Master of Science in Applied Data Science

Portfolio Milestone Report

In

Syracuse University

School of Information studies (iSchool)

Chiau Yin Yang | SUID# 638291989 | cyang38@syr.edu

Table of content

"How has the data science master program cultivated me?"

Abstract

This paper is the Graduation Portfolio Milestone for my master's degree of Applied Data Science at Syracuse University, it is also a summary of my growth and reflection regarding my experience in the last 2 years. It will showcase my knowledge about data science and machine learning through the major academic projects I did in the previous coursework. Within the 2-year timeframe, the program has provided various courses to prepare me to take on a variety of problems - text mining, time series data analysis, information visualization, data warehousing and so on. I cherry picked academic projects that cover different areas of expertise and skills sets, including python, spark, SQL, and R. You can follow my journey of knowledge and critical thinking regarding data science, and the technical skills that are required to succeed in this field have progressed along the way. My portfolio milestone will cover my learning reflection from the 6 projects I picked, my latest resume, and the github repository with the related files and codes.

"How has the data science master program cultivated me?"

This paper is basically a summary of what I have specialized in during the 2 years in the master's program of Applied Data Science - the knowledge and critical thinking regarding data science, and the technical skills that are required to succeed in this field.

Data science, without a doubt, is the hottest buzzword in the past decade, it further creates the sexiest job in the current society. It makes businesses devise their marketing strategies more targeted and effective, helps banks and associated authorities identify and even prevent any potential frauds, and supports researchers to spot brand-new patterns or characteristics that might bring about a medical breakthrough. Data science, namely, is a discipline about data. The topic involves a wide range of focuses, including data analysis, text mining, business intelligence, deep learning, machine learning and many more. Each area of study can further break down to a variety of tasks, such as data collection, exploration analysis, data transformation and preprocess, model development, derived business insights and advanced actions. Data science is a big and complex topic that cannot be capsuled in this paper, but I picked six academic projects that cover different areas of expertise - text mining, time series big data analysis, information visualization, data warehousing and data mining. I would introduce them in chronological order so that you can see my knowledge and skills progress.

Academic Projects

Semester	Course Name	Project overview	Skills
2018 Fall	IST687 - Introduction to data science	**Airline Satisfaction Analysis** This project is to find out the possible factors to drive the overall customers' satisfaction score from the airline survey data using R and other machine learning algorithms.	R, Machine learning
2019 Spring	IST719 - Information Visualization	**San Francisco Crime Analysis Infographic** This project is to practice data processing and information arrangement, layout in the poster using programming language, R, and illustrator graphic tool. Topic is San Francisco 2018 crime statistics.	R, illustrator, information arrangement, poster creation
2019 Spring	IST718 - Big Data Analysis	**San Francisco crime instance resolution prediction** This project focuses on whether a crime instance could be resolved based on the location, time, type of crime instance. The tools used to perform the analysis are Python in PySpark cluster environment, and classification machine learning algorithms.	Apache Spark, Python, Machine Learning, Pipeline, Statistical Learning, time-series data analysis, databricks
2019 Spring	IST652 - Scripting for data analysis	**News Article Text mining** The project analyzes news articles from different publishers, and whether the machine can summarize major topics using LDA and NMF topic modeling without reading the content.	Python, Pandas, Topic modeling, Data analysis, word cloud, visualization in Python (Seaborn, matplotlib)
2019 Summer	IST736 - Text Mining	**Drug Review Analysis** The project studies reviews regarding various drugs from a pharmaceutical website (drugs.com). Topic modeling, wordcloud, and other NLP techniques were used to enrich the research.	Python, Machine learning, Natural language processing, Scikit-learn

2019 Fall	IST722 - Data Warehouse	**Data Warehousing Project** This project is to get familiar with data warehouse creation from relational databases, and star schema and cubes for further Business Intelligence reporting in all kinds of business scenarios.	Database, Data warehousing, ETL tools, business intelligence (Power BI), SQL Server, Microsoft SQL Server Data Tools (SSDT)

Airline Satisfaction Analysis

(Github link: https://github.com/cyang38/MS-ADS-PortfolioMilestone/tree/master/Airline-satisfaction-analysis-master)

The first project I have participated in is to analyze customers' satisfaction in the airline industry. The project was conducted in a group setting. The survey dataset used in the research was provided by the instructor. It contains 130,000 data points, and includes information regarding the customers - age, gender; airline and flight information - airline class, origin and destination city, number of flights prior; and the customer satisfaction ratings about the experience. The score ranges from 1 to 5, and 5 being the most satisfied. In order to seek useful insights and make business recommendations, we started with descriptive analysis. We used a programming language - R - to perform a series of exploratory data analysis on all variables. After having basic understanding about the data, we came to the census on how to handle some bad data, such as missing data handling and outliers removal. We also decided to only focus on a subset of the data, which is "cheapseats airline" based on the law of large number. The final number of observations we used is approximately 26,000.

