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

By
Ralph Parlin

Capstone Advisor
Professor Yang Yang, Ph.D.

Syracuse New York
April 20th 2021

Table of Contents

List of Projects Referenced

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

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.

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 will remain listed on the market for more than 60 days, having a franchise dealer involved, whether the vehicle is listed as a cab or not, the vehicle's type of body, and price are the leading attributes that predict if a vehicle will remain listed that amount of time on the market
- 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

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 FastFly'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. Adding a "Senior Aline Status" membership where senior travelers receive unique benefits (e-readers, books, early boarding, closer seating to the front, reserved overhead storage in front of planes, free drinks, etc.) would help improve FastFly's reputation amongst Senior travelers and better align their NPS status proportions to levels amongst our competitors

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 SOllution 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

Project 6 – IST 719 Information Visualization; Frank Marullo

Project Title: Recreational Boating Accidents: Prevention Through Awareness

Task: Create a high quality visualization poster that summarizes a semester long project that requires: 1) Picking and preparing a dataset with R; 2) Defining an audience and requirements for the visualizations; 3) Experimenting with different visual encoding and graphic design choices, 4) Executing the visualization, and 5) Presenting the final product.

Purpose: Exercise skills learned in the course with an emphasis on R programming and Adobe Illustrator

Method: Presentation and Poster

Insights: Our analysis provides evidence that:

- Generally, boating accidents, regardless of outcome, most often result in a collision and are primarily caused by operator inattention or improper lookout.
- For accidents involving injury, most often an individual experiences a cut or broken bone.
- For accidents involving death, the leading cause of the accident is alcohol, which most often results in death from drowning.
- In summary, the best way to boat safely is to pay attention, assign a lookout, wear a life vest, and to drink alcohol responsibly.

1. Lead-in

I recently overheard my wife trying to explain what it is I do for work. As I heard the question asked, I couldn't help but turn my attention to her conversation. As I listened, I became sadden to learn she really has no idea what it is I do. She knows I've been practicing data science for a while, and now I'm in graduate school studying data science, but she really has no idea what "data science" is. Although this breaks my heart slightly, and certainly forced me to have a heartfelt conversation with her about it, it helps bring this portfolio into perspective. My wife is a Senior Marketing Director for a global machine tool manufacturing company—she should know better. But the fact is, many people—including business professionals—lack a good understanding of what data science is, and how it can help them accomplish their organization's objectives.

2. Introduction

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. At the strategic level in the immediate Office of the Secretary of Defense, my team's ability to synthesize critically important information from across the Department directly impacted policy decisions and national security.

At every level I served, my professional witness to the transformative power of timely and relevant information compelled me to want to lead change in this arena within the Army by improving our ability to better integrate data technology into enterprise process and warfighting. To do so required me to make a professional pivot and become an Operations Research Systems Analyst (ORSA), the data scientist of the Army responsible for the curation of information.

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. My professional experience and academic competencies have allowed me to make an initial positive impact in the ORSA community. However, 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. I was selected by the Army to enroll in the Master's in Applied Data Science program at Syracuse University. My academic goal here on campus has been to master the skills needed to bridge traditional ORSA techniques with evolving data science practices.

This capstone portfolio is a small summary of what I have learned at Syracuse University during my study of applied data science. It reflects my current understanding of the field garnered from my professional and academic experience. This paper uses several examples of my academic work explored through the lens of the program goals that aims to demonstrate my understanding and position of several data science topics.

3. Methodology

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 .

4. "The Practice of Data Science in the Wild"

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

According to some researchers, it is estimate that our digital universe is growing by 40% a year. While there is some debate on the true rates of growth, one fact can be agreed on: 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. The conditions to apply data-driven techniques to help organizations achieve their goals has never been better.

The practice areas of data science are limited only by the imagination. In its simplest form, data science is the science (and a little art) of decision making. Have a decision to make? A data scientist can help. 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.

While studying at Syracuse I have had an opportunity to work on a collection of practice areas of data science, with respect to industry sectors, organization types, and organization goals. I have also been afforded the opportunity to practice the many broad ways to apply data science ranging from simple descriptive statistics to complex machine learning techniques.

