|  |
| --- |
| Master of Science in Applied Data Science |
| Portfolio Summary |
| Syracuse University |

|  |
| --- |
| Scott Snow  6-18-2019 |

The field of data science can be constructed as a ladder. This ladder had a very wide base but gets more, narrow as one goes up. Data scientists can either be professionals who climb and work at varying points on the ladder, or they can work on a very specific part. The top of the ladder is the action that is taken based on everything done previously. The goal of data science is to act based off all the other aspects.

In this domain, taking no action can be also considered an action but that decision is made based on the work that was done. The first step is to gather data. In certain workplaces, this is done manually with a constant input of data from workers and managers. Certain data that eventually is analyzed starts as blanks on a form. For my initial projects in my first term, this was how I acquired my data. I didn’t really know much yet and I didn’t realize both how difficult it is to effectively communicate data collection needs to other parties and how big of an issue missing data could cause in analysis. It was however a great learning experience as I gained valuable tools in SQL and Excel. For the Excel work, I was required to create a storyboard to present my data acquisition and the analysis that I did.

Other forms of data collection come from machine input such as scanners, sensors, and automated cameras. Most recently I used battery data to perform a time series visualization. The data was from a full lifespan test on a battery showing its charge and drain cycles. Outside of that, I have only done basic SQL work with grocery store data.

The most common form of data collection is from the internet. This is data that comes from websites or IoT devices. My final project for Scripting for Data Analysis (IST 652) involved web-scraping multiple Wikipedia pages to put together a list of songs. My most recent project for Visualization involved getting real time data on video Streams from the website Twitch.tv using their API. Other projects

Ultimately, in an education setting, these data collection processes resulted in only subset of the data that I cleaned, analyzed and used for data mining. The remaining data that I used during my education came from websites such as Kaggle.com which included ready to be imported data frames with either numerical data or already cleaned unstructured data that only required a small amount of effort to scrub and transform into structured data.

My first project and homework in Data Analysis and Decision Making (MBC 638) refreshed much of my knowledge on statistical analysis including T-Test, normality, Correlation, and Chi-Squared test for normality. This knowledge was utilized again in preparing my final group project for Introduction to Data Science (IST 687) where we used a T-Test to compare non-finalist teams with a player replaced with the MVP or a finalist player had an aggregate measure for key metrics outside of the confidence interval or inside. Inside 95% we were confident that the given player could have been changed out without a drastic change in overall score. Outside 95% we concluded that that specific player was essential to their team’s finals appearance and that said team would not have played in the finals with an average replacement player.

Overall, much of my statistical work has been done in my business classes while much of my School of Information coursework has been devoted to data cleanliness and learning models. An example of some of my early work can be seen in my project for Data Analysis and Decision Making. For that project, students were individually tasked with developing a business or business-like problem. At the time, I was interested in using new skills to improve efficiency at my current place of employment. For that purpose, I formulated a process improvement problem where we took our current process for responding to floor defects, of which there were many undefined decision points, and attempted to improve them. First the response time was measured with no change, and then definitions for some of the decision points were implemented and the response times were measured for subsequent defects. Due to the current efficiency of the manufacturing process, defects were uncommon. As such, the data set only contains around 13 data points accumulated over a three-month span. Despite this dearth of data, I was able to apply numerous statistical methods and create a variety of visualizations for my final presentation. This experience created a solid foundation on which to build my future learning. I as able to utilize the techniques used in future classes and gain more understanding about how they could have been better used on this project.

My first introduction to information science and structure was in the opening class of Database Administration and Management (IST 659). Previously, I had only worked with class structures within a programming sense such as linked lists. This was much different than that. For my data source, I wanted to see if I could improve the efficiency of the business I had worked at. I was specifically trying to create a database that tracked daily production output and which employees were at which station during each day. My first task was to develop a pipeline to obtain the data. Included in the data collection was the names of my fellow employees. These were the only identifiers that I collected for this project. I let each of them know that only their name would be included in this database and that the code that included their names would potentially only be posted online in one location for including the project in a portfolio. None of my colleagues objected to this. Following, I was required to map out the relationships between each field that I was collecting. This helped me understand how data is collected and organized. After the data was collected and the database was created, I was tasked with creating interface query forms in Access and reports. These displayed statistical measures of the data that was collected which helped the user identify patterns. While the project was never actually used in production, I did create multiple user accounts that gave specific groups within the organization access to only certain parts of the database.