Next, we split the tasks by the models we ran in order to investigate what causes higher satisfaction scores (4 and 5), and the models we ran were - (1) simple and multiple linear regression (2) association rules (3) Support Vector Machine (SVM). I volunteered to study the cheapseats airline data using association rules, mainly because this mining method is new to me, so it would be a great opportunity to gain hands-on experience with it. I then categorized the target variable, the satisfaction score, into 2 groups - happy customers (scoring 4 and 5), and unhappy customers (scoring 1 and 2). I used "arules" and "aruleViz" packages in R to search for interesting relationship between the 2 target groups and 20 other variables, including travel type, price sensitivity, flight cancellation, shopping and food at airport and many others information regarding the flights and customer status with cheapseats airline. Before modeling, I converted numeric columns into level bins - high, average and low based on its quantiles. After data munging, I started to tune the hyperparameters, such as support, confidence and lift, to narrow down the rules that give us more insights. I set happy customers and unhappy customers at the "right-hand side" and sorted "lift" value to find out what might lead to lower or higher satisfaction scores.

In this practice, I definitely gained lots of knowledge about how the model works, tested a few hypotheses towards what I thought was obvious, and got my hands dirty to preprocess and handle data, eventually generate and interpret the results, and made business recommendations. This project serves as a perfect first step into the data science field. From my perspective, the most valuable reflection from this project is to put the data science theory learned from the course into realization. It helped to build a comprehensive understanding when I focused on one algorithm. I apprehended many useful concepts by actually following through the data science process from data preprocessing to data modeling, to results interpretation and finally derived actions from the insights.

San Francisco Crime Analysis Infographic

(Github link: https://github.com/cyang38/MS-ADS-PortfolioMilestone/blob/master/IST719_Poster_v4_0214201.pdf)

The second project I did is San Francisco crime analysis. I used the same dataset and researched with two different techniques - one is information visualization, another one is big data analysis in PySpark. This dataset was collected by the San Francisco Police department or submitted via online reporting by the public since 2008. For the information visualization project, I only considered data points in 2018, and presented the information as it is a recap of the year. The biggest section is the San Francisco map color-coded by the number of instances of the neighborhood. For this project, I analyzed the data with one-dimension and multi-dimensional plots using R, and integrated it into a poster using illustrator. For example, I broke those instances down by day of the week, hour of the day, and daily and monthly distribution. Finally, I carefully picked each plot to best demonstrate the information in an effective and clear way and conveyed a uniform message. It definitely took me lots of thinking to dissect the data and arrange the layout in the poster.

San Francisco crime instance resolution prediction

(Github link: https://github.com/cyang38/MS-ADS-PortfolioMilestone/tree/master/SF_Crime_Analysis-master)

As for the big data project, my teammates and I selected 10-year data from 2008 to 2017, approximately 1.4 million rows in total. With this data size, it became harder to run on regular python. Therefore, the whole project was run on Apache Spark. We took advantage of a third party software, Databricks, and performed all of the work there. In this project, our goal is to predict whether a specific crime case can be resolved by the police based on the information of that crime instance, such as time, location, type of crime or combined. We tried a few different preprocessing methods in a pipeline and trained the models with three algorithms - logistic regression, random forest, and gradient-boosted tree classifier. Along the process, we have gone through a thorough data science process: raw data collection → data munging → feature engineering → feature selection → model training in the pipeline (with vector assemble) with three different algorithms for feature selection → model hyperparameters selection with validation set → model evaluation on test set → inference. It involved numerous techniques and trial and error to produce credible outcomes. It was a tedious and iterative process, and we have re-done and re-adjusted our settings multiple times. To be honest, it was quite easy to be lost in the process of model training and tuning in order to pursue the highest accuracy rate. However, knowing when to stop and how to interpret your findings are identically important as the goal is to solve business problems, not to chase the best performance blindly. I struggled a lot in the beginning in that I wanted to grasp everything before working on the data. Apparently, there are countless theory and academic principles about machine learning and predictive analysis, but nothing beats an exhaustive hands-on experience. This project was absolutely a precious episode in the journey, and I felt a great sense of achievement to actually derive some useful insights.

