Capstone Portfolio

By Ralph Parlin
rsparlin@syr.edu

"Applied Data Science: A Reflection of Academic Objective Attainment"

A Capstone Portfolio submitted to the Faculty of the iSchool at Syracuse University in partial fulfillment of the requirements for the degree of Master of Applied Data Science

Capstone Advisor
Professor Yang Yang, Ph.D.

April 2021

Capstone Portfolio

Agenda

- A Little About Me
- Methodology
- Program Goals (1-7)
- Goal to Project Mapping
- Concluding Comments

Capstone Portfolio

About Me

"As an Infantry Company Commander in Afghanistan, I experienced first-hand how having the right information at the right time can be the difference between saving and costing lives."

- Army Infantry Officer from 2005 to 2017
- 2017 (to present) made a professional pivot and become an Operations Research Systems Analyst (ORSA) Army Officer
- Operations Research Systems Analysts (ORSA) introduce quantitative and qualitative analysis to military research and decision-making processes by developing and applying probability models, statistical inference, simulations, and optimization models.
- The rapidly changing information landscape has expanded the traditional role of ORSAs, requiring the adoption of new tools and techniques like machine learning and artificial intelligence.
- Selected by the Army to enroll in the Master's in Applied Data Science program at Syracuse University.
- My academic goal has been to master the skills needed to bridge traditional ORSA techniques with evolving data science practices.
- B.S. Marketing with a minor in Economics (SUNY Oswego) – 2005
- Master's in Policy Management (Georgetown University) – 2015

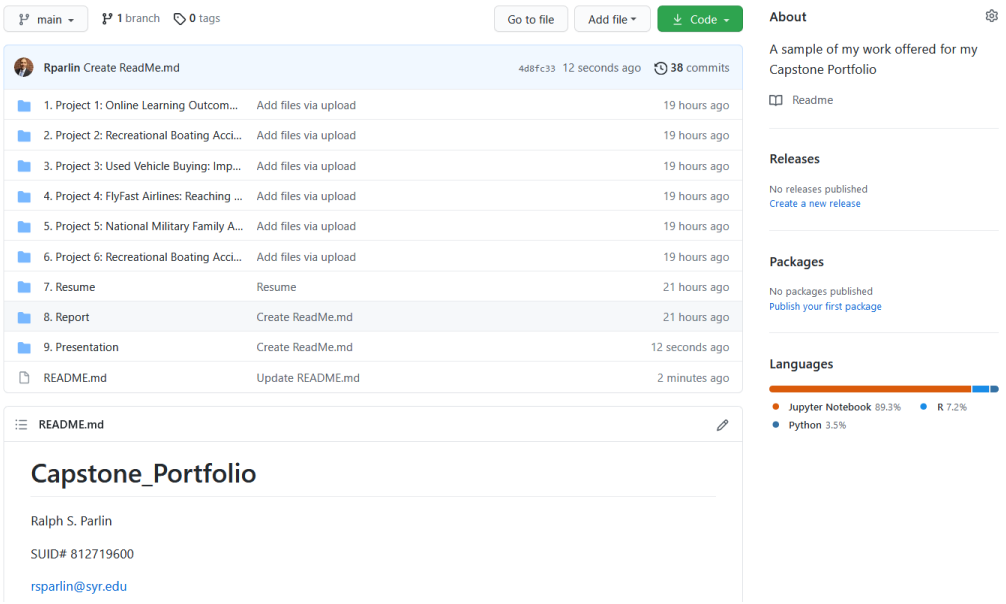
Capstone Portfolio

Methodology

"Applied Data Science: A Reflection of Academic Objective Attainment"

- This paper is structured so that each program goal within the Applied Data Science curriculum is briefly discussed in a way that provides the reader an understanding of why each program goal is important and uses projects I have completed during my time at Syracuse University to demonstrate how it was applied.

