Michael Morales

September 16, 2019

Portfolio Milestone

***Introduction***

My journey to becoming a data scientist began with a work training presentation covering a tuition reimbursement program. The presenter was explaining how classes needed to be presented individually for reimbursement rather than as a program in whole. She used data science as an example, saying that a program may not be relevant to one’s current career track, but classes in the program could potentially be eligible. The term ‘data scientist’ became a fascination for me. There was something about it that spoke to my analytical mind and made me feel I had discovered something new and largely unknown. I needed to know more.

Weeks of research into the many applications of data science and the jargon associated with it were a preparation for the data gathering process. Those weeks foreshadowed the work of familiarizing myself with an unknown domain in preparation for interpreting prediction model results. Collecting and curating personal data for multiple application processes and awaiting feedback was an exercise in preparation for the many final projects required in the program, as well as for the real-world application of the knowledge gained.

In the course of this report, I will apply the training I have received to validate the expertise I have achieved as a result of my participation in the MS in Applied Data Science program at Syracuse University. It is my hope that I will successfully defend my knowledge and competence by demonstrating how each course and the final project requirements have met or exceeded the overall objectives of the Applied Data Science program.

***Program Objectives and Relevant Projects***

1. ***Describe a broad overview of the major practice areas in DS***

Data science is a multi-disciplinary field with many areas of application. While some data scientists may have a broad focus and apply many aspects in one job, others may find themselves working in one narrow field. Knowledge of the domain in which one is working is key, and the focus will change accordingly. Some may, after learning how to apply machine learning and statistical methods, focus solely on data engineering and warehousing.

In my studies in IST722 – Data Warehousing, I found that I learned a great deal about how data warehousing is used in my job at Schneider Electric. I tell my clients almost daily that they will see updates in our customer-facing websites after an update overnight, a process used in data warehousing to keep from stressing the main database with continuous queries. I also learned in Data Warehousing that schemas are used to make data useful for analysis. Without data warehousing, the work of gathering and making data useful would be much more difficult. The benefit of having a data science background as a data engineer is having the knowledge to organize the data in a way that will make it useful for its purpose, whether it be machine learning, exploratory analysis, or statistical analysis.

Data mining is another important area of data science. In my IST707 and IST736 projects, I learned to use data mining methods to translate business problems into useful data questions and then create prediction models to answer those questions and provide useful insights. These classes were especially useful in teaching me how to tell a story that can be understood by laypeople unfamiliar with data science processes and methods.

Business intelligence and strategy are areas related to data science. I learned to use BI methods in projects I completed for MAR653 – Marketing Analytics, MBC638 – Data Analysis and Decision Making, and SCM651 – Business Analytics. For MBC638, I completed a process improvement project that had a direct impact on my job. I observed the work of an intern and timed his processes, and was able to provide him with insight into how to streamline his workflow and work more efficiently. For MAR653, I completed an analysis of data from the website Board Game Geek, which is a forum for users to rate board games. In this analysis, I found that board games are experiencing a “golden age” and user ratings are useful tools for learning about what makes highly-rated games so popular. Some of the same methods applied in SCM651, where price optimization and Google Ads were the subject of group projects.

Data visualization and story-telling are perhaps the most important aspects of data science. The ability to share findings with laypeople is the catalyst to making data science useful in the world beyond the computer screen. Predictions, analysis, and recommendations are useless if they cannot be communicated to stakeholders up to and including the C-suite. This was stressed most vigorously by Dr. Ami Gates in IST707. Dr. Gates provided useful feedback and shared stories from her career experience emphasizing the importance of good writing and communication.

In my project for IST707, I analyzed World Cup results to try to form a hypothesis about whether it is financially a boon or boondoggle to host the tournament, and whether winning and placing can be correlated to financial benefit. While I found that there was insufficient data to make any meaningful predictions, I felt confident in my ability to share my findings and explain the reasons why the data were too skewed and sparse to be useful. I also found that FIFA does a very good job of applying lovely window dressing to its financials to hide the fact that the tournament tends to leave a path of destruction in its wake. Ultimately, the project was a very useful exercise in learning to do deep research on a domain to know it well enough to provide insight. Data science cannot exist in a vacuum and be beneficial to the world. This project also taught me about operations-focused data analysis. Data science is a very technical field, but it also has room for practitioners who bring an eye for problem-solving and processes to the table, but who may not be as proficient in the technical aspects.