5. "Gather the Data, Leverage the Tools: Data collection and organization"

Goal 2: Collect and organize data

All data science endeavors begin with a clear understanding of the business or research questions that must be answered. This process is informed by working with relevant stakeholders to develop the business and research questions. Armed with these questions, a data scientist can then develop a data collection plan that allows them to gather the data needed to answer or offer insights that inform business or research questions.

<u>Data Acquisition</u>

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. My studies have offered me the opportunity to explore many of these techniques but the most challenging data acquisitions involved migrating data out of databases and into more usable analysis formats. An example of this can be seen in the work my team did during our Linear Statistical Modeling - MAS 766 course (Project 1).

This project was titled "Online Learning Outcomes: Impacts of Demographic, Socioeconomic, and Student Behavior". During this course, the world was undergoing the largest online learning academic endeavor ever recorded. Around the globe, students and teachers from kindergarten through college, with little warning, were forced to migrated their classrooms from physical to virtual environments. Our research explored the impacts on learning as a result of this massive on-line migration. Through regression analysis, the aim of research was to determine how demographic, regional socioeconomic and academic behavior factors impact a student's online performance. Specifically, we wanted to know how these factors impact two outcomes: one, a student's average academic numeric score across all classes; and second, how these factors impact the likelihood a student will withdraw from online education.

To answer these questions we identified a database that contained anonymized data about courses, students and their interactions with a Virtual Learning Environment (VLE) for seven courses. The database is open to the public and is called the Open University Learning Analytics Dataset (OULAD). The OULAD Dataset contains information on students, courses, assessments, registration,

and interactions with the online learning environment for 32,594 student records and 173,913 academic assessments across 22 courses.

The challenge was that data was organized in 3rd normal form since it was housed in database. As such, we needed to map out a data schema that allowed us to migrate the data features and observations required for our analysis. Armed with our plan, we extracted the necessary database tables and records and moved them into second normal form to develop the datasets needed to conduct this analysis.

Preparing Data for Analysis

Once a data scientist has acquired their data, it's time to get cleaning. This is where the data scientists lay a foundation of work that helps ensure their 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.

Data cleaning and the software we use to perform these functions is a critical skill I have honed over the course of this program. My ability to programmatically prepare data for analysis has allowed me to create custom functions, routines, and processes that make the data cleaning period both fun and efficient.

The data cleaning process is typically described as data munging, data wrangling, or simply data cleaning. This phase of analysis is not only required for proper modeling and statistical analysis, but this phase is where the data scientist becomes familiar with the data. Although there are a host of activities that can occur during the cleaning process, a lot of effort is spent on addressing missing values, outliers, duplicate records, and feature engineering. Each of these cleaning efforts are important for different reasons and the decisions one applies to each is usually driven by the business questions or the data itself. Although data cleaning was a major part of nearly every project I have participated in, a specific example that helps illustrate my competency in preparing data properly can be seen by the work I did in my Scripting for Data Analysis - IST 652 class with a project titled "Recreational Boating Accidents: Causes, Insights, and Ways to Improve Boater Safety" (Project 2).

The motivation for this project came from my love of boating. As an avid boater and lover of data, I could think of no better way to center this project than on the boating industry. The annual economic impact of recreational boating averages around $170.3 billion a year. It also supports more than 691,149 workers and 35,277 business. It is estimated that 141.6 million Americans go boating each year. However, being on the water, although fun, is also dangerous. In 2018 there were over 4,415 boating accidents on US waters. These accidents resulted in $45.9 million worth of damages, and more importantly, lead to 633 deaths and 2,511 serious injuries.

What are the leading causes of boating accidents? How do certain maritime factors interact to explain the variation of accident outcome: injury or death. What can we learn from modeling this data that may help the Coast Guard teach prevention methods, or be better prepared when responding to these unfortunate incidents? These are just a few of the questions my analysis attempted to shed light on.

Through data visualization, statistical analysis, and clustering analysis my report discovered patterns in the data that provided insight and helped inform classification efforts. Using regression analysis, the analysis determined which factors are significant when trying to predict the financial cost of boating accidents. Using decision tree classification, this analysis modeled whether the outcome of an accident results in death.

