Team Project Instructions

Introduction

Throughout this project, you will act out a scenario in which you and your team are analysts retained by a company (large or small) or funding source (e.g., a VC firm or incubator) who have asked you to use data science techniques to address a chosen business problem. You and your team will design the data science task, analyze the data, and describe your results.

Your own data and results need not be on par with actual industry results—the goal is for you to get a realistic hands-on experience given the constraints and techniques you've learned. Don't worry too much about coming up with a novel idea; it's more important to develop the idea well (within the scope of what we've discussed in class).

You must choose a **classification problem** (we will discuss this in Module 2) and use the "data science process" to structure your research and write-up. Remember that proceeding linearly through the steps may be ineffective, and this may need to be reflected in your analysis.

You should interact with the instructor from preparing your initial ideas through your write-up, just as a consulting group would interact with a firm or funding source in preparing a report. Feel free to ask the instructor to help fill in any gaps between the available material and what you would like to discover.

Schedule of Deliverables

Use this schedule of deliverables as a birds-eye view of what items will be due in each module.

Module	Deliverable
Module 1	Nothing due
Module 2	Nothing due (form your team and choose a dataset)
Module 3	Proposal
Module 4	Nothing due
Module 5	Nothing due
Module 6	Status Report/Outline
Module 7	Nothing Due
Module 8	Report Presentation slides R Markdown file and Artifacts

Deliverables in Details

Each deliverable will build upon the last, culminating in the final report and presentation. Please take time to review the instructions for each deliverable.

Week 2: Form your team, Choose a Dataset and Business Question

Teams will select their data set via a Google Sheet under Modules > Project > Dataset. You will not need to submit anything via Canvas this week. However, there will still be work to do: after choosing your dataset, formulate your main business question.

Another option for your project data is to select one from the following possible dataset links:

Possible datasets:

- UCI Machine Learning Repository
- Kaggle
- KD Nuggets
- R Bloggers
- FiveThirtyEight

Also, you may use a dataset from your own work as long as the data can be shared without any NDA, IRB, etc., involved. Please contact your instructor if you decide to use a dataset from your work.

M3 Proposal

Teams will complete a proposal that addresses their understanding of the dataset and formulated business question by answering the following:

- What is your business problem and business question?
- What is the data instance (entity)? (That is, what does each data point represent? A customer, a country, a product, etc.)
- What might be the target variable (the variable of interest to be predicted)?
- What is your proposed method to test or examine the problem?
- How will the results be used, and how will they provide business value?

M6 Outline (Status Report)

Teams will submit an outline of their work so far. Please make sure to include feedback from the proposal submission and the following:

- Business problem and question
- Data understanding and prep process
- Identify the target variable
- Preliminary results (if any)
- Work plan for the next two weeks

M8 Report

The write-up should be a maximum of 10 pages, single-spaced, including any appendices you would like to include (where appropriate). Use external sources where appropriate, with clear citations and a bibliography.

Your report should follow the rubric and contain all the steps in the data science process:

- Business understanding and business question
- Data understanding and data preparation, including identifying a target variable, data visualization, some descriptive summary statistics, attribute understanding, etc.
- Modeling and results
- Model tuning and evaluation
- Discussion and limitations, including deployment-related issues, the potential hazards, and bias your finding might result in if deployed
- All group members should contribute to the analysis and report.

M8 Presentation

Each team will present their project to their employers in the classroom. The details for the time limit and others will be announced in the class.

All group members must contribute (have face-time) during the presentation.

M8 Presentation Slide Deck

You may choose to use PDF or PPT; the length (number of pages) or master design is not restricted.

M8 Supporting Data and Artifacts

This includes

- R Markdown file that contains all your codes and comments for them
- Data source if your group selected a dataset from your own work or Potential sources outside the list in the Google sheet (on Canvas Project module)

Rubric

Assessment Criteria	Very Good	Good	Not Good Enough
Business Understanding (15 pts)	• Thorough and clear discussion of the business question and objective. • The target variable was appropriate for the business problem and was clearly defined to be readily used for analysis.	• Demonstrate good understanding of business problem and reasonably clear objective of the project. • The target variable was appropriate for the business problem.	• The business question and objective are not clear. • No or limited discussion of the background of the business problem. • The target variable is not clearly defined and is irrelevant to the business problem.
Data Understanding & Visualization (20 pts)	• Clear and effective description of data. • Appropriate selection of data for the business problem at hand. • The data collection and preparation procedures (and the sources) were clearly described. • Creative and very effective data visualization that directly guides analysis.	• Clear and effective description of data and data collection procedure. • Appropriate selection of data for the business problem at hand. • The data collection and preparation procedures (and the sources) were clearly described. • Effectively utilized data visualizations to describe the data and guide some analysis.	• The data are not clearly described. • The data are not appropriate to address the business question being analyzed. • The data collection and preparation procedures (and the sources) are not clearly described. • No or very limited attempt to visualize data.
Modeling (15 pts)	• Insightful and thorough analysis of all possible issues in the modeling stage. • The team provided sufficient discussion on the choice of model (e.g., alternative model specifications, the pros and cons of their model). • The application of algorithm and the interpretations of the results were accurate and clearly explained.	• The choice of model is well discussed. • Reasonably complete analysis of the issues. • The model is applied appropriately and the interpretation of the result is accurate. • The team provided sufficient discussion on the choice of model (e.g., alternative model specifications, the pros and cons of their model).	• The team provided insufficient discussion on the choice of model (e.g., alternative model specifications, the pros and cons of their model). • The model is applied inappropriately and the interpretation of results is flawed.

		• The application of algorithm and the interpretations of the results were accurate and clearly explained.	
Evaluation (15 pts)	• The team clearly demonstrated the generalization performance of their model. (I.e., how was the model evaluated?) • The analysis of expected benefit follows logical development and is supported by data.	• The team clearly demonstrated the generalization performance of their model. (I.e., how was the model evaluated?) • The analysis of expected benefit follows logical development and is supported by data.	• No or very limited attempt to evaluate model performance. • Poor flow of reasoning or logic.
Deployment (15 pts)	• Clear and thorough demonstration of the use scenario of the result. • Well-reasoned and thoughtful guidelines and recommendations for deployment. • Recognize obstacles, challenges, and risks. • Thoughtful discussion of the potential issues associated with the proposed plan. • Comprehensive discussion on potential mitigation strategies.	• Reasonably complete demonstration of the use scenario of the result. • Demonstrated recognition of potential issues associated with the proposed plan and provided potential mitigation strategies.	• Superficial, obvious, or inappropriate demonstration of the use scenario of the result. • No or very limited awareness of potential issues. • No or very limited offering of strategies to address issues.
Presentation/Report (20 points)	• Exceptionally well organized and easy-to-follow structure.	• Reasonably well organized and easy-to-follow structure. • The presenter made good use of time.	• Overall lack of organization and structure of the content. • The presentation or report was unstructured and difficult to follow. • The presentation was over time limit.