Master of Science in Applied Data Science
Portfolio Milestone Report

In

Syracuse University
School of Information Studies (iSchool)

Bing-Je Wu

Bwu117@syr.edu

Table of content

Abstract

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, and information science. To apply the data science philosophy, other domain knowledges should be combined as well to help a company grow or making profit.

As a Syracuse University student in the Applied Data Science program, I, BING-JE WU, have completed the 12 courses (36 credits) offered by the School of Information Studies and Whitman School of management, including Primary Core courses (18 credits), Analytics Application Core course (3 credits), and Elective Courses (15 credits). In this program, I have concentrated mostly on the mathematics, statistics and machine learning tracks. The overview courses taken are shown below:

Primary Core	Analytics Application Core	Elective Courses
IST 687 - Introduction to Data Science	MAR 653 - Marketing Analytics	IST 652 - Scripting for Data Analysis
MBC 638 - Data Analysis and Decision Making		IST 772 - Quantitative Reasoning for Data Science
IST 659 - Data Administration Concepts and Database Management		IST 769 - Advanced Database Administration Concepts and Database Management
SCM 651 - Business Analytics		IST 664 - Natural Language Processing
IST 707 - Data Analytics		IST 736 - Text Mining
IST 718 - Big Data Analytics		

Program Overview

The Applied Data Science program, as an interdisciplinary program, provides students the opportunity to learn in a broad range of areas related to data science. As a successful student, the following objectives should be learned throughout the program:

1. Describe a broad overview of the major practice areas of data science.
2. Collect and organize data; identify patterns in data via visualization, statistical analysis, and data mining.
3. Develop alternative strategies or a plan of action to implement the business decisions derived from the analyses.
4. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
5. Synthesize the ethical dimensions of data science practice

During the program, by taking the courses mentioned above, many of objectives have been achieved. In the class of Introduction to Data Science, IST687, I got an overall understanding how people in the data science field work on a data science project. In practice, we always have a problem or an issue that we want to solve or to improve. This is the part of the business understanding. In order to achieve the business goal, the relative data should be collected or obtained for further analysis which involved in Data Acquisition and understanding. With the necessary data in hand, the next step is data cleansing, cleaning the data, formatting the data, and dealing with missing values. By finishing the first step, data cleansing, a further action, data wrangling, process of transforming and mapping data from raw data form into another form with the intent of making it more appropriate and valuable for various tasks, can be done properly. Those two steps are crucially important for a data science project. It normally takes up to 70% - 90% of overall project time. Therefore, knowing the right tool to help the process is important. And that is why I took Scripting for Data Analysis, IST 652, to learn how to use python, learning various python packages, using python and python libraries to build a data pipeline in aid of data analysis, visualization, and data mining processes. Apart from that, I also learned who to debug the codes by myself, going through the web searching, documentation review, and online resource tutorial. It improves my coding ability and let me realize how to organize my code not only for readability but also for reproducibility. In this way, others can read my code clearly and have the ability to reproduce the same result as mine. The course project I did in the course, IST 652 - Scripting for Data Analysis, provides me an opportunity to practice the process of data collecting, data cleaning, and data wrangling. Because it was a group project, I had to work with others. In this opportunity, I had practiced sharing code with others and organizing the codes for clear understanding. It was a great experience since I have not had this kind of experience at work because of the nature of my position. Also, through this opportunity, I have built the foundation of creating plots for data exploration analysis and for reports. I even implemented these skills at work to build a data pipeline for data-driven report to reduce the repeated work. So far, I have successfully built two pipeline projects at work and brought values to my team and department.