News Article Text mining

(Github link: https://github.com/cyang38/MS-ADS-PortfolioMilestone/tree/master/News-Article-Text-Mining-master)

Since most of the datasets I selected previously were structured data or tabular format, I chose to focus on something about unstructured data. The course is regarding data analysis with python scripting language, and it taught me how to analyze data using python and associated packages, like pandas, matplotlib and so on. The topic we chose is to explore news articles and summarize what topics were there without actually reading these articles. Two teammates and I found this dataset on Kaggle, and it includes around 143,000 news articles that are from 15 different publications. Those news were dated 2014 to 2017. We chose NPR news to be our research focus as it has the third largest number of articles, and the number of articles per author and number of authors are around 50% percentile compared to other publishers. Each author has produced a consistent number of articles. We ran the descriptive analysis using seaborn and matplotlib to visualize monthly productivity per author. Next, we suspected that each author should have distinct focus and style in their articles. In order to test this hypothesis, I used POS tagging from the NLTK package and filtered nouns and adjectives by each author. The nouns could help us identify what topics this author focuses on, whereas adjectives could imply the direction the author went into - positivity or strong negation. Before categorizing tags, basic preprocessing methods were used, including stopwords and punctuation removal and lower-case conversion. I used word clouds to demonstrate what the most frequent nouns and adjectives each author used in their articles, and it is clear that each author has different preferred themes. Some like to write about politics and international affairs, others prefer lifestyle, like books, music and so on. I also used bi-gram and tri-grams with frequency scores to help us understand the content, and they could potentially become hashtag recommendations. Interestingly, you can see that the top 50 most frequent nouns match the major events at that time, such as the 2016 US presidential election. This approach was surely refreshing for me because it deviates from my normal practice, even though it could be straightforward and even obvious to look at the data from this perspective.

Furthermore, we ran topic modeling using LDA (Latent Dirichlet Allocation) and NMF (Non-negative Matrix Factorization), and the machine suggested 4 and 6 different topics respectively. We tried different matrices, such as term frequency, document frequency and TF-IDF and compared their results. It took more effort for us to interpret the results since topic modeling is a means of unsupervised learning. For instance, we had to label each topic group based on the members in the group. Therefore, the inference process undeniably required considerable thinking and discussion. As far as I am concerned, inference and result interpretation are the most challenging yet essential part, because it demands decent understanding of the domain knowledge and proper awareness of statistics and mathematics. There are many uncertain elements that could cause poor products even with all the facts we know. Plus, there is never a straight guidelines or answer to any particular problems, but that is what makes data science a worth-exploring field.

Drug Review Analysis

(Github link: https://github.com/cyang38/MS-ADS-PortfolioMilestone/tree/master/Drug-review-master)

Later in the summer of 2019, I took the text mining course in order to develop proper knowledge regarding unstructured data. The course concentrates on common applications and machine learning algorithms specifically for text. I found this review data from drugs.com, and it contains approximately 53, 000 reviews on over 2,500 different drugs. Patients leave reviews with ratings ranging from 1 to 10 based on their experience and satisfactions. It was my first time working on reviews with satisfaction scores, and the main reason why I picked this dataset is I wish to be familiar with the healthcare or medical industry a bit. Since sentiment analysis is often conducted on reviews, I determined to perform a classification prediction on the usefulness of the reviews. The website serves as a forum where any patients who have tried specific drugs can leave their comments with a score, and other people can vote whether the review is useful for them or not. First thing I speculate is that the longer a review is, the more useful counts that review might get. However, that was not the case, I did not observe significant correlation between two variables, maybe because there was another spurious relationship. The longer a review had been posted on the website, the more counts it could accumulate. I then further tested another hypothesis that whether the machine can predict a review to be useful or not based on the content of the review. I categorized the usefulness based on the useful count by the quartile into 4 classes: 0 being less useful and 3 being the most useful. To normalize the count by the factor of time, I divided the useful count by the number of months since the review was posted. The overall accuracy rate was not as high as expected, especially for the 2 groups in the middle considering the fact that there are 14 duplicates in their top 20 most frequent bi-gram pairs. I recognized that it is harder for the machine to distinguish between mild differences (the middle groups) for multi-classification problems. Besides, the results from LDA topic modeling somehow matched a bit on the top 10 conditions from the data, and I believe it is due to most people mentioning the effects or the conditions they were having in the reviews.

