Course Syllabus for DS 785: Capstone

NOTE: This syllabus document contains the basic information of this course. The most current syllabus is available in the full course.

Course Description

This course describes the premise of the capstone—what it entails, its purpose, and an outline of work required to fulfill the capstone project requirements.

Students are provided with an overview of the capstone course objectives—how to prepare and organize for a semester-long project, the methods used to develop a project, descriptions of project options, and the supporting work that culminates in a final project.

This course provides the information and steps needed to select a topic and a format and then prepare the project proposal that is required in the second week of enrollment.

There are formal assignments within the capstone to keep you and the instructor aware of your progress. Students can contact the instructor if clarification is needed, questions arise, or there is an interest in project topic discussion and refinement.

Prerequisites: DS715 or DS716, DS730, DS740, DS750 or completion of 27 credits.

Course Objectives

By the end of this course, you will be able to:

- Identify and assess the needs of an organization for a data science action.
- Collect and manage data to devise solutions for data science tasks.
- Select, apply, and evaluate models to devise solutions for data science tasks.
- Interpret data science analysis outcomes.
- Effectively communicate data science information effectively in various formats to appropriate audiences.
- Transform findings from data resources into recommended future steps.

Course Components

Discussion

There are several discussion posts that are required assignments within the course. Each of the discussions has a different objective. More details on the content suggested or required for the discussions can be found in the Lesson Module where they are assigned.

- Introduction Discussion: This discussion is an opportunity to introduce yourself to your peers in the course. This discussion is not about your project but rather who

you are and your professional history and/or goals. You are required to reply to at least two of your class peers.

- Project Discussion: This discussion is where you will present your project to the entire class. This discussion will detail the scope/objectives of the project, the specific data science methodology that will be used, and the overall significance of successfully completing the project. You are required to reply to at least two of your peers.

- Peer Discussion and Feedback: There are 4 peer discussions and feedbacks that align with the first 4 presentations. For these discussions you will provide an overview of what the members of your peer group will see in your presentation and discuss challenges or questions that you may have. You will also indicate specific feedback that you would like to receive on the presentation that will assist you in revising your presentation before submission to the instructor for grading. You are required to provide feedback to ALL members of your group.

Presentation Content

This document outlines the purpose and content expectations for each of the five required presentations. These presentations are designed to demonstrate your understanding of the end-to-end lifecycle of a data science project, particularly as it applies to solving real-world problems in a business or industry setting (see Life Cycle of a Data Science Project.). While the suggested content provides a structure for your work, you are encouraged to adapt your presentation to suit the needs of your specific project. Final evaluations will be based on your inclusion of the core components outlined below.

Presentation 1: Problem Definition and Planning

This initial presentation sets the foundation for the project by articulating the business need and defining the scope of your analysis.

Key Components:

- Business Use Case: Clearly describe the business problem or question being addressed. Provide industry context for significance.
- Success Criteria: Specify objectives and define measurable outcomes using KPIs where appropriate.
- Resources, Constraints, and Assumptions: Detail the data sources, tools, time, and constraints.
- Project Plan & Timeline: Outline a realistic schedule with milestones.
- Risk Assessment: Identify challenges and mitigation strategies.

Presentation 2: Data Collection, Cleaning and Preprocessing

This presentation focuses on sourcing, validating, and preparing the data for analysis. Note that each project and data set will have different characteristics. Not all components listed under Data Cleaning and Preprocessing may apply to every dataset. Likewise, there may be datasets that require cleaning and processing steps that have not been listed. Use the bulleted list under Cleaning and Preprocessing as a guide to develop your narrative.

Key Components:

Data Collection

- Data Source Identification: Specify data origin and access methods. Be specific and give URL's of open-source data if used. Discuss and legal issues or limitations on use of the data.
- ETL (extract, transform, load) Process: Describe how data was extracted, transformed, and loaded. Note that some projects will require the use of SQL, web scraping or API calls, other projects may simply be uploading CSV files. There is no requirement of advanced ETL for the project.
- Data Integration: Explain merging of multiple datasets.