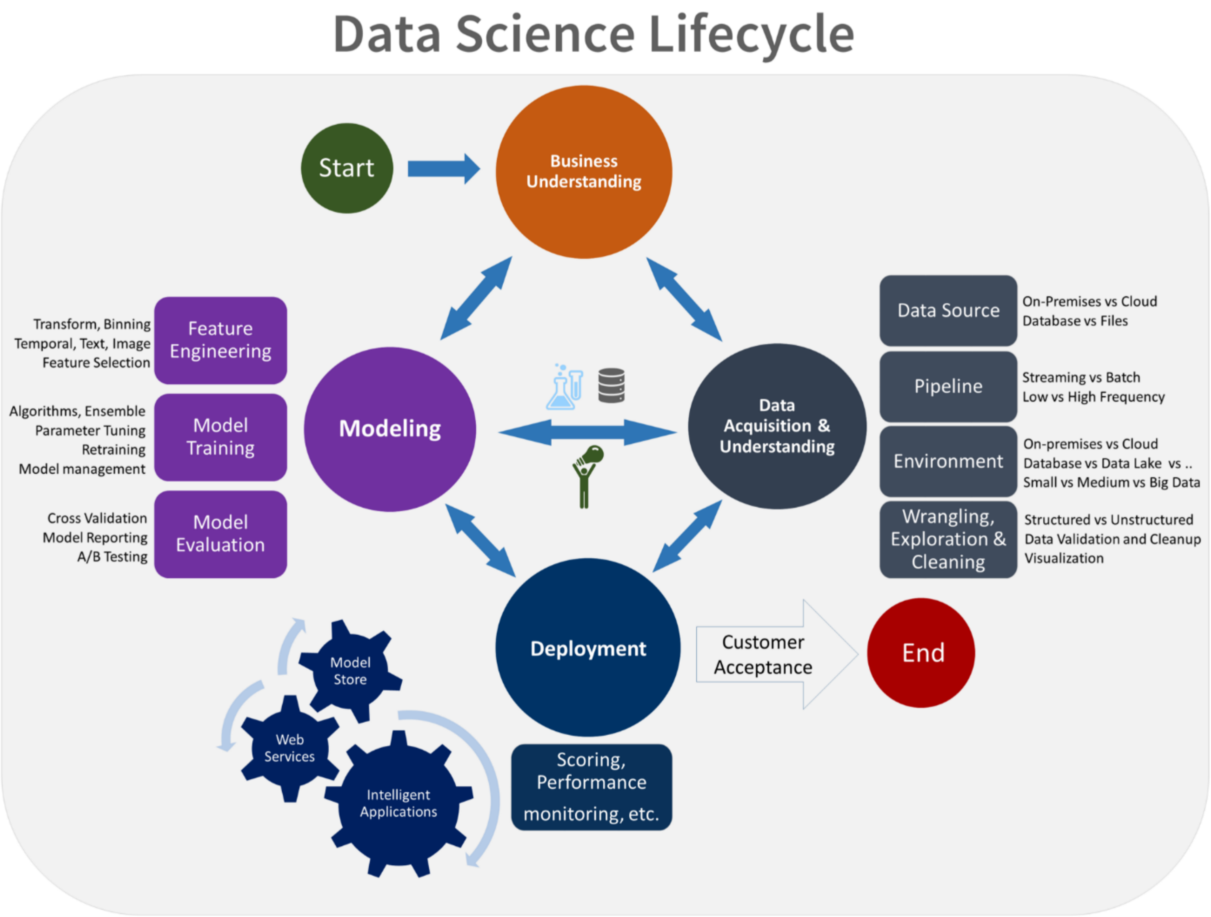

Image Source : https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

So far, I have covered the basic life cycle of a data science project. To bring a data science project up to the next level, some further statistics, mathematics knowledge and skills are required. Those skills are the fundamental of modeling process. And by having courses, such as MBC 638 - Data Analysis and Decision Making, IST 707 - Data Analytics, IST 772 - Quant Reasoning for Data Science, IST 718 - Big Data Analytics and IST 769 - Advanced Database Administration Concepts and Database Management, many of tools and skills have been added to my data analysis arsenal. Tools are including statistics test, predictive modeling, machine learning algorithm in both supervised learning and unsupervised learning, Bayesian thinking, time series analysis and varies form of database systems. The major technical skills that I have learned are how to organize data, how to make strategies for building models and how to evaluate models in order to put the models in the production. In the project of the IST 707, our group utilized the unsupervised learning algorithm and supervised learning algorithm to solve our research question, what factors affects the student performance. For the project of MBC 638, Data Analysis and Decision Making, I have experienced using statistical tools to improve a process. It requires the skills of determining which process needs to be improved and how to improve. With the process improvement plan proposed, we can collect data and show the before and after results as the evidence to support that the improvement plan is valid before putting too much investment on the improvement project.