For this individual project, I practiced to write and organize my discovery in a research paper format. It was new for me to compose a cohesive flow in a research paper, which had to be easy for people to follow and understand. I started with some conspicuous assumptions and gradually tested them out. There were times where the results challenge what you thought you knew, so you were forced to change your perspective and explain according to the findings. Progressively, I have uncovered that it is helpful to be flexible, open-minded, and to throw away any presumptions or opinions towards any topics, or even anything you thought you knew.

Data Warehousing Project

(Github link: https://github.com/cyang38/MS-ADS-PortfolioMilestone/tree/master/IST722_Deliverables_Team1)

After a few exercises in data science, I switched my focus a little onto database management and how data circulates in the process. Therefore, last semester, I took a data warehouse course. The course equipped us with how to build a data warehouse with ETL tooling in a variety of business scenarios. The project we did was to construct and implement two business processes using the fictitious Fudgemart and Fudgeflix databases. The organizational activities my team came up with were (1) shipment monitoring management and (2) risk management. The shipment management system, namely, is to assist business users track shipment status so that it is convenient for both customers and associated business management to monitor and create reporting. On the other hand, the risk management system is to analyze the time difference between customers' orders and their credit card expiration date to avoid potential business loss and detect suspicious activities. We used 5 business days as the parameters, if the credit card expiration date is within 5 business days from the order date, the purchase order will not go through or ship out. Meanwhile, the system would calculate the dollar amount of orders that were accidentally shipped, and eventually generated a blacklist of customers.

In this data warehousing project, we initiated the project with enterprise bus metrics in order to build a star schema. In the bus metrics process, the sheet helped us to list the dimensions needed and the major fact tables. Next, we followed through the detailed dimensional modeling sheet to create the tables in the database using SQL scripts. Once the tables were established in the database, we designed an ETL source-to-target map to aid our actual ETL (Extract, Transform, and Load) process. With the map, we implemented the whole process in the Microsoft visual studio SSIS tool from source to stage, and eventually to data warehouse. In the end, we built an interactive business intelligence dashboard so that the business users could perform advanced analysis and business reporting.

Among all the projects I have participated in, this data warehouse project is probably the most time-consuming and complex because, unlike other data analysis projects, there were strict rules to follow. If we did not execute certain tasks as planned, it would take us more efforts and time to fix afterwards. Compared to data science, data warehousing requires lesser creativity but more consistency and attention to details. Generally, when it comes to data science, people instantly think of machine learning, modeling or algorithms. However, having at least basic familiarity of how data flows, and when and why some organizations choose to build a data warehouse environment is undoubtedly a great quality. It enhances my landscape towards data science, and boosts a better understanding of the data flow in order to help me make better decisions in the future. In my opinion, how the data is stored, transformd, and moved is tightly associated with data analysis because we need good data to produce accurate insights, but most people just choose to neglect. All in all, I would not say this is a typical data science related subject, yet it is just as crucial.

Conclusion and reflection

In summary, during my time with iSchool in Syracuse University, I had conducted many researches on various areas and topics, both individually and within a group. From my viewpoints, despite these projects being diverse from one another, the core principles of data science are similar and they can be broken down into below stages:

1. Define problem statement and goals;
2. Collect relevant data;
3. Understand data - exploratory data analysis
4. Preprocess data;
5. Engineer features and select features
6. Model development

7. Model evaluation and adjustment
8. Result interpretation and inference
9. Business actions

This program not only cultivated my technical abilities, such as programming languages, statistics, and machine learning algorithms, but it also nurtured my soft skills, including teamwork, analytical thinking, and perceptive and open-minded. I came into the applied data science master program with little knowledge about machine learning, yet I graduated from the program with abundant hands-on experience and expertises in many areas. From those experiences, I discover that knowledge is rigid, however, how you apply and interpret are powerful and agile. Though the professors taught me plentiful concepts, putting them into real-world applications yourself is far more beneficial. I learned to always remain curious and suspicious of what you know and what you see, and never jump to conclusions too quickly. It is vital to look at things with an open mind, and embrace what the data may lead you. Last but not least, when conducting a project, plan ahead but also leave space for imagination as the things do not always go as planned.

Professional Resume

Here you can find my latest resume:
https://github.com/cyang38/MS-ADS-PortfolioMilestone/blob/master/Resume_chiauyinyang_021420.pdf

Author Information

- Author: Chiau Yin Yang
- Github Repository: https://github.com/cyang38/MS-ADS-PortfolioMilestone
- SUID#: 638291989
- Email: cyang38@syr.edu
- Program: M.S. in Applied Data Science at Syracuse University Information studies
- LinkedIn: https://www.linkedin.com/in/joy-yang11/