The Data

The US Coast Guard is the authoritative data source for boating incidents on US waterways and they collects very specific details on every boating accident they are called to. Day, time, location, weather conditions, as well as many other features were used to model specific outcomes. To develop a single datafile that included information from each of the four data tables, using MS Access I had to move the

tables to second normal form by creating SQL queries to merge the data. After joining the tables I was left with 11,720 observations across 35 columns of data. I exported the table as a MS Excel file for upload to Python for wrangling and analysis.

Data Cleaning: Treatment for missing Values

The data cleaning phase of this analysis revealed the dataset had a total of 85,183 missing values. There were a total of 11,534 records with missing data and 13 columns with missing data. However, this is to be expected after moving the database to second normal form. For example, not all accidents resulted in injury or death so for each of those observations, a great deal of data was missing. I began by addressing the columns that are unique to all accidents: OperatorAge, Length, Yearbuilt, NumberPeopleOnboard.

For OperatorAge there were 2,118 records (approximately 18% of the data) missing data. A distribution plot shows the distribution is roughly normal so I decided to use the mean age of 44 to replace the NA values. For Length there were 588 records (approximately 5%) missing data. A distribution plot of this feature shows a right skewed distribution so I used the median of 20 feet for replacement. For Yearbuilt there were 1,189 records (approximately 10%) missing data. A distribution plot of this feature shows a left skewed distribution so I used the median of 2004 for replacement. Lastly, for NumberPeopleOnboard there were 472 records (approximately 4%) missing data. A distribution plot of this feature shows a right skewed distribution so I used the median of 2 people for replacement.

For categorical values I recorded NAs as “Unknown” rather than interpolate from the mode. This decision is based on the business rules used by the Coast Guard where if the reporting does not provide a value for a given field, they typically record it as “Unknown”. Many of these categorical features also required additional cleaning for consistency. For example, the female gender was recorded as both “f” and “F”, so cleaning was required to make the values consistent.

For the features missing data that are conditional on the circumstances of the accident I filtered for “Death” and “Injured”, check for missing values, and replaced the missing values with mean and median estimates where appropriate.

Data Cleaning: Duplicates and Outliers

An inspection for duplicates resulted in the finding of 94 duplicate records. The duplicate records were dropped from the dataframe resulting in a new shape of 11,626 rows by 35 columns. Next I needed to check for and address the numeric columns for potential outliers. The mitigating actions I took varied by feature, but generally I employed a IQR fence technique, or some sort of median/mean interpolation depending on the shapes of the distribution.

At the beginning of this academic program, the data cleaning process was challenging. However, the more classes I took and the more repetitions I got, the awareness of the different techniques used during this activity increased. Additionally, and perhaps more importantly, the trip wires and trade-offs one must consider when dealing with some of the cleaning tasks, such as dealing with missing data, became much more intuitive.

6. “Seeing the Data: The Science and Art”

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

Patterns in data are very telling and often provide great insight toward answering business and research questions. In fact, in the early stages of any data mining endeavor, looking for patterns in the data is arguably among the most critical first steps. The most common times to look for pattens in the data are during exploratory data analysis (EDA) and post hoc analysis, and the best ways to discover these patterns is through data visualization and statistical analysis.

Just like “a picture is worth a thousands words”, so to is a good visual representation of data. With a datafile with millions of observations across many features, very little coding can rapidly create a few quick visualizations packed with information. What is the distribution of the data? How do the frequencies of certain values compare? Through tools like histograms, bar charts, and scatterplots, so much can be learned about the data though simple visualizations. Even an effective line chart can help illustrate the stationary or non-stationary aspects of the data showing the existence or lack of seasonality, trend, etc.

Another powerful, and simple way to get a better understanding of the data is to use statistical analysis. Statistical analysis clearly extends well beyond identifying patterns in the data by providing robust hypothesis testing techniques, but even simple summary tables and statistics about features can say much about the data and inform the subsequent strategies and model efforts. Reviewing measures of center such as mean, median, mode, standard deviation tell us much about the dispersion of the data. Looking at the range of values across given features also helps inform decisions about if and how to scale values. Even simple statistical procedures such as covariance and correlation inform an understanding of patterns in the data.

