IDC 6940: Capstone in Data Science (1-3 cr)

Course Description

This is a capstone project course using Python, SQL, R, and/or other specialized analysis toolkits to synthesize concepts from data analytics and visualization as applied to industry-relevant projects.

The goal of IDC 6940 is to carry out an industry-relevant **project** in applied Data Science that synthesizes concepts from databases, modeling, analytics, visualization and management of data. Given the professional nature of the MS degree program in Data Science, it is essential that students have experience with analyzing real data sets.

Every capstone project requires a project mentor. The project mentor can assist in identifying, planning and/or executing the data science project. Students will meet periodically with their project mentor(s) to discuss project progress and results, and to troubleshoot. Projects will be implemented in Python, SQL, R, and/or using other specialized analysis toolkits used by Data Scientists.

Projects may involve individual or team effort.

Students will be evaluated by a committee of faculty members and assigned a letter grade. The course will have a coordinator in addition to the mentors/supervisors for individual projects.

The class will meet biweekly to learn from analysis case histories, monitor project progress, have class presentations, and evaluate project progress reports.

Credit Hours

The MS-DS degree requires 3 credits of IDC 6940. Currently, the course has been approved only as a 3-credit course. However, we anticipate that in the near future, the course may be taken for 1-3 credits in any given semester. In other words, the required 3 credits IDC 6940 may be spread out over more than one semester. Students are encouraged to spread out IDC 6940 over more than one semester to enable completion of substantial and meaningful projects.

Learning Outcomes

Students will synthesize concepts from data science, including data analytics and visualization.

Students will learn to identify good data sets and good questions to explore the data.

Students will learn to strategize how to address the goals of the data exploration.

Students will learn to apply the concepts to industry-relevant projects.

Students will learn how to communicate the results via oral presentations and written reports.

Sample Syllabus

The class will meet biweekly to learn from case histories of data analysis and will have invited speakers from the industry.

The class will also be used to monitor project progress, have class presentations, and evaluate project progress reports.

Capstone Process

Students will be provided a list of faculty members who can be faculty mentors for the capstone project in IDC 6940. This will also be provided on the course website.

Students are encouraged to identify an external mentor in addition to their project mentor from FIU. The external mentor may be from the industry and may be more knowledgeable about the project domain. The external mentor may help in identifying good data sets, may help in guiding the student to ask industry-relevant questions and may help in interpreting and evaluating results of the project.

At the end of the project, students will make a 15 to 30 minute oral presentation and submit a detailed written project report, including links to relevant data sets and code (which can be shared via a service such as github). If the students are working in teams, only one joint presentation and report is required.

A committee of three will evaluate the projects. This committee will include the track coordinator, the faculty mentor and the external mentor. If the project does not have an external mentor or if the track coordinator is the faculty mentor, then a third committee member will be invited from the list of approved project mentors.

Sample projects can be found at data analysis challenge websites like http://www.kaggle.com and http://dreamchallenges.org/.

Suggested Timeline

The following is a suggested timeline for students to complete the capstone project. Note: students should plan to complete the capstone course in their final 1 to 2 semesters before graduation.

Step 1: Selecting Mentors (Semester 1)

Step 2: Selecting a Dataset (Semester 1)

Step 3: Planning the Project (Understanding the domain, identifying data analysis questions, identifying analysis tools, writing a proposal) (Semester 1)

Step 4: Pre-Project Review (oral presentation of planned project and incorporating feedback into project) (Semester 1)

Step 5: Project Implementation (Applying analysis tools, preparing initial report, meeting with domain experts for preliminary evaluation of results, interpreting results with help of domain experts, re-analyzing data after discussion and feedback with experts) (Semester 2)

Step 6: Oral Presentation (Semester 2)

Step 7: Final Report Submission (Semester 2)

FIU Engineering & Computing

Project Guidelines

Students are encouraged to find projects from their professional area or from their domain of interest. This is best achieved by talking to domain experts from industry. Faculty mentors may assist in this process.