Data Cleaning and Preprocessing (after initial data profiling, discuss what is relevant from the remaining bulleted items)

- Initial Data Profiling: Overview of data structure and completeness.
- Missing Data: Address how missing values were handled.
- Formatting and Consistency: Correct data types and formats.
- Outlier Detection: Identify and manage outliers.
- Variable Reduction: Justify removal of unnecessary variables (note: discussion of dimensionality reduction using methods such as PCA should be left for the EDA presentation).
- Scaling and Normalization: Explain any transformations and their purposes.
- Encoding Categorical Data: Discuss encoding strategies.
- Handling Imbalanced Classes: Outline techniques used.
- Data Splitting: Explain dataset partitioning.

Presentation 3: Exploratory Data Analysis (EDA) and Feature Engineering

Understand data structure and engineer features to enhance model performance.

Key components:

- Statistical Analysis: Present descriptive, inferential and correlation statistics and any other statistical measures that were used to understand patterns, distributions, frequencies, and correlations. Discuss the significance of the findings in the context of the model building. For example, if the distributions are skewed, will that affect the efficacy of the models? How was this handled?

- Visualizations: Include graphs, scatterplots, heat maps, charts, and other plots that give insight into trends, patterns, irregularities, outliers, and potential relationships among variables.
- Feature Creation: Describe new variables, how they were developed (e.g. ratios of variables, etc.) and their rationale.
- Dimensionality Reduction: Explain use of PCA or similar techniques for reducing complexity. Visualization of PCA (if used) could be included.
- Business Interpretability: Discuss how features are interpretable and useful for decision-making

Presentation 4: Model Building and Evaluation

Train models, tune parameters, and evaluate their performance. Include visualizations of model results. For example, include a visualization of clusters resulting from application of KNN. Note that the results of modeling must be presented and discussed in this presentation.

Key Components:

- Model Selection: Justify algorithm choices in the context of the business problem/challenge and the data. Identify how this compares to approaches that have been used to address the challenge/problem historically, in research or by similar businesses or industries.
- Baseline Model: Present a simple comparison model if appropriate. If regression techniques are used, a simple base line model might be linear regression.
- Hyperparameter Tuning: Describe optimization methods used.
- Cross-Validation: Explain your validation strategy if CV was used. Discuss how the number of folds (k-folds) was selected.
- Performance Metrics: Discuss the choice of metrics and how they are appropriate to the business context. Discuss the evaluation of the models using the metrics. Note - you must present the results of the mode(s) and the metrics in this presentation!
- Error Analysis: Highlight sources of model error. Examine misclassifications or high-error points to refine understanding
- Model Interpretability: Discuss insights from feature importance determined from models.

Final Presentation: Results, Insights, and Recommendations

Summarize findings and propose business actions.

- Project Overview: Recap the business problem and objectives.
- Data & Methods Recap: Summarize datasets and techniques.
- Visual Summaries: Include visual support for findings.
- Key Insights: Provide a concise summary of analysis results.
- Business Impact: Estimate value or significance/consequences of findings. If project objectives were not met, discuss reasons and the resulting limitations of

the results of the project. Note that not meeting a project objectives does not mean it was a "failed" project. For example, a project that was developed to test new approaches to increase accuracy of current models may result in showing that the current models are the best models. This is a valuable result!

- Next Steps: Discuss potential deployment or operationalization of insights or discuss project extensions (e.g. continued model development, testing of other models, further research).

Project Paper

Students must submit a professionally written paper of their project at the end of the course. Note that except for the Conclusion and References, the sections of the paper align with the five presentations that were required during the course. While the presentations require in-depth discussion of each of the stages of your project, the paper should present the key ideas/findings/results from each section. Inclusion of the full narrative from each presentation will likely result in a paper that exceeds the page limitation. Review the examples of papers in the Resources Module to gain understanding of the depth and breadth for each of the sections of the paper.

Grading

Your mastery of course content is assessed using a variety of methods:

Activity	**Percentage**
Project Idea Submission	10 points
Introduction Discussion	10 points
Capstone Project Proposal & Timeline	30 points
Project Discussion (Full class)	10 points
Group Selection Preference Form	0 points
Presentation 1-4 Discussions @30 points	120 points
Final Presentation Discussion	20 points
Presentations (80, 4@100 points)	480 points
Project Paper Submission	100 points
Code Submission (Pass or Fail Course)	
Self Reflection	10 points
Capstone Website Form	10 points
TOTAL	**800 Points**

Final grades are assigned using the following scale:

90–100%	A
80–89%	B
60–79%	C
0–59%	F