis for your reader, and explicitly state your research question. A good introduction not only gives background material, but it also dives into how data analysis principles specifically apply to the situation.

Milestone Three: Develop Data Analysis Model

In this Milestone, you will draft your decision tree or other machine learning model. This task presupposes a data set, a viable data analysis research question, and the necessary data prep. To complete this milestone, you *may* have to experiment with different modeling styles. The main objective is to draft your model, explain what you did, and explain why it is the best model for your research question.

Milestone Four: Revise and Evaluate Data Analysis Model

In this Milestone, you will revise and evaluate your machine learning model based on the feedback you received from the instructor for the previous milestone. Evaluation in this case could mean a few different things. For example, if you are performing a bottom-up style recursive partitioning analysis with decision trees, you should report on the error rate and variable selection; you might also consider alternative variable categorizations to improve your model. As another example, if you are performing a top-down decision tree modeling exercise, what are the threshold values that cause the tree to flip? You should perform sensitivity analysis on the critical variables in your tree and report what those sensitivity analyses are telling you. For any style of modeling, you should consider things like what makes your model strong, what breaks your model, etc.

Final Submission: Data Analysis Model and Report

In the **Final Module**, you will submit your data analysis model and report, compiling all the components used to develop the model and produce the report, as well as a leading abstract, table of contents, and in a format that addresses all of the critical elements in the instructions. The project should include sections that detail the limitations and justification for your analysis. You will probably be compressing what you wrote for your introduction to make it fit within the eight-page limit. You should also take the time to address any ethical or legal issues that connect with your results or decisions being analyzed. Lastly, you should address the agility of your analysis and how it might be applied to future uses. **This assignment is graded with the Final Project Rubric.**

Deliverables

Milestone	Deliverables	Grading
One	Research Question	Graded separately; Milestone One Rubric
Two	Write Introduction	Graded separately; Milestone Two Rubric
Three	Develop Data Analysis Model	Graded separately; Milestone Three Rubric
Four	Revise and Evaluate Model	Graded separately; Milestone Four Rubric
	Data Analysis Model and Report	Graded separately; Final Project Rubric

Final Project Rubric

Guidelines for Submission: The final report will be a 15–20 page research paper, double-spaced, in 12-point Times New Roman font with one-inch margins all around and APA citations. Title page, abstract, appendices and bibliography of sources are extra beyond the 15–20 pages of the report. You may include one page or less of annotated/captioned graphics as part of the report. The purpose of the limits is to keep the discussions compact and to maintain the integrity of publication-quality research.

Instructor Feedback: This activity uses an integrated rubric in Blackboard. Students can view instructor feedback in the Grade Center.

Critical Elements	Exemplary (100%)	Proficient (90%)	Needs Improvement (70%)	Not Evident (0%)	Value
Introduction	Meets “Proficient” criteria and cites specific, relevant examples to establish a robust context for the data-mining analysis plan	The purpose, type, intended populations, and uses of the analysis report are analyzed to establish an appropriate context for the data-mining analysis plan	The purpose, type, intended populations, and uses of the analysis report are not sufficiently analyzed to establish an appropriate context for the data-mining analysis plan	Either the purpose, type, intended populations, or uses of the analysis report are not analyzed	6.25
Data Appraisal: Characterize	Meets “Proficient” criteria and claims are qualified with source evidence or examples	Makes accurate claims about the general use of the dataset(s) and the intended purpose of the data	Not all claims about the general use of the dataset(s) and the intended purpose of the data is accurate given the available evidence	Does not make claims about the general use of the dataset(s) and the intended purpose of the data	6.25
Data Appraisal: Context	Meets “Proficient” criteria and qualifies claims specific to discrete needs of the organization	Makes accurate claims about the data within industry standards and the context of the problem to be solved	Not all claims about the data are accurate based on industry standards and the context of the problem to be solved	Does not make claims about the data based on the context of the problem to be solved and industry standards	6.25
Data Appraisal: Measurable Utilities	Meets “Proficient” criteria and supporting explanation is qualified with examples or research evidence	Makes accurate determination and thoroughly explains the measurable utilities and how the data supports that choice	Determination of unit of analysis is not entirely accurate or explanation does not thoroughly explain how the data supports measurable utilities determination	Does not determine a measurable utilities	6.25
Select Appropriate Techniques: Preparation	Meets “Proficient” criteria and quality of explanation allows for a seamless delivery of the initial molding process	Makes appropriate analysis step selections and explains the process for preparing the raw data	Not all analysis step selections are appropriate for preparing the raw data, or not all step processes are sufficiently explained	Does not select and explain analysis steps for preparing raw data into a useable form	6.25