Projects need to be substantive and meaningful. Data Analysis projects may be designed to test one or more hypotheses (e.g., does factor X cause event Y), or may be exploratory in nature (e.g., what factors may be responsible for event Y). Data analysis projects must explain the choice of approach, tools and visualization. In many cases, different approaches applied to the same data may shed different light on the datasets and it may be reasonable to apply more than one approach. In many cases, different visualization approaches can help highlight different results and conclusions. Where appropriate, statistically sound analyses should be performed. Statistical significance of conclusions should be inferred, where appropriate. Domain-specific interpretations must be made from the results with the help of the mentors. Re-analysis of the data may be necessary after discussion with the domain experts. Sufficient time should be set aside to allow for an iterative process of refining the data analysis and interpretation.

The Mid-point Review will involve a presentation of the proposed data set and the analytical questions that will be pursued in the project. A one-page proposal will be submitted by each project team and will orally defend the proposal in front of the evaluation committee. The committee will examine the proposal for the nature of the project, the tools to be used, and the potential for successful completion, and will provide feedback to the project team. The oral presentation for the mid-point review should explain the tools and methods to be used and the processing for arriving at the conclusions. The final oral presentation should explain the methods used and the conclusions made. The final report must be detailed and comprehensive and written in a form that the work can be reproduced. Supplementary material, including source code, executables and results must also be submitted for evaluation.

The rules for plagiarism will be discussed and provided at the start of the class or on the course website.

Authors

This manual was developed by Profs. Miguel Alonso and Giri Narasimhan in 2018.

FIU Engineering & Computing

Appendix

Project Rubric

Category	Criteria	Meets Criteria	Score 1- 10
Project Definition	Project Overview	Student provides a high-level overview of the project in layman's terms. Background information such as the problem domain, the project origin, and related data sets or input data is given.	
	Problem Statement	The problem which needs to be solved is clearly defined. A strategy for solving the problem, including discussion of the expected solution, has been made.	
	Metrics	Metrics used to measure performance of a model or result are clearly defined. Metrics are justified based on the characteristics of the problem.	
Analysis	Data Exploration	If a dataset is present, features and calculated statistics relevant to the problem have been reported and discussed, along with a sampling of the data. In lieu of a dataset, a thorough description of the input space or input data has been made. Abnormalities or characteristics about the data or input that need to be addressed have been identified.	
	Exploratory Visualization	A visualization has been provided that summarizes or extracts a relevant characteristic or feature about the dataset or input data with thorough discussion. Visual cues are clearly defined.	
	Algorithms and Techniques	Algorithms and techniques used in the project are thoroughly discussed and properly justified based on the characteristics of the problem.	
	Benchmark	Student clearly defines a benchmark result or threshold for comparing performances of solutions obtained.	
Methodology	Data Preprocessing	All preprocessing steps have been clearly documented. Abnormalities or characteristics about the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified.	
	Implementation	The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed.	
	Refinement	The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary.	
Results	Model Evaluation and Validation	The final model's qualities — such as parameters — are evaluated in detail. Some type of analysis is used to validate the robustness of the model's solution.	

	Justification	The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem.	
Conclusion	Free-form Visualization	A visualization has been provided that emphasizes an important quality about the project with thorough discussion. Visual cues are clearly defined.	
	Reflection	Student adequately summarizes the end-to-end problem solution and discusses one or two particular aspects of the project they found interesting or difficult.	
	Improvement	Discussion is made as to how one aspect of the implementation could be improved. Potential solutions resulting from these improvements are considered and compared/contrasted to the current solution.	
Overall Quality	Presentation	Project report follows a well-organized structure and would be readily understood by its intended audience. Each section is written in a clear, concise and specific manner. Few grammatical and spelling mistakes are present. All resources used to complete the project are cited and referenced.	
	Functionality	Code is formatted neatly with comments that effectively explain complex implementations. Output produces similar results and solutions as to those discussed in the project.	

Oral Presentation Rubric (to be developed)