Review: Final Team Project Instructions

Review: Final Team Project Instructions

Watch this overview video for additional understanding of expectations.

- Audio (08:57)
- Transcript

Review this GitHub repository as an example of an excellent final project.

Introduction:

Two primary skills that data science recruiters seek are the ability to solve business challenges using technology and the ability to communicate technical solutions to executive leadership. This project is an accumulation of skills learned in all seven modules and is specifically designed to reflect all 5 course learning outcomes. You will begin your final project during Module 2, and as you learn new skills and concepts, apply them to your project week over week.

With your team, you will invent a hypothetical company. You will then collect possible data sources (3-5 data files with two having *at least* 10,000 records in the raw dataset) using the provided resources and formulate a challenge your business might be facing, which could be solved using the identified data. You should complete the provided data science design document, which includes plans for data ingestion, exploration, and training. From the design document, you and your team should organize a PowerPoint presentation that indicates the phases of the project, as might be presented to peer data scientists in a project retrospective. The slides should contain links to the code used to build the project, which should be stored in a Git repository. Finally, you and your team will create a video presentation for your hypothetical board of directors assessing the business challenge, justifying your recommendation for action and any nuances the board should consider.

Project Timeline:

- Module 2 (by the end of Week 2): The course instructor will group students into teams of two to three members. Each team will complete the Abstract, Background, Goals, Non-Goals, and Data Sources sections of the Data Science Design Document.
- Module 3 (by the end of Week 3): Each team will complete the Data Exploration, Measuring Impact and Security Checklist, Privacy, and Other Risks sections of the Data Science Design Document.
- Module 4 (by the end of Week 4): Each team will complete the Data Preparation section of the Data Science Design Document and implement code in GitHub.
- Module 5 (by the end of Week 5): Each team will complete the Model Training section of the Data Science Design Document and implement code in GitHub.
- Module 6 (by the end of Week 6): Each team will complete the Future Enhancements section of the Data Science Design Document and implement code in GitHub.

- Module 7 (by the end of Week 7): Each team will submit the following deliverables in the final week:

1.

    1. **Technical Design:** Your completed Data Science Design Document submitted as a **PDF** document containing your implementation plan, including goals, documentation of data sources, a detailed description of how you explored the data, an overview of how you prepared and trained your model, a step by step guide on how you would put your code into production.
    2. **GitHub repository:** Containing your project technical notebook(s) that demonstrates team collaboration. In your final technical design submitted on Blackboard, include a link to your GitHub repository.
    3. **Presentation (PowerPoint/Other):** Present your "Data Story." One PDF or PPT document (at most 10 slides) containing slides prepared for a nontechnical executive audience, with an executive summary of the project, your actionable insights, your business recommendations, a call to action, and the next steps for future work.
    4. **Video:** One 5 to 7 minutes video presentation as an **mp4** by all team members. Your video should address business executives as your audience, using your PowerPoint presentation to tell your data story that includes recommendations and/or next steps. Points will be deducted if the video is under 5 minutes or over 7 minutes.

It is critical to note that no extensions will be given for any of the final projects' due dates for any reason, and final projects submitted after the final due date will not be graded.

Project Datasets:

Choose from one of the data sources below (3-5 data files with two having *at least* 10,000 records) and formulate a challenge your business might be facing, which could be solved using the identified data.

- AWS Open Data
- Kaggle.com
- Data.gov
- Census.gov
- Awesome Public Data Sets

Requirements:

Divide the work equally between the team members for the following steps, and everyone needs to work on at least one portion of the project in SageMaker. You are expected to write high-quality, efficient, and readable code in Python. This project requires that you and your team create a business brief, conduct a video presentation, submit a technical design document with a link to your GitHub repository, and an executive summary presentation using PowerPoint. Your GitHub repository should be clean and professional, your data files should be stored in GitHub or S3, and you should use at least two

AWS services in addition to SageMaker. You must use a SageMaker service for the Training portion of the project, i.e., you can use *sagemaker.sklearn.estimator,* but should not use *sklearn* directly as the goal of this course is to leverage cloud technology for distributed processing.