Data science is often sector-specific. Many of the machine learning methods I have learned are used extensively in epidemiology, astronomy, biology, and many other scientific and business fields. In my job, linear and logistic regressions are important tools, along with time-series analysis and correlation models. In the correlation model I used for the Board Game Geek analysis in MAR653, I found that complexity and a game’s popularity were closely related, and was able to support my hypothesis that since popularity was on the rise, complexity was also increasing with time.

1. ***Collect and organize data***

The projects I have completed during my time in the program have consisted of analyzing data available from the classes themselves or as spreadsheets from online sources. Collection of data has involved research of websites such as Kaggle and others where json data are available. In my final project for IST652 – Scripting for Data Analysis, I collected json data from reddit.com. Reddit allows the collection json data without the need for an API by sampling adding the .json extension to any of their sub URLs. These subject-specific subsites are known as subreddits, and are forums where subjects of interest can be discussed, debated, and shared. Though I did not employ this data in the final analysis, I was able to employ the knowledge learned during the course to collect json data.

Organization data is always necessary, as data are not necessarily suited to analysis. For my IST687 – Applied Data Science final project, I was required to remove several columns and rows from a dataset of more than 300,000 lifts from powerlifting meets. The goal of the analysis was to identify relationships between different lifts in the powerlifting repertoire, the relationship between bodyweight and the Wilks coefficient score, and which lifts influence the Wilks score most. While this data went back to 1964, the Wilks score has not been in wide use for the last 55 years. I made the decision to only use data from the last 30 years, as I found in my research that the first serious studies of the Wilks score were not conducted until the mid-nineties. This project also required the use of R to remove columns not of use to the analysis. Columns describing the class of the lifters, the federation sanctioning the meets, whether the lifter was tested for performance-enhancing drugs, and the results of other scoring methods were removed to further pare the data down to a manageable and useful set.

I employed similar methods in my projects for IST652 and IST707.

1. ***Identify patterns in data visualization, statistical analysis, and data mining***

Patterns are one of the keys to useful data science. For my final project in IST736 – Text Mining, I employed the use of Latent Dirichlet Analysis, which measures patterns in word usage to assign weight to each word and model topics based on them. The goal of the project was to help Quora, a question and answer forum, identify insincere, toxic, and abusive questions. I found that LDA was useful in that it provided insight into the kinds o f topics popular among insincere questions. Topics such as race, religion, and politics stood out as bastions of insincerity and toxicity. This pattern-based finding could be useful. Quora could consider using topic modeling to categorize questions and flag them for further analysis.

1. ***Develop alternative strategies based on the data***

The ability to be agile is important in most situations, but it is essential in data science. In my World Cup analysis for IST707, my original intent was to attempt to predict the winner of the next tournament. I found that the data was sparse and heavily skewed. Of the 79 national teams that have competed in the World Cup since its inception in 1930, only 8 teams have ever won. In my project for IST652, I analyzed hiking data from hikr.org. I had intended to include json data from Reddit, but the goal of the project was to see if I could build a model that would predict the difficulty rating of a hike, and Reddit data did not include a consistent difficulty rating.

1. ***Develop a plan of action to implement the business decisions derived from the analyses***

My data warehouse project in IST722 involved the creation of a data dashboard stakeholders could use to increase marketing effectiveness and monitor KPIs across the enterprise. Though the business, Fudgeflix and Fudgecorp, were not real-world organizations, the scenario provided a situation that is often encountered in the real world: a merger. Mergers require stakeholders to find client crossover, to analyze sales performance separately and as an aggregate, and to adjust their offering to cater to a new client base. A key to accomplishing these goals is to look at historical data and get an idea of trends, which my dashboard provided. It also provided a resource to judge which products were most successful in the past and which are currently trending upward.

