Discovering, Learning, and Performing Data Science

by Christy Sato

GitHub Link to Project Deliverables: https://github.com/christysato/MS-ADS-Portfolio

I am not your “typical” Applied Data Science Master’s student. In most cases, Master students have studied a similar area in the field in their undergraduate program, whether that be computer science, analytics, or programming. But for me, I actually came from a Communications and Business background. While these are great fields to go into, the reason why I decided to pursue my Master’s in Applied Data Science was because I felt that I was still lacking the technical skills that would help me understand the business or market data in a greater depth. I can honestly say that I had no real definitive technical skills prior to applying to the Applied Data Science program at Syracuse University. I knew the Master’s program was something I wanted to pursue my education in and that it would be challenging, for that fact that I would have to work extra hard to catch up with other students who came into the program with a greater knowledge already. Thankfully, Syracuse University accepted me even with my little to no background in the field, which I am forever grateful for. Flash-forward to now, as I am in my last semester, I can reflect back to how much I have learned and grown as a data scientist. The classes and projects I have completed, demonstrate the vast knowledge I have gained just within one year of my studies. The program’s various areas of practice prepared me for the working industry and has given me the constant learning opportunities in the life-long career in the Big Data Era.

The Importance of Reviewing Data

As I took 4 classes per semester, I started to recognize that data science is not just one concrete solution or simple process. Data science requires a process of various practice areas like retrieving correct data, cleaning, evaluating, understanding, and testing before even reaching the point of deriving successful business decisions. The most important and time consuming part before starting any data science project is the data collection and organization of it. Datasets can contain over thousands of rows and columns, but in many cases, data is not clean, accurate, or organized in a sufficient way. Before we can complete any analysis, you need to ensure that the data is accurate, clean, and consistent. One would think that data collection is easy, especially when there are many online sources that can provide a multitude of datasets. But in reality, I noticed when picking a dataset for a project from these sources, some of these datasets do not have sufficient rows or columns. I learned that using small datasets cannot provide as many useful insights or solutions to business problems, since there is not enough data to understand why

things are happening. Another problem I have run into is that since these datasets are created by other users, in many cases, there is not a metadata file that will explain what each column means. This becomes an issue when a column is named, for example, “ProName”. It is not completely clear what this can mean. It is always possible to assume it means Product Name if the dataset was transactional in nature, but it is not the best practice to assume with data. In these cases, it is always best to go through the dataset, making sure it has enough data and contains a metadata file, rather than just picking one without any thought process.

Collection of Data

The most interesting way I have collected data through a project was through a Spotify and Twitter API. For my Scripting for Data Analysis project, I decided I wanted to accomplish something I have never done before, but in doing so, learn what features in songs, like danceability, energy, etc, make it a good study song in a Spotify Study Playlist. Using Python in Jupyter Notebook, I was able to collect the features for each song in different playlists to not only look at study playlists, but other genre playlists, to compare the two. In doing so, I noticed significant patterns in these playlists through analysis and data visualization. By first just comparing each of the study playlists; a couple of them had a very similar low average in danceability, energy, valence (how happy/positive a song may seem), and loudness. But then comparing it to other genre playlists, you see a high average in these levels. These were the distinguishing factors that made these playlists differ. To visualize the difference in study and non-study playlists, I imported the seaborn package to create bar charts that further narrate this analysis. I was also able to collect data from Twitter users who have used the “#StudyPlaylist” in their tweets. I wanted to see what Twitter users were saying about study playlists in real time data.

This project was the best way for me to learn Python scripting and the use of API’s as a data collecting tool. As a beginner in Python, I received the necessary practice in weekly homework assignments and a final project that I struggled with at first, but conquered eventually. I believe that this practice will help me in the long run because I received the basic skill in Python coding, that I plan to improve to become more advanced in, but also the soft skill of asking my Professor for help, and not being afraid of asking a “dumb” question. I think some professionals tend to be afraid of asking superiors or coworkers for help because they do not want to seem they are incapable. But when I tried to first work on the API’s, I had no clue what I was doing. I continued to work on it by myself which not only was time consuming but I was frustrated and

almost wanted to give up. However, not until I finally asked my Professor for help, that it was a lot easier than I expected. Not only did I actually understand what I was doing wrong, but I also realized that if I went to the Professor earlier, I could have completed more analysis on the dataset. Everyone in the working field will be stuck on certain problems and in best case scenarios they figure them out and can move on. But when you are stuck and cannot seem to figure it out, we should not be afraid to ask someone for help. This worthwhile lesson can be applied to future school projects or even in a professional manner.