Another useful technique that helps show patterns in the data is modeling with Cluster analysis. Clustering is perhaps best described using the popular adage “birds of a feather flock together.” Humans naturally like to group things together based on shared values, attributes, etc. However, doing this with datasets containing vast amounts of features and records is nearly impossible as it is often too difficult to see the underlying patterns and relationships that would group data instances together. But though the help of machine learning and algorithms like K-means, this grouping, or clustering, can be done near instantaneously.

An example of visualization, statistical analysis, and clustering pattern identification techniques can be seen by revisiting my Project 2. After inspecting, wrangling, and cleaning the dataset, the next step I took was to search for patterns in the data using exploratory data analysis (EDA) techniques. The exploratory data phase consisted of three distinct activities. The first was to create visualizations for both numeric and categorical features, the second was to provide summary statistics for each of the features in the numeric dataset, and the third was to examine potential clustering. Each of these actives helped better understand the features, their relationships, and provided intuition during the modeling phase.

Data Visualization

Through a collection of data visualizations I learned that collisions with other vessels is the leading accident event. To better understand why boats are colliding I first explored visibility conditions and learned most often accidents occur when the weather is good. However, if visualized in terms of damages, we see a different story with respect to damages, as poor visibility plays a big role. These and the many other visualization efforts really helped me better understand the patterns in the data.

Summary Statistics

Examining numeric features is another very informative step. In this project I created a collection of summary tables that provide “8 number summaries” offering the count, mean, standard deviation, minimum, 25% quantile, median, 75% quantile, and max values of the data. Although not as visually appealing as the data visualizations, these summary charts provide a quick and clear understanding of the data. Sub-setting and filtering the statistics in different ways is also very revealing of patterns in the data. For this project there did not appear to be any major correlations between the numeric variables. These sorts of discoveries are good to know prior to regression modeling as highly correlated features often lead to a regression model having high multi-collinearity.

Cluster Analysis

Although technically a modeling effort rather than EDA, cluster analysis is another very useful way to identify patterns in the data. The objective of the cluster analysis in this project was to increase familiarization with the data, and to uncover potential unseen relationships in the data that may not have been recognized during EDA. For this cluster analysis I used K-means as part of the sklearn.cluster submodule. After conducting feature selection, I normalized the data using standardize Z transformation. This is an important step as clustering analysis is grounded in distance formulas and having a wide range of values between numeric features in the dataset may result in some features having more weight than they should. After normalizing, dummy variables were added to the dataset because I was curious on how these features impact the clustering of the data.

Next I wanted to identify how many clusters to use in the clustering analysis. The technique I used was to initialize the analysis by iterating over several combinations of cluster (1-10) and compare the MSEs in a plot. I used a plot to identify a good "elbow" in the curve which appeared at the 3 cluster mark. As such, I decided to run my cluster analysis with K = 3. I fit a new cluster analysis model with K=3, and merged the cluster assignment of each observation back to the dataframe. As a function of unsupervised learning, the labels for our data and clusters are unknown. However, taking a look at a few of the features of interest with respect to their assigned cluster will tell us a little bit 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.

7. "When Data Speaks, the Analysis Strategy Must Listen"

Goal 4: Develop alternative strategies based on the data

The analysis strategy is initially shaped from the business or research questions. Trying to answer those questions are the motivation for the endeavor. However, the data itself also shapes the strategies one can employ. Tying into the previous section, the patterns discovered during EDA help inform the strategy pursued for analysis. For example, the detection of outliers may cause the analysis or the business and research questions to be re-scoped. Or what is learned may help point toward different modeling strategies to use, e.g. inference, prediction, or classification techniques.

An example of how strategies can change, emerge, or develop from what is learned from the data can be seen in the work my team did during our Big Data Analytics - IST 718 (Project 3) where our EDA brought to light interesting insights which caused us to amend the business questions which in turn forced us to develop alternate modeling strategies based on the data. This project was titled "Used Vehicle Buying: Improving the Buyer and Seller experience through inference and prediction models".