- A brief overview of each project is offered to set the stage in the discussion of relevant points of each program goal. The complete details of each project, to included course title, professor, data files, source code, and reports, can be accessed through my Github repository located here: (https://github.com/Rparlin/Capstone_Portfolio)



Capstone Portfolio

Program Goals

- **Goal 1:** Describe a broad overview of the major practice areas in data science
- **Goal 2:** Collect and organize data
- **Goal 3:** Identify patterns in data via visualization, statistical analysis and data mining
- **Goal 4:** Develop alternative strategies based on the data
- **Goal 5:** Develop a plan of action to implement the business decisions derived from the analysis
- **Goal 6:** Demonstrate communication skills regarding data and its analysis for managers, IT Professionals, programmers, statisticians, and other relevant stakeholders/professionals in their organization
- **Goal 7:** Synthesize the ethical dimensions for data science practice (e.g. privacy)

Capstone Portfolio

Goal 1: Describe a broad overview of the major practice areas in data science

The Practice of Data Science in the Wild

- With more devices, connected to more people, more often, it's clear that the amount of data we collect, store and process will continue to grow
- In its simplest form, data science is the science (and a little art) of decision making
- Through a wide collection of tools founded primarily in mathematics, a data scientist applies descriptive, predictive, and prescriptive techniques that bring empirical evidence and scientific methods to decision-making processes
- The variety of my projects is evidence of this:
  - A range of industry sectors, organization types, and organization goals.
  - Applications ranging from simple descriptive statistics to complex machine learning techniques

Capstone Portfolio

Goal 2: Collect and organize data

Gather the Data, Leverage the Tools: Data Collection and Organization

- Data Acquisition
  - Acquiring data to perform required analysis ranges from the ease of being provided a dataset, to the challenge of retrieving the data from Enterprise Resource Planning (ERP) systems, survey data, web scraping or databases.

- Preparing Data for Analysis
  - Lay a foundation of work to ensure success in later analysis and modeling efforts.
  - It is estimated that as much as 70% of a data scientists time is spent in this portion of any data analysis project.
  - Includes tasks such as treatment for missing values, addressing duplicates and outliers, feature engineering, etc.

Capstone Portfolio

Goal 2: Collect and organize data

Gather the Data, Leverage the Tools: Data Collection and Organization

- Data Acquisition

Project 1 - MAS 766 Linear Statistical Model I; Raja Velu, Ph. D

Project Title: “Online Learning Outcomes: Impacts of Demographic, Socioeconomic, and Student Behavior”

Task: Research a topic of choice and conducting analysis using techniques covered in class culminating in a written report and presentation to top management or researcher

Purpose: As a member of a team, employ the linear statistical modeling techniques studied in the course

Method: Linear statistical modeling using R, Python and Excel

Insights: Our analysis provides evidence that:

- A student’s demographics, their socioeconomic status and their academic behavior play a role in determining online academic outcomes
- Students from a wealthier regions have advantages, and how access to broadband matters
- Number of courses a student takes is a leading indicator in helping to predict if a student will withdraw from a program before completing

Capstone Portfolio

Goal 2: Collect and organize data

Gather the Data, Leverage the Tools: Data Collection and Organization

- Preparing Data for Analysis

Project 2 – IST 652 Scripting for Data Analysis; Ying Lin, Ph. D

Project Title: "Recreational Boating Accidents: Causes, Insights, and Ways to Improve Boater Safety"

Task: Conduct an individual project programed using Python

Purpose: Apply the data analytics and machine learning knowledge learned from the course to solve a real-world problem set.

Method: Data Analysis using Python

Insights: My analysis provides evidence that:

- Nearly half of all boating accidents reported to the Coast Guard result in serious injury; 12% result in death
- Rough seas, poor viability and human error are leading causes of catastrophic outcomes
- Operators can improve outcomes if they pay attention, properly assign lookouts to help keep watch for other boats and obstacles, or elect to stay at the dock when weather and visibility are degraded.