Data Visualization and Cleaning

Once the data is collected and cleaned, it is ready for identifying patterns in the data. Data visuals are the best way to show customers, fellow employees, or stakeholders on what the patterns in the data are telling us. Visuals are easy to comprehend at a first glance, and people do not have to be technically savvy to create or understand them. In many of my projects, data visualization is the second step into explaining what the data has. This is important because it can show the distribution of data, if there are outliers or missing values in the dataset, and any key takeaways that can be further looked into. These are important to know as a data scientist because if data has missing values or outliers, there are steps into dealing with them. Just like I said previously, finding a good dataset is hard, so illustrating the flaws visually can be helpful to figure out what the next steps are for the data set. In my Big Data project, even though our loan default dataset had over 100 variables and 300,000 rows, we also noticed that most of the variables had over 100,000 observations with missing data. Additionally, our dataset was unbalanced, meaning that 92% of the observations were repaid, while only 8% were defaulted. We wanted to create a machine learning algorithm that would use this data to predict if a future home loaner would default or not. But when the target variable was clearly unbalanced, we ran into the issue of a bias that could lead to our predictions to always predict the customer would repay. The challenge was, how to make sure that bias is reduced and if missing values were completely removed, that we still had enough data to implement our machine learning. But what I have learned from this is that while it may seem better and quicker to just find another dataset, we were able to move past these issues. My team and I decided to drop the columns that had more than 60% of data missing and to balance the data by randomly selecting observations that repaid their loans and match that to the number of observations that defaulted. This class and project has prepared me with important concepts of machine learning, Python coding to create machine learning models, and dealing with these very common issues in a dataset, especially in

the real professional world. I learned how to deal with both supervised and unsupervised learning plus, different techniques in both classification and regression problems that are necessary for my career.

Statistical Analysis

Statistical analysis is very useful in data science because it can see which variables are affecting another variable. Using statistical knowledge like R-squared, p-value, coefficients and their slopes, data scientists can see what variables are strong predictors, strongly significant, or by how much it can affect a target variable. In my Data Analysis and Decision Making project, using statistical analysis was our way of figuring out what various variables were affecting life expectancy on a global aspect. We implemented linear regressions as an equation but also as a scatterplot visual so see if variables were negatively or positively affecting life expectancy. By seeing the scatterplots with the linear line, we can see how strongly correlated the two variables were and make really surprising discoveries into the dataset. For example, my project team and I assumed that alcohol consumption would negatively affect life expectancy. But to our surprise, the data told us that alcohol consumption actually has a positive effect on life expectancy. However, when it comes to these regressions the most important thing I have learned in all my classes is that correlation does not necessarily mean causation. So by looking at the relationship between alcohol consumption and life expectancy, we can actually take a closer look at the R-squared which was pretty low. This means that the correlation is not that strong and there are probably other factors involved that are causing this unexpected positive effect. This class's deliverables, like quizzes and homework, were a super effective way for me to have a deeper understanding of statistical business concepts like trends, seasonality, and forecasting which is used in any business that wants to predict their future revenue, stocks, or number of customers. I have actually already used many of these concepts and techniques in my summer internship and continue to reapply it in other projects. Now looking back at the class, I see the potential Excel has in these techniques and how much I have used it in the past year. Before I graduate, I want to get my Excel Certification to prove that my knowledge and skills in Excel are advanced.

Data Mining

Both my Introduction to Data Science and Data Analysis projects consisted with several data mining techniques. In the Introduction to Data Science project, our business goal was to

improve airline companies' customer satisfaction. To do this, we ran a support vector machine, association rules mining, and linear and logistic regressions. With the SVM, our major challenge of using it was due to the size of the dataset, but found the results to be interesting. My team and I wanted to see how the model would predict if a customer is satisfied or not. Overall, it would predict with 80% accuracy which was relatively good. Unfortunately, it took the code to run all through the night for us to finally get a complete output. Association Rules mining is usually used in transactional data to make product recommendations at a grocery store. But with this scenario, we wanted to see what factors are leading to a higher customer satisfaction rating of 4 or 5. It was really interesting to see that the age between 53 to 85 years old, business class, and male gender had a role in high customer satisfaction. Maybe more older men travel on a business trip than women, which is why their overall experience with flying is more pleasant and even paid for by the company, so they can comfortably sit in business class. With this same approach, we also analyzed what factors are leading to lower satisfaction. According to our findings, we saw women in their 50s to 80s who flew economy were the least satisfied with their flights. Using both the linear and logistic regressions, we had a closer look at what is negatively or positively affecting customer's satisfaction. In our final conclusion to the airline company, we advised them to lower delay times in their flights, develop strategies to improve overall satisfaction for women in their 50s to 80s flying economy, and continue to market airlines status' for gold and silver who generally gave a high satisfaction rate.