Having the ability to deal with data and build models are not enough for a data scientist. To become a successful data scientist, we need to be able to convey our finding and observation to both technical persons and non-technical persons. Thus, writing reports is another art that I have to master. Luckily, some of the courses have provided the training for writing a report in academic format and non-academic format. In the class of Quant Reasoning for Data Science, IST 772, I have learned to simplify the technical terms to simple words. For example, the midterm and final exam of the IST772 required me to use statistical test to find if a new treatment is better than the old treatment. It is easy to use the statistic tools, such as t-test in R. It provides a quick way to run the test and show the result, but it is hard to translate those result to non-technical audiences. With the experience of writing the report as the midterm exam, I have learned how to give suggestions without using jargons to the managers or non-technical persons based on the findings and results.

For now, I am just a half-baked data scientist candidate. In order to push myself into next level, with the basic understanding of the life cycle of a data science project, advanced knowledge of data mining and communication skills, the last thing is to put everything together in practice. And, that is what I am doing for the Text Mining class, IST 736. This course provides a perfect opportunity for me to synthesize everything I have learned together and demonstrate it to audiences. A perfect chance to mimic a data science life cycle, having a problem or a goal to solve or achieve, collecting necessary data for modeling, making everything ready for deployment. If there is any problem or defect on the product, the life cycle will be circle around again to get the intended result.

Me and my team are currently working on a project to build a text prediction application. It is a huge challenge for us since we need to pull off the project within 11 weeks. This currently working project needs many requirements that include project management skill, data collection, data cleansing, data wrangling, data visualization, data analysis, models building, models evaluation, integration models in production, product testing, product finalization and final presentation of the product. We have defined our business goal, having an application to predict the next word and perform sentiment analysis on the typed content for a movie review. Two major predicted models were successfully built and ready for the production. A web application prototype also was created for testing and refining. In the last of few weeks, we are planning to try different models and have A/B testing on our product and find the best solution for the product in terms of computation cost and prediction efficiency. With all the skill sets that we have possessed, I believe we can present a satisfying product in the end. My goal is to put myself in the position as the center of the graph shown below.

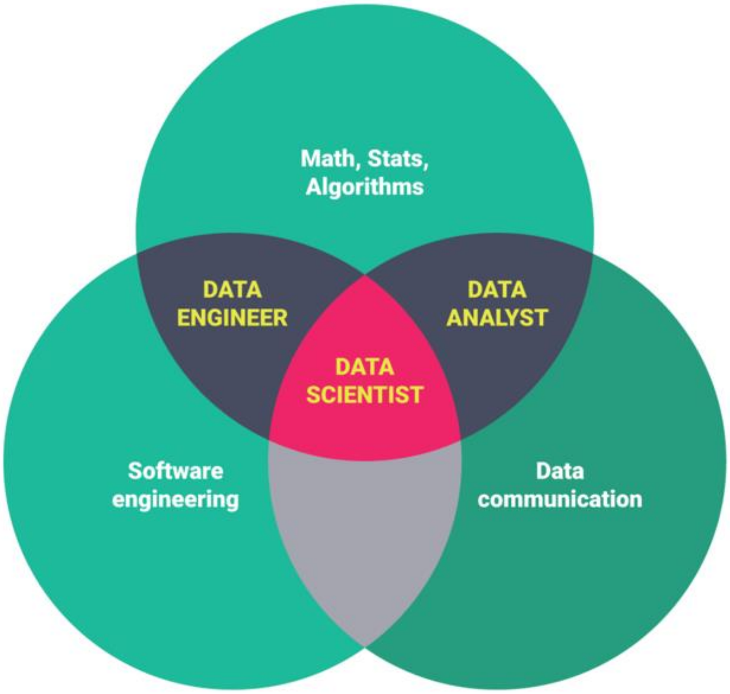

Image Source : https://www.springboard.com/blog/data-science-career-paths-different-roles-industry/