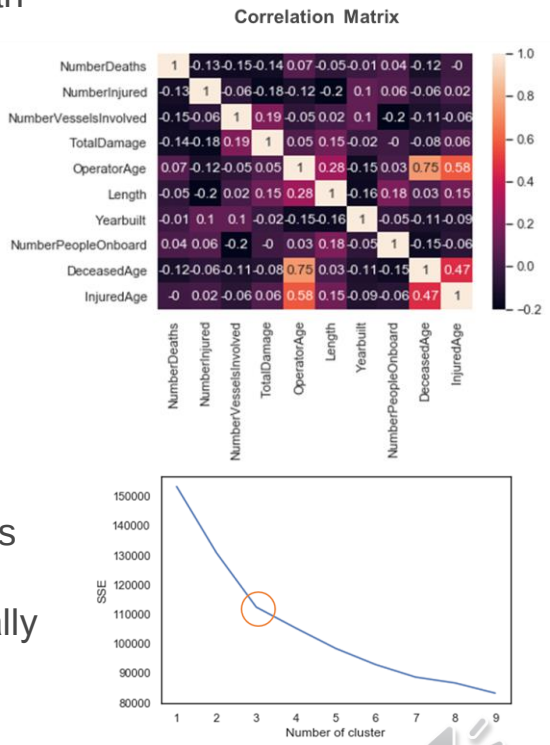
Capstone Portfolio

Goal 3: Identify patterns in data via visualization, statistical analysis and data mining

Seeing the Data: The Science and Art

Project 2 Again: Exploratory data analysis (EDA) with three distinct phases: Data Visualization, Summary Statistics, and Cluster Analysis

- Data Visualization
  - Through tools like histograms, bar charts, and scatterplots, so much can be learned about the data though simple visualizations
- Summary Statistics
  - Summary tables that provide “8 number summaries” offer a quick and clear understanding of the data.
  - The review of the data and the underlying relationships is key to informing the modeling effort
- Cluster Analysis
  - Although technically a modeling effort rather than EDA, cluster analysis is another very useful way to identify patterns in the data.
  - Reviewing features of interest with respect to their assigned cluster tells us about how the data are grouped together. I often refer to these as “personas” and this project yielded three clusters, or personas, that really helped tell the story of interesting patters in the data.



Capstone Portfolio

Goal 4: Develop alternative strategies based on the data

When Data Speaks, the Analysis Strategy Must Listen

- Although driven by the business or research question, the data itself also shapes the analysis strategy
- What's learned during EDA may modify the scope or strategy, or refine the business or research question
- It may also help point toward different modeling strategies to use (inference, prediction, classification, etc.)

Capstone Portfolio

Goal 4: Develop alternative strategies based on the data

When Data Speaks, the Analysis Strategy Must Listen

Project 3 – IST 718 Big Data Analytics; Willard Williamson

Project Title: "Used Vehicle Buying: Improving the Buyer and Seller experience through inference and prediction models"

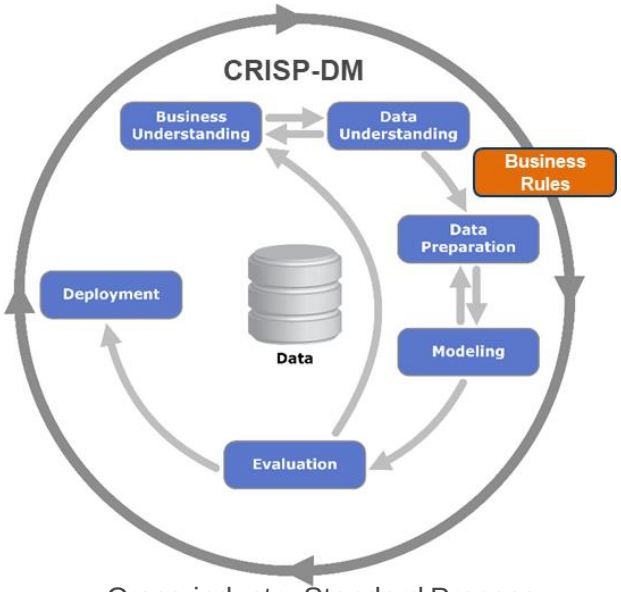
Task: Research a topic of choice and conduct analysis using techniques covered in class culminating in a written report and presentation

Purpose: As a member of a team, employ the machine learning concepts covered in the course

Method: Data Analysis using Python and Spark distributed processing

Insights: Our analysis provides evidence that:

- With respect to predict price, a vehicle's horsepower, milage, and engine displacement are the leading vehicle attributes that predict price
- Predicting price is best achieved through our Random Forest model which rendered our lowest MSE with features that explain over 90% of the variation in price
- With respect to predicting if a vehicle was ever part of a commercial fleet, the vehicle's model year, the vehicle's mileage, and the vehicle's owner count are the leading attributes that predict if it was part of a fleet