Buying used vehicles can be challenging. Buyers are unaware of what sellers will take as a reasonable offer for their vehicle. Sellers, to include dealers, are often unaware of what price they call sell their vehicle for. This tension on price is muddied further by factors such as vehicle features, geographic region, time of year the sale is offered and alike. Our intuition combined with what we saw in the data helped us arrive at the following problem statement: "Current processes for determining the price of a used vehicle creates pain points for both U.S. vehicle buyers and sellers as they struggle to determine a vehicle's fair market value. However, the valuation process can be improved through effective modeling." To address the problem statement, we developed several research questions that allowed us to group analysis efforts along three primary characteristics of buying and selling vehicles: the price, how long a vehicle has sat on the market, and how the vehicle was used during its lifetime.

Our goal was to predict price, whether a vehicle would remain on the market more than 60 days, and predict vehicle history with respect to fleet use. With respect to inference, our goal was to determine which vehicle attributes matter most with respect to price, which vehicle attributes are most important in ensuring used vehicles don't sit on the market, and which vehicle attributes are most important with

respect to a vehicle's fleet class. To answer our research questions and meet our prediction and inference goals, models were created using a collection of machine learners including Regression, Logistic Regression and Random Forest.

The EDA phase of this this analysis was critical in scoping the project and it greatly informed our analysis. For example, the original dataset included used vehicles with prices well beyond $1 million dollars. To provide a more useful model to average consumers, we capped the values of a vehicle at $90,000. Our data exploration also led us to discover that many of the observations included in the data were for vehicles categorized as "new". These and other discoveries we made from the data greatly impacted our analysis and modeling strategy.

8. **"From Analysis to Action: Implementation to drive organizational decisions"**
   Goal 5 Requirement: Develop a plan of action to implement the business decisions derived from the analysis

To be value added to an organization, the 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 is the desired endstate of my work. An example of this can be seen in the work my team did during our Data Analytics - IST 707 (Project 4). This project was titled "FlyFast Airlines: Reaching New Heights in Customer Satisfaction". The purpose of this analysis was to identify the primary drivers of an airline passenger's likelihood to recommend, gain a better understanding of FlyFast Airway's market position, and offer a broad plan that leverages these findings to improve FlyFast Airway's customer satisfaction.

The data used for this analysis was collected via a survey of airline passengers traveling with FlyFast, Cheapseats, and Sigma airlines from January 2014 to March 2014. The survey recorded 4,985 observations across 23 variables pertaining to traveler demographics, purchase behavior, flight information and most importantly, their likelihood to recommend the airline. Each traveler's likelihood to recommend an airline is provided on a scale from 1 (unlikely) to 10 (very likely). For this analysis, each traveler's likelihood to recommend measure was discretized into "Detractor", "Passive" and "Promoter" statuses. This transformation provided us a discrete measure of traveler satisfaction and served as the cornerstone of this analysis allowing us to align our findings with the industry standard of Net Promoter Score (NPS) ratings.

One of the primarily deliverables from this analysis was providing the client an updated Marketing Plan based on the findings from our analysis and modeling. Using market position analysis, as well as Strength, Weakness, Opportunities, and Threats (SWOT) analysis, our report helped paint for FlyFast airways a clear picture of where they sit with respect to competing airlines, and actions they can take to reinforce their strengths, seize opportunities, and mitigate threats. Based on the collective understanding provided by the analysis and market positions identified in the report, our team was able to produce four actionable recommendations to increase FlyFast's NPS status that were backed by rigorous analysis. I am confident had the scenario not been notional, these recommendations would have had the desired impact that organization was trying to achieve.

9. **"Communication: The Leading Differentiator Between Success and Failure"**
   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 is the glue that holds all data science projects together—from start to finish. Having 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. And lastly, all the good data science in the world means nothing if the insights learned can't be communicated effectively to stakeholders.

An example of my proficiency with communication is evidenced in the Database Management System I made with another student in my Data Administration Concepts and Database Management - IST 659 class. For our project we developed a database solution for the Marketing Manager at the National Military Family Association (Project 5).

This project assisted the Marketing department at the National Military Family Association (NMFA) establish a single Database Management System. The National Military Family Association (NMFA) is the leading 501(c)(3) non-profit association serving the families of the currently serving, veteran, retired, wounded or fallen members of the Army, Navy, Marine Corps, Air Force, and Coast Guard.