Projects

Project 1. Google Play Store Apps

IST 687 - Introduction to Data Science

Learning Objectives:

- Essential concepts and characteristics of data
- Scripting/code development for data management using R and R-Studio
- Principles and practices in data screening, cleaning, and linking
- Communication of results to decision makers

Skill Sets:

- Identify a problem and the data needed for addressing the problem
- Perform basic computational scripting using R and other optional tools
- Transform data through processing, linking, aggregation, summarization, and searching
- Organize and manage data at various stages of a project life-cycle
- Determine appropriate techniques for analyzing data

The goal for this project is to implement the skills learnt from the course and extracts valuable insight from a data set through statistics analysis and visualization, and Machine Learning algorithm by using R studio.

In this project, I have understood how to manipulate data in R and clean the data into a desired format to perform statistics analysis and visualization. Further, with the introduction of Machine Learning mentioned in the course, I have a fundamental understanding of how to implement a Machine Learning algorithm to build a model and get prediction.

The data set is Google Play Store Apps dataset from Kaggle (https://www.kaggle.com/lava18/google-play-store-apps). We have done many data cleansing steps on the dataset. We removed the missing values; converted ‘Size’ variable into the same unit and categorized it into different levels; removed duplicates; split ‘Genre’ variable into two; converted a variable from string type to numeric type and categorized it. We produced graphs to answer the business questions. And further, we implemented random forest algorithm and support vector machine algorithm to build a model to predict if a published app will be popular or not (installation more than 100,000 times).

Code and report: GitHub Link

Project 2. Student Performance

IST 707 - Data Analytics

Learning Objectives:

- Document, analyze, and translate data mining needs into technical designs and solutions
- Apply data mining concepts, algorithms, and evaluation methods to real-world problems
- Employ data storytelling and dive into the data, find useful patterns, and articulate what patterns have been found, how they are found, and why they are valuable and trustworthy

Skill sets:

- Data mining, including data preparation, concept description, association rule mining, classification, clustering, evaluation and analysis
- Be able to apply data mining skills to business, science or other organizational problems

The goal for this project is to extract valuable insight from a data set through implementing the data mining skills, including Data Preprocessing, Visualization, and Machine Learning Algorithms.

In this project, I have learned how to perform Data Preprocessing in R and implement transform on data to perform Machine Learning Algorithms in order to get desired predication outcome. The Machine Learning Algorithms implemented in the project include supervising learning such as association rule mining and K-means clustering, and supervising learning such as random forest, Naïve Bayes and Support Vector Machine.

The Student Grade Prediction dataset is from Kaggle (https://www.kaggle.com/dipam7/student-grade-prediction). We have done preprocessing on the dataset, as well as the data transformation. We converted the dataset into desired format for each algorithm. For clustering analysis, we converted nominal and ordinal variables into numeric variables. And for association rule mining analysis, we converted all attributes into factor variables. We produced several plots, and used Machine Learning Techniques, such as clustering analysis and association rule mining analysis, to answer our research questions. And further, we successfully build a Support Vector Machine model with 96% of accuracy on predicting if a student is going to pass or fail the final exam.

Code and report: GitHub Link

Project 3. Health Condition Improvement Project

MBC 638 - Data Analysis and Decision Making

Learning Objectives:

- Help students understand the value of data collection and analysis in acquiring knowledge and making decisions in today's business environment.
- Students will be able to identify and apply the appropriate statistical technique for a given set of conditions in order to answer a question.