Understanding the data science process took multiple courses. In some early work such as the final project for Intro to Data Science in which my team analyzed NBA player data there was a sense during completion that we were improvising the entire process. Reflecting on this project and considering the knowledge and skills that I have now, I still get that sense. The project did, however allow us the opportunity to practice the skills we had learned and try to be successful in the task we had chosen for ourselves which was to predict Championship and MVP caliber players based on salary and regular season floor production and to predict teams who may be over or underspending on their roster. In addition to attempts at a linear model, we explored many visualizations that showed some players to be overpaid and some teams to be overpaying for their rosters or who are simply not being coached well enough.

This specific term had an interesting juxtaposition of hands off and guided learning. While I felt that Intro to Data Science allowed for independent exploration, the other course I had taken that term, Business Analytics (SCM 651), provided much more guided learning. In the long term, both were useful as I have specific skills and a verbose understanding of the entire analytical process. One assignment that stands out was using the tools available on Google Analytics to look at a marketing campaign that the Whitman School had done for their online MBA program. Much of this work was done in Excel and my team looked at traffic patterns for the average client, most common time of day for access, and volume patterns overtime, and breakdown by keywords used that resulted in clicks. The results of our exploration were documented in an outlined report.

During this term, I also read an article that resonated with me. Previously, I had read about housing and health care using analytics which while intriguing, was not something that surprising to me. The article I read was from the MIT Sloan Management Review titled “Data Analytics in Sports” by Thomas H. Davenport. Up to that point, I hadn’t seen the movie Moneyball. The movie’s story was mentioned in the article and it really invigorated my imagination for the potential for machine learning. Many of the projects I undertake after this point ask increasingly challenging questions.

By the third term, I had become more comfortable with obtaining pre-formatted data from various sources and performing analysis. Data Analytics (IST 707) furthered my knowledge and comfortability with the data science processing. Additionally, it really furthered my knowledge on data mining among other subjects. The final project was once again larger in scope than the skills available at the time. Working with one other person, we still attempted to do what we could but ultimately the result was like the project done in Intro to Data Science, only with better visuals and better constructed and utilized models. Each assignment was also too specific and guided to effectively demonstrate my progress towards the program’s learning directions. Data Analytics mostly functioned in my learning as a bridge from the Intro to the final course, Big Data (IST 718).

Much of what I learned in Data Analytics however, I was putting to use in my concurrent course at the time, Marketing Analytics (MAR 653). The assignments in both courses allowed me to contribute heavily to the term project. For this project, my team located housing data from Ames, Iowa. Much of this data was centered around the housing crisis. Initially we were not sure what approach we should take. After analyzing the data, I noticed a price difference in houses that had been remodeled vs homes that hadn’t been. Using this knowledge, we developed an approach where we created a linear model using the homes that had been remodeled and using features which were using to improve with a remodel such as attached garage, livable basement, attached pool etc. along with the continuous measurables of the house. Then using the linear combination from that model, we predicted prices on homes that hadn’t been remodeled. Comparing those predictions to the actual listing prices we identified 13 houses that our model indicated would be worth more if they were remodeled. The business approach was that we were a real-estate investment company looking to purchase properties, remodel and resell them for profit. The model would hypothetically be used to predict further listings and be reran as more remodeled homes were listed.

This project was a definite turning point where I began to intuitively grasp the process of collecting the data, cleaning, wrangling and analyzing it, developing an approach to reach a desired piece of information and communicating the results. I felt much more confident in my coursework going forward. In the following term I studied Text Mining (IST 736) and Scripting for Data Analysis. This was a great term as I was using Python in both terms have previously only used R. For scripting, the data source I chose was information from Wikipedia and Lyrical text data from the website Genius.com. For both sources, I used web scraping packages to compile a list of songs belonging to the rap/hip-hop genre. First from winners and nominees for Best Rap Song, and then from songs released in the appropriate time frame for this award. The primary focus of the project was collecting the data and performing preliminary analysis. The conclusion was that word and vocabulary complexity seemed to be the biggest factors in determining winners as each successive group had a smaller average ratio of total words to occurrences of each word meaning that there was more use of uncommon words in the songwriting process. This is one of many topics I would am interested in revisiting post-graduation.

While my course work in Scripting was primarily focused on data acquisition and cleaning, Text Mining had similar elements but also place great emphasis on patterns and communication skills. I feel that the Instruction in this course and Big Data have most shaped my view and approach as a data science professional. An example of this can be seen in the final homework assignment for the term. In this assignment we were tasked with using Latent Dirichlet allocation via the software Mallet to group speeches made by members of the 110th House of Representatives. I was able to use mallet to determine groupings based on keywords used. I utilized multiple charts to visualize these groupings and make conclusions based on what the visualizations showed. The format for the report was provided by the Professor and I thoroughly enjoyed it. Even though not all reports I will write as a data scientist in the future can be put into the same format, I saw the benefit of having a clear and concise progression as I completed the assignments over the course of the term.