Select Appropriate Techniques: Manipulation	Meets “Proficient” criteria and step selection and explanations are seamlessly integrated into a clear process	Makes appropriate step selections and explains the process of steps for in-depth analysis and manipulation of the data to support organizational decision making	Not all steps are appropriate for in-depth analysis and manipulation in support of organizational decision making or not all steps are explained in terms of process	Does not select and explain in-depth analysis and manipulation steps for decision support	6.25
Select Appropriate Techniques: Checkpoints	Meets “Proficient” criteria and the explanations of the selections provide clear and seamless integration of steps into the overall manipulation process	Makes appropriate algorithm selections, and explains the process of the selections, for the optimization, risk assessment, and built-in check points to ensure the success of data analysis and manipulation	Not all algorithm selections and explanations of process for optimization, risk assessment, and built-in check points are appropriate to ensure successful data analysis and manipulation, or key valuable methods are missed	Does not select and explain the process of algorithm selections for optimization, risk assessment, and built-in checkpoints	6.25
Select Appropriate Techniques: Defend	Meets “Proficient” criteria and substantiates claims with scholarly research evidencing considerations of social responsibility	Makes and justifies claims about the ethical and legal issues related to the use, interpretation, and manipulation of the data for the decisions being made, based on industry standards, laws, and organizational policies	Not all claims about the ethical and legal issues related to the use, interpretation, and manipulation of the data for the decisions being made are justifiable based on industry standards, laws, and organizational policies	Does not make claims about the ethical and legal issues related to the use, interpretation, and manipulation of the data for the decisions being made	6.25
Defend and Evaluate Choices: Best	Meets “Proficient” criteria and substantiates claims with research in specific support of the decisions/problem at hand	Makes and justifies claims about the appropriateness of the methods for manipulation and algorithm selections made for decision support based on analysis of industry standards and valid research	Not all claims about the appropriateness of the methods for manipulation and algorithm selections made are justifiable based on analysis of industry standards and valid research	Does not make and justify claims about the appropriateness of the methods for manipulation and algorithm selections made	6.25
Defend and Evaluate Choices: Agility	Meets “Proficient” criteria and substantiates claims with scholarly research and real world examples	Makes and justifies claims about the agility of the choices made for decision support in various industries, projects, and organizations with research and relevant examples	Not all claims about the agility of the choices made for decision support in various industries, projects, and organizations are justifiable based on the provided research and examples	Does not make claims about the agility of the choices made for decision support in various industries, projects, and organizations	6.25

Defend and Evaluate Choices: Address Issues	Meets “Proficient” criteria and the details of the explanation expound upon social responsibility and industry standards	Details the ethical considerations that should be made about use of the results of the solution and how ethical use can be ensured	Explains ethical considerations for use of the results of the solution, but lacks detail or does not explain how ethical use can be ensured	Does not explain ethical considerations for use of solution results	6.25
Machine Learning or Decision Tree Model: Implement	Meets “Proficient” criteria and performance of mining process and accuracy of decision solution evidence appropriate planning and implementation of plan within the context of the selected topic	Correctly performs the data mining process and creates an accurate decision support solution	Performs the data mining process and creates a decision support solution, but solution is not accurate	Does not perform the data mining process and create a decision support solution	6.25
Machine Learning or Decision Tree Model: Structure	Meets “Proficient” criteria and substantiates claims with scholarly evidence and real world examples	Makes and justifies claims about the overall structure and purpose of model for organizational decision support based on specific examples and research	Not all claims about the overall structure and purpose of model for decisions support are justifiable	Does not make claims about the overall structure and purpose of model for organizational decision support	6.25
Machine Learning or Decision Tree Model: Documentation	Meets “Proficient” criteria and model is of quality to allow others to develop further, more detailed models to address possible issues	Outline effectively acts as process documentation for addressing potential complications during implementation of the analysis plan	Not all aspects of outline would be effective in addressing the potential complications of implementation, or common major issues are not addressed	Does not include an outline for addressing potential complications during implementation	6.25
Machine Learning or Decision Tree Model: Results	Meets “Proficient” criteria and comprehensively evaluates against criteria above the given criteria and specifically relevant to the context of the selected topic	Accurately evaluates the results of the decision tree model against the given criteria	Evaluates the results against the given criteria, but with gaps in accuracy	Does not evaluate the results against the given criteria	6.25
Articulation of Response	Submission is free of errors related to citations, grammar, spelling, syntax, and organization and is presented in a professional and easy to read format	Submission utilizes visualization options that effectively address the needs of the audience and has no major errors related to citations, grammar, spelling, syntax, or organization	Submission utilizes various visualization options that don’t effectively address the needs of the audience or has major errors related to citations, grammar, spelling, syntax, or organization that negatively impact readability and articulation of main ideas	Submission does not utilize visualization options for the audience or has critical errors related to citations, grammar, spelling, syntax, or organization that prevent understanding of ideas	6.25
Earned Total					100%