Skill sets:

- Identify the problem of the process
- Implement DMAIC method to improve the process
- Utilize statistics tools to analyze and compare before and after

In this project, I have learned how to implement DMAIC method for a process improvement project. In the Define phase, I was able to determine walking distance variable as my output (y) and set up a goal for 4 miles per day.

In the Measure phase, I have decided how will collect the data of my process and clearly stated out that 9 day of data are history data. And, I identified some potential variables that could affect my output (y) by using Data Stratification Tree. Also, I used statistics tool to estimate my sample size and calculated the margin of error of the sample mean of output (y).

Several tools were used for the Analyze phase, such as SQL, Pareto Chart, Hypothesis Test, Confidence Interval, Multiple Linear Regression and Control Chart. SQL level has been calculated as 0.2 which indicates that my process is very poor. Pareto Chart shows that I should be more active. The Hypothesis

Test proves that performance of my process did not reach my goal. Confidence Interval states that my process performance is below my goal. Multiple Linear Regression shows that number of floors variable, lunch walk variable, workout variable, and number of active hours variable are statistically significant on my output y. The Control Charts implies that I may need to review my process based on dramatic shifts on Moving Range Chart.

In the Improve phase, I have determined the improvement actions based on the analysis from Analyze phase. The SQL level has been improved from 0.2 to 1.3. The Hypothesis test proves that the performance of my improved process still not reach my goal. The confidence interval states that my target is within the interval. The control chart has been more moderated and random.

For the Control phase, I made a conclusion that I have improved my process and the average output is closer to my target goal.

The Report: GitHub Link

Project 4. Pokémon

IST 652 - Scripting for Data Analysis

Course Learning Objectives:

- Write python scripts to access and amass data from fields in structured data, access fields in Semi-structured data, and define and find patterns of data in unstructured data
- Prepare and transform data to produce data summaries, lists, and networks
- Analyze and solve data access problems for the three types of data and to find and deploy appropriate software packages that can be integrated into the problem solution
- Frame real-world data questions and show how they can be answered from data

Skill Sets:

- Data wrangling, scripting needed to solve problems of accessing and preparing data in a variety of formats and situations
- Data science pipelines, from acquiring and cleaning data to accessing data and transforming data for analysis or visualization

The goal for this project is to work on the semi-structures and unstructured data, and to extract valuable insights from a web site or social media through data wrangling and data science pipelines.

In this project, I have learned how to scrape a table from a web page through web scrapping packages, how to access data from a social media platform through the social media API, and how to manipulate with JSON structured data, Python dictionaries and lists.

The Pokémon project contains four datasets from different data sources. The first dataset, as the main dataset, is from Kaggle, (https://www.kaggle.com/abcsds/Pokémon). Two datasets are Pokémon GO dataset and Pokémon moves dataset. Both datasets are from the same Pokémon database website, (https://pokemondb.net/go/pokedex) and (https://pokemondb.net/move). The last dataset is Tweets that contain Pokémon names, collected from Twitter. We used packages to scrape tables from the webs and utilized Twitter API to collect tweets. Then, we converted those collected data into structured data and semi-structured data for analysis. Data science pipelines were built to clean and transform data for visualization and for loops statements. We created functions, such as splitting a string into two for a column, inserting a bracket for special names, and breaking a string on camel case (a lowercase followed by an upper case). Data type and missing values were converted and replaced with new value. Some additional columns were added to answer the research questions. Further, plots and for loop outputs were implemented to answer the research questions.

Code and report: GitHub Link

Conclusion

Data science is an endless learning journey. All the skills, tools that I have learned during the program are just a tip of iceberg. There are still many skills, tools, theories and concepts need to be learned, adapted and implemented in our daily work.

My next goal after the graduation is to learn more about big data cloud computing. As of today, it has become an industry standard. The trend of cloud computing is growing. More and more companies are going to store their data in the cloud instead of having their own data center/warehouse, not only because of the cost but also lacking talents. Especially, we are entering the big data world. In the future, there will be more need on dealing with huge amount of data. Having the ability to extract business insights from a terabyte, a petabyte or even an exabyte data could be crucial in any business.