Cross-industry Standard Process for Data Mining (CRISP-DM)

Image Source: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Capstone Portfolio

Goal 5: Develop a plan of action to implement the business decisions derived from the analysis

From Analysis to Action: Implementation to drive organizational decisions

- Work of the data scientist must be actionable and executable
- Turning insights into action requires the data scientist to fully understand the domain in which it is applied
- Developing action plans that drive decisions in organizations

Capstone Portfolio

Goal 5: Develop a plan of action to implement the business decisions derived from the analysis

From Analysis to Action: Implementation to drive organizational decisions

Project 4 – IST 707 Data Analytics; Steve Wallace

Project Title: "FlyFast Airlines: Reaching New Heights In Customer Satisfaction"

Task: Identify a real-world problem and develop an analysis report and briefing complete with recommended actions based on findings

Purpose: As a member of a team, apply analytics and machine learning concepts covered in the course

Method: Data Analysis using Orange, R, and Weka

Insights: Our analysis provides evidence that:

- Generally, the air travel market is split evenly across NPS status with Promoter taking the majority stake at 35.73%, Passives at 34.24%, and Detractors at 30.03%
- Adult and Senior Female travelers are a high-payoff market segment and with their needs were met, they could radically improve FlyFast's Promoter numbers
- Prioritize marketing and testing resources on moving Detractors to Promoters before trying to move Passives to Promoters
- Offer free Airline Status upgrades to Seniors even if this means increasing costs slightly as Senior are only moderately price sensitive.

Capstone Portfolio

Goal 6: Demonstrate communication skills regarding data and its analysis for managers, IT Professionals, programmers, statisticians, and other relevant stakeholders/professionals in their organization

Communication: The Leading Differentiator Between Success and Failure

- Communication is the glue that holds all data science projects together—from start to finish.
- A clear understanding of the problem requires effective communication. Absent this, one may spend countless hours trying to solve the wrong problem.
- Communication during the duration of the project is also essential as requirements may change, or the data is presenting unseen obstacles or opportunities not previously known.
- All the good data science in the world means nothing if the insights learned can't be communicated effectively to stakeholders.

Capstone Portfolio

Goal 6: Demonstrate communication skills regarding data and its analysis for managers, IT Professionals, programmers, statisticians, and other relevant stakeholders/professionals in their organization

Communication: The Leading Differentiator Between Success and Failure

Project 5 – IST 659 Data Administration Concepts and Database Management; Hernando A. Hoyas

Project Title: "National Military Family Association Database Management System"

Task: Identify a data management problem in an organization and propose a solution to solve the problem using database technology.

Purpose: Exercise the database administration skills learned in the course and build a functional database using SQL

Method: Through a series of five deliverables (Proposal, Database Design Report, Database Implementation Report, Database Demonstration, and Bug Report) use SQL and the techniques learned in class to provide a Database Management Solution to our client

Insights: Our Database Management Solution allowed the marketing manager at the National Military Family Association to:

- Increase the effectiveness of marketing activity through relevant marketing communications
- Provide better management of marketing programs and efforts
- Effectively measure the impact of marketing to subscribers, donors and members

Capstone Portfolio

Goal 7: Synthesize the ethical dimensions for data science practice

Do no harm: Managing data sensitivity and preventing bias

- Like good communication weaving its way through all aspects of a data science endeavor, so too are the ethical dimensions
  - Sensitive information
  - Proprietary information
  - Personal identifiable information
- Ethical considerations and how they impact modeling
  - Model bias – towards a people or an outcome
  - Are they perpetuating a social phenomenon based on historic precedence embedded in historic data?
- These are the types of questions we must asks ourselves to ensure we are maintain ethical standards in our work

Capstone Portfolio

Concluding Comments

- Headed to Austin Texas and join a team of data scientists working in the Decision and Data Science Directorate at Army Futures Command
- My work will focus on accelerating the delivery and adoption of machine learning and artificial intelligence into enterprise process
- I feel what I have learned during my time at Syracuse University studying Applied Data Science was a leading factor in being selected for this job

Thank you

Ralph Parlin
rsparlin@syr.edu