My most recent term brought the process to a close. I took Big Data and Data Visualization (IST 719) together which made for a great learning experience. First, I used R in Data Visualization and Python in Big Data. An immediate struggle was moving back and forth between the two languages. As I dealt with this learning curve, I was able to achieve the most learning due to the foundation of the previous courses. For Visualization, this was most apparent in my final project which consisted of a poster detailing acquisition and analysis to answer questions.

For this project, I looked at another of my interests, the expansion of E-Sports within the Entertainment industry. Specifically, I wished to look at the intersection between popular and competitive titles. What I found was that only a handful of games were equally competitive and popular while some titles such as Grand Theft Auto V were popular within the streaming space but had practically no competition. Oppositely, Counter-Strike: Global Offensive is the most competitive title, but it is not as comparatively popular. To go further into a potential advertising possibility, I also looked at the top streamers, the games they played and how often they streamed. Finally, I looked at the financials of each competitive game. Logically, if competitions for a title have a large prize pool, that will drive players to continue competing in that game. This combination of factors Identified “Fortnite”, “League of Legends” and “DOTA 2” as the best games to partner with or advertise adjacent to. A fourth game, “Super Smash Brothers Brawl”, had a very passionate fan base. The title did appear as a top competitive game but its most distinguishing feature was its average viewers per stream for the Brawl streams in my dataset. Ultimately, visualization up to that point had been a combination of improvisation and searching for solutions online. The previous coursework did cover some visualization packages, but know I feel confident communicating results in a clean manner in the future.

While visualization is important, it only provides an overall notion. The concise results that provide actionable insight are commonly obtained through data mining. I achieved the most straight forward results from data mining in the first two assignments of Big Data. The first of three lab assignments for this course consisted of collecting a variety of data to predict whether Syracuse should pay a hypothetical new coach based off stadium size and attendance, academic performance, and revenue. Using this collection of data, I ran a variety of linear models to achieve the model that accounted for the most variance of Coaches salary for all universities who fielded football teams in the Football-Bowl-Subdivision for Division I athletics. Given this output, I determined that the coach for the Syracuse football team should be paid about 3.2 million US dollars. This is much more than the salary for current coach, Dino Baber’s, whose salary is 2.4 million according to USA Today. While this assignment purely for educational practice. Similar real-life results would be highly important to both parties in helping them make decisions and negotiations regarding salary.

The second of three lab assignments looked once again at housing data. Given the housing crisis in 2008, analyzing housing data is potentially one of the most important solutions that data science has provided consumers (Zillow AI Blog, “Data Science at Zillow”). This assignment used housing market data from Zillow to look at a few choice locations and then attempt to predict the postal areas that will experience the highest growth within the next year. While, using time-based data to forecast was a subject covered in Data Analysis and Decision Making, this was the first assignment entirely dedicated to forecasting. The goal in the assignment was to determine the top three postal areas in which to purchase real estate. The report detailed the market data and forecast for four postal areas in Arkansas. Then it looked at the top 9 areas for projected growth. This was measured as the ratio between the forecasted market average one year from now and the current housing market average. These areas were compared to income distribution. This distribution was aggregated with data collected from data.gov. I identified this as an important factor because a high proportion of low-income residents means that such a population is less likely to purchase a higher cost residence and is more likely to move away from the area given the forecasted increase which inherently means more risk. The three areas with the best combination of low risk and decent ROI were two postal areas in Nashville, TN and Bradenton, FL.

Ultimately, I am very confident in my skills and knowledge as I seek to transition into a professional role. My final term includes courses that suite both of my potential professional goals, Natural Language Processing and Financial Analytics. While I look forward to learning these specific skillsets, I would be comfortable now taking on any professional role. Whether I am developing a data pipeline, cleaning and wrangling data, exploring data and communicating the patterns in the data visually, developing a machine learning model the provide actionable insight into the data, or some combination of all of these, I feel well equipped. Having done analytics in a variety of fields and received instruction from two different educational bodies within Syracuse, I feel confident to ensure proper security for sensitive data, and to communicate insights to the many different professionals that I will be working with in the future. While my most immediate prospect is to work

Berkowitz, Steve, and Jim Varney. “USA TODAY Sports.” USA Today, Gannett Satellite Information Network, sports.usatoday.com/ncaa/salaries/.

“Data Science at Zillow.” Zillow AI Blog, 18 June 2018, www.zillow.com/data-science/data-science-overview-2017/.

Davenport, T. H. (2014). What Businesses Can Learn From Sport Analytics. *MITSloan Management REview,* 55410th ser., 10-13.