For MAR653, the Retail Relay exercise involved an analysis of sales data to make a recommendation regarding future marketing efforts. I was able to determine that a subscription service would be an excellent customer retention tool, as the customers with the longest histories were the most likely to be retained. In my final project for Marketing Analytics, I was able to identify the very strong correlation between board game complexity and high ratings, which leads to a recommendation that game creators focus their efforts on creating new titles with higher complexity and less repeatability.

1. ***Demonstrate communications skills regarding the data and its analysis for managers, IT***

***professionals, programmers, statisticians, and other relevant professionals in their organization.***

IST707 was taught by Dr. Ami Gates, who repeatedly stressed the gravity of communicating effectively to people who do not understand data science practices, and who may struggle to understand results. In IST736, Dr. Jeremy Bolton equally stressed the use of visualization and lay language to communicate results. Thanks to their combined efforts, I have been putting my English degree to use in homework and final reports. The use of creative writing is essential in data science as it can help a data scientist to engage the audience and keep them interested in the subject matter. Without the ability to effectively communicate results and recommendations, analyses become useless. Without good grammar, the audience can become distracted. And in the absence of outlined, thesis-driven language, the message can become lost.

In my IST736 final project, I had to explain in detail the kind of discourse Quora wanted to encourage, and contrast it with the toxic queries they want to mitigate. This requires creativity, as facts cannot simply be stated one after the other else a report become a long bulleted list. Models needed to be explained in simple, but detailed phraseology that allowed their complexity to be understood. I had to explain the relationships between words that made up topics in a way that would make sense to someone completely unfamiliar with topic modeling.

1. ***Synthesize the ethical dimensions of data science practice (e.g. privacy)***

Private customer data and customer identities should remain private. As such, the data I worked with in the program was anonymized. Data should be anonymized whenever possible, even if not required by law. In situations where the law requires anonymity, special care needs to be taken to ensure compliance. If it is necessary to analyze non-anonymized data, the dataset should always be kept secured to avoid data leaks. Generally speaking, analysis of data using data science techniques and practices is possible without personal data being included.

In the event that personal information must be included in a dataset for analysis, transparency should be maintained with the people who have allowed their information to be used. They should have input into the way the data is being shared and stored so they are always aware of who has access to it. In my 651 project, where I used data from my job, I was careful to keep any proprietary information out of the assessment and deliverable, and I kept the identity of the intern I sat with private.

In the famous Target example where the father of a college-age woman discovered her pregnancy inadvertently based on advertisements sent to his house, Target data scientists failed to consider the unintended consequences of their recommendations. While they saw their prediction models as offering a useful service to women preparing for the birth of their children, and while this may have been true, they failed to consider that pregnancy can be an intensely personal experience for parents-to-be and mail is not necessarily private in cohabitive situations. Data science can be a powerful tool, and that power needs to be implemented in a way that does not cause harm.

Perhaps the most important ethical principle that practitioners should keep in mind is that machine learning algorithms are trained to make predictions on the information in the data. This means that any biases that have made their way into the data will pass through to the predictions. If the data is discriminatory, the predictions will discriminate. If the data is colored by a sexist bias against women, it will predict misogynistically. If the data includes training samples that suggest to the models that the elderly should be treated with disdain, the models will result in recommendations that disadvantage senior citizens. These are areas where knowledge of the domain is of the utmost importance, as it is in nearly every step of the process of analysis.

***Conclusion***

My time in the applied data science program has been filled with challenges. Learning new ways to analyze information to formulate questions I never thought I would ask has been demanding. Learning new ways of communicating that allows humans and computers to talk to each other has been stimulating, but incredibly difficult. And learning to look for meaningful patterns has been strenuous, but it has provoked the most thought.

Patterns are all around us all the time, but they are not always meaningful. The applied data science program at Syracuse has taught me to seek meaning and insight in patterns, but to always be skeptical of them. Prior to my education in data science, I saw correlations and assumed they were always meaningful, but now I know that relationships can exist where none were intended or nourished.

Most of all, I have learned that research is crucial. An analysis cannot be effective without domain knowledge, just as a master mechanic trained by Ford cannot necessarily apply his knowledge to Porsche race cars. The skills of applied data science are useless when they exist in a black box and cannot be interpreted and communicated in a meaningful way. The result of a prediction model is not simply a spreadsheet. It must be understood, interpreted, and communicated.