I recall when presenting this project in class, my team and I did not finish the presentation because of our time management. If one presenter spoke for longer than expected, we did not react in a way to complete the presentation in time. During a job presentation, it is important to consider time restraints, especially in a fast paced environment where people need to be done with meetings at a certain time. In this case, this was a lesson learned from our mistake. As a team, we need to practice our timings, our slides, and overall presentation skills. Another skill I learned that I will continue to carry into my career is the R programming language. It was another beginner's course to programming, but I learned so much from it within one semester. Always referring back to Saltz's and Stanton's "An Introduction to Data Science" was extremely useful during my lab work and project. I can officially say I am confident in coding in R and would love to implement it into my future work.

Completion of Learning Goals

As I am finishing up this last semester, I have realized that applying to Syracuse University has been one of my best decisions for my education and career. Just within a year's worth of school, I have experienced and accomplished everything I need to prepare me as a data analyst. With the help of all my projects, I fully understand that data science involves different practice areas, like data collection, cleaning, munging, visualization, statistical analysis, data mining, and so much more. The best way for me to perform all these data science areas is almost like a step by step approach. I first collect data either through API's (which I completed in Scripting for Data Analysis project), surveys, or online data sources like Kaggle. Then, with this data, I have to make sure that it is clean by checking if there are NA's, outliers, and other misleading data that could be fixed, replaced, or removed. I performed this step in all of my projects. Next, I identify patterns using R or Python to visualize the data given to me. Most of my projects included visuals in bar charts, scatterplots, line graphs, pie charts, box plots, etc. so I can see the distribution of data and understand if there are relationships between variables. I then can perform statistical analysis to see positive or negative correlations between features or using linear or logit regressions to derive equations that can help with seeing an outcome depending on several inputs. This was mostly used in my Data Analysis and Decision Making project. Since I now can understand the variables' relationships with one another, I can execute various data mining approaches to gain machine learning understanding of the data. I have performed support vector machines, association rules mining, and regressions on my Introduction to Data Science, Big Data, and Data Analysis projects. My Data Analysis project delves more into decision trees and Naïve Bayes Classifier which are other helpful machine learning solutions to a classification problem. Depending on the business problem, there are alternative strategies for machine learning. For example, I learned from Big Data that if there is an unsupervised problem, which means that the dataset has no definitive output variable, that one solution is clustering that can group data into similar groups. Combining all of this knowledge, I can finally develop strategies and plans that can help the problems that are being faced. Just like in my Introduction to Data Science project, my team decided that the best solutions were to improve delay time for flights, continue marketing and gaining more customers to join the gold or silver airline status, and improve the economy class to provide better comfort for females who are in their 50s to 80s.

In conclusion, the Applied Data Science program has really given me the foundation of all the technical skills I need to thrive in this career. But it has also introduced me to the communication skills and ethical practices when it comes to data and technology in general. The electives I have chosen has bridged those important factors into data science. The Information

Security class showed me how important protecting and securing data is for companies, since breaches are very common and can damage the lives of the company and the customers with it. Information Policy opened my eyes to all ethics involved with the advancement of technology. Issues like the government allowing access to our private accounts and phones, how to make sure AI machines are programmed ethically, and how policies are important and completely involved in technology ethics. My Data Warehouse class goes into depth into the importance of how to manage, organize, and implement data warehousing techniques at an internal level. All the preparation is for the business users that just want to see the facts and analysis. But data warehousing is the first step to ensuring data efficiency. These subjects are all vital to data science because it all revolves around one thing; data. Staying up to date and educated on these aspects make us stronger data scientists because it is never ending. Improvements in all these fields are necessary as more data and more technology continue to grow. Data science is a never-ending learning field, it involves collaboration with others, trial and error, and a collection of external knowledge and sources. These were the very reasons why I decided I would be a great fit as a data scientist.