Building the public's awareness of this non-profit is the primary job of the Marketing department. The Marketing department advances the NMFA brand through a collection of techniques and tactics, and our project was designed to help them with their primary tool: email. The Marketing department is solely responsible for sending all electronic communications for NMFA. As part of this responsibility, the Marketing department must help the Development Department connect with donors so they can be made aware of and contribute to donation campaigns. The Marketing department also assists the Programs department with touting the scholarship and family programs NMFA offers by emailing potential military family candidates.

The problem is that performing the support needed to the Development and Programs department requires the Marketing manager to have access to the audiences each respective department works with. However, under the current structure, each of the departments maintains their own sperate system and process for capturing, storing, and managing data about the people with whom they interact. These disparate systems coupled with no consistency in data integrity presents significant challenges to the Marketing manager when conducting the necessary data merging tasks required to support the Development and Programs departments. Simply trying to consolidate or segment email lists across the departments takes hours of reconciliation when this should be near instantaneous.

The data integration required across the Development, Programs, and Marketing departments requires an extensive Data Lake/Data Warehouse architecture with an extract, transform, load (ETL) capability and it exceeds the scope of this project. Instead, our proposed system focused on addressing the challenges faced by the Marketing department and provided a proof of concept for NMFA to consider as a model for a broader database solution. Our system resulted in the Marketing manager being able to manage targeted messaging for specific people in the database. It allowed them to build records for each marketing campaign they develop, as well as glean insights from marketing and donation campaign trends to assist the Development and Programs departments.

A critical first step in this process was to ensure each department had a clear, and shared understanding of the different audiences each department works with. Furthermore, each department had subscribe to the idea that there needs to be select data consistency across each department system. These understanding were codified in unifying business rules and process we developed and effectively communicated to their team. These business rules were the perquisites needed to provide the Marketing manager a system that allowed her to integrate the Development and Program department data into the more comprehensive database management system.

This project required extensive communication with the project sponsor. On a near weekly basis, my team held teleconference meetings that helped us understand the problem and develop a solution that was achievable in the time we had available. Our communication plan employed techniques we learned from Professor Saltz in our Introduction to Data Science course. We leveraged his lessons to focus on

talking to subject matter experts (SMEs) and had them communicate their issues by telling stories, explain exceptional cases, and by asking questions about risks and uncertainty.

Our rigorous communication plan is what allowed my team to understand the problem and back-brief the project sponsor in a way that provided her confidence we were on the right track. Additionally, by clearing communicating the facts, assumptions, and limitations throughout the project timeline, each party's expectations were managed properly and as a result, the project was highly successful.

Another example of effective communication was the data visualization my classmates and I created in Information Visualization – IST 719. This project used the U.S. Recreational Boating datafile from Project 2 and advanced the analysis with respect to visualization culminating in a poster to help communicate critical findings (Project 6). For this project we used R to clean and prepare the data from the U.S. Recreational Boating datafile and used Adobe Illustrator to produce the final poster file. The poster was built in accordance with best practices we learned from the course and offered a well planned information hierarchy which made our key findings easy to see and understand.

10. "Do no harm: Managing data sensitivity and preventing bias"

Goal 7: Synthesize the ethical dimensions for data science practice (e.g. privacy)

Similar to the importance of good communication weaving its way through all aspects of a data science endeavor, so too are the ethical dimensions. Often times the data we work with are sensitive. The data may include proprietary information that is sensitive to the competitive standing of the organization. The data could include personally identifiable information of people captured in the data file. This was the case with the Coast Guard data requiring us to strip certain features of the data to ensure the analysis could not be traced back to an individual based on Hull Identification Numbers.

Another ethical consideration is the how the data we use can impact modeling. Do our models have bias? 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.

11. Concluding Comments

Upon completion of my degree I will head to Austin Texas and join a team of data scientists working in the Decision and Data Science Directorate at Army Futures Command where my work will focus on accelerating the delivery and adoption of machine learning and artificial intelligence. 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. Each course I have taken has enabled me to be professionally prepared for this opportunity. This report, and the collection of my work featured in its Github repository, is evidence to what I learned and how I have been shaped by this experience.