For the Technical Design Document:

- Include a link to your team's GitHub repository. All of your code should be stored in GitHub in a clean and professional manner. Notebooks should be stored in .ipynb format.
  - Your code should be clean, have useful comments, and only include code that builds towards the project goal.
  - Your data should be stored in S3 and documented in your GitHub repository
  - Any graphics, such as charts/graphs, that help explain your data should be included in your .ipynb files.
  - Your code should be comprehensive and complete with supporting your design document.
  - All team members should contribute to the GitHub repository, and commit history will be available to the instructor for review.
- Include a clearly defined problem statement. This should be one to two paragraphs.
- Include clearly defined goals and non-goals. This should be three to five goals, each one to two sentences.
- Include a clear description on how you will measure the impact of your work; this should tie directly to the goals. This should be one to two paragraphs.
- Complete the security checklist and describe any risks surrounding sensitive data, bias, and ethical concerns.
- Describe your implementation of the solution. Each section should be two to three paragraphs and should describe your findings from the code in your GitHub repository. Your answers should be detailed and explain your rationale for the decisions you have made.
  - Data Sources
  - Data Exploration
  - Data Preparation
  - Data Training
- This section is intended for a technical audience and must be written in a clear, organized fashion.
- You can follow the Data Science Design Document to be sure all sections are met.

For the PowerPoint presentation:

- Should be **no longer** than 10 slides.
- Should include the following:
    - Introduce the problem statement.
    - Describe your implementation at a high level to a non-technical audience.
    - Describe the conclusions of your data analysis to a non-technical audience.
    - Describe your recommendation to the board based on your findings.
- Should be intelligible to a person who does not know predictive modeling techniques.
    - Suppose you are submitting a report to a Director/VP/Executive-level audience who are not familiar with statistical terminology and predictive modeling methods. This can be seen as the executive summary/introduction of your report.
    - You will want to present your analysis as a "Data Story" instead of a technical report. You should articulate your final model results for everyone to understand with clear recommendations and actions.
    - Emphasis is on how you present your findings and recommendations vs. just the content of your slides.

For the video presentation:

- Give a 5 to 7 minutes presentation covering your PowerPoint presentation, and record the Powerpoint and audio as you present.
- Your video should be attached to your final submission an **.mp4**
- Should include the following:
    - Introduce the problem statement.
    - Describe your implementation at a high level to a non-technical audience.
    - Describe the conclusions of your data analysis to a non-technical audience.
    - Describe your recommendation to the board based on your findings.

Project Deliverables and Submission Format:

- Implement the project in Python language using AWS SageMaker. Within your code, for each stage, import all packages used for the project in the first cell, use code cells for code and comments, and use markdown cells for headings and descriptions. Generate a **PDF document** for submission with code, comments, and results within the notebook. **Attach this document as an Appendix to your Data Science Design Document.**
- Prepare and submit your Executive Summary and recommendation slides in **PPT or PDF format**.
- Prepare a recorded video presentation of your project (slides are needed for the business audience) using a screencasting tool, such as Screencast-O-Matic or Zoom, to record your screen

and provide a voice narration. Ensure that the sound quality of your video is good and each member presents an equal portion of the presentation. Export the video file to an **mp4 format**.

  - You may use any recording software you wish. You may want to utilize Screencast-O-Matic, which is integrated with Blackboard and linked below. You can access it using your USDOne account login. View the Recording Video Presentation and Submission Guidelines for MS-ADS Students guide for additional recording instructions.

- Submit the 5 to 7 minutes presentation mp4 video file, technical design document PDF, and executive summary PowerPoint/PDF slides on the final team project submission page of Blackboard. You will use the naming convention **Final Project Report-Team Number.pdf** (e.g., Final Project Report-Team 1.pdf and Executive Summary Presentation-Team 1.pptx). **Only one member of your team will need to submit these deliverables.**

**Plagiarism, or passing another person's code/work/answers off as one's own either by directly copying or even paraphrasing it without proper citation, is a serious offense and can result in sanctions, including grade reductions, course failures, and even expulsion from the university. For more information, please see the USD Code of Honor.

To understand how your work will be assessed, view the assignment rubric.