Applied Data Science Portfolio Milestone Learning Goals

Name: Yodit Ayalew
NetID: yyayalew
SUID: 773636611
Email: yyayalew@syr.edu
Resume: https://www.linkedin.com/in/yodit-a-17000a20/?jobid=1234
Course: IST 782 Applied Data Science Portfolio Milestone
Repository: https://github.com/yoditayalew/ADS-Portfolio-Milestone.git

Applied Data Science Portfolio

• • •

I.Introduction

II.Learning Objectives

2.1. Describe a broad overview of the major practice areas of data science

2.2 Collect and Organize Data

2.3 Identify patterns in data via visualization, statistical analysis, and data mining

2.4 Develop alternative strategies based on the data

2.5 Develop a plan of action to implement the business decisions derived from the analyses

2.6 Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization

2.7 Synthesize the ethical dimensions of data science practice

III.Conclusion

IV.Applied Data Science Courses Included in the Portfolio

4.1 SCM 651 Business Analytics: Housing Price Analysis

4.2 MBC 638 Data Analysis & Decision Making: Process Improvement

4.3 IST 659 Data Admin Concepts & Database Management: Designing database management system (DBMS) for Oncologists

4.4 IST 722 Data Warehousing: Analyze business processes of Fudgeflix and Fudgemart to capitalize on synergies of the combined company.

4.5 IST 718 Big Data Analytics: Book Recommendation System

4.7 IST 652 Scripting for Data Analysis: Semi structured Data Analysis

4.7 IST 600 Managing Data Science Projects

V.References: Wikipedia, Course asynchronous, Course books…etc.

Applied Data Science Portfolio

I. Introduction

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data (Data science - Wikipedia).

I have a 10+ years of experience in the IT industry mostly software development, Data Governance, Data Quality, Data Architecture etc. Prior to joining the ADS program at Syracuse, I thought data science is just being proficient in Python and know some machine learning and AI, what I didn't know is how diverse data science can be. The ADS program really opened my eyes where I took courses like Business Analysis, Data Analysis and decision making, Quantitative reasoning, Big data analytics, Data Warehousing, Data Admin concepts and Database Management, Scripting for Data Analysis, Information Security and Managing Data Science Projects where after gathering the different disciplines of ADS this course helped me understand how Data Science Projects differ from other IT projects and so how it should be managed. In the ADS program, there are several goals the program looks to achieve which includes

- Describing a broad overview of the major practice areas of data science
- Collect and Organize Data
- Identify patterns in data via visualization, statistical analysis, and data mining
- Develop alternative strategies based on the data
- Develop a plan of action to implement the business decisions derived from the analyses
- Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization
- Synthesize the ethical dimensions of data science practice

In this paper I have selected 7 projects and/or assignments that I think helped me achieve the goals mentioned above in which I will discuss below

II. Learning Objectives

Goal 1. Describe a broad overview of the major practice areas of data science

Data science and the many areas of data science therein, broadly defined, is the science, study, and use of data in a digital platform. As this is a massive and growing area of expertise today, there are naturally a growing number of specializations to be found within the field. The following are some of the important, specialized areas of expertise in data science right now. Below is the list with short description of what each of them means

- Statistics and Probability
- Python
- Machine Learning
- Data Processing
- Data Visualization
- Data Mining
- Predictive Analytics
- Big Data
- Modeling
- Data Consultancy

Statistics and probability: represent a considerable area of mathematics that also greatly impacts data science. This specialty area is all about establishing and working with finite figures as well as the effects of the ever-present factor of "chance" in all things. Those additionally learned in this particular area are a great asset to general and specialized areas of the data science industry today.

Python: While understanding the ins and outs of Python isn't always required in data science jobs, it is a growing necessity that is a valuable commodity for the worker here to present with. Python was created several decades ago but remains an incredibly important programming language used in countless computer applications today. In addition, applications that do not utilize Python often require interpretation so that they may work in tandem with those programs that do. In the end, Python is a valuable specialty asset to know in data science.

Machine learning: is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Data processing: at its core, is the term used to describe the various processes computers use in the handling of data. Most people understand the premise of data and its simple storage, but beyond these basics, data is moved, encrypted, translated, compressed and decompressed, and much more. Data processing, subsequently, is

the specialty knowledge area of data science that specifically handles all of these data processes.

Data Visualization: As its name suggests, data visualization is the data science specialty area focused on how data can be potentially presented in a visual manner. A large portion of computer use today has to provide the end user with a way with which to see and visualize the data being presented. Examples of data visualization concepts include readable text, holograms, interactive data displays, and on-screen charts and graphs. This specialty field works on making new ways of visualization as well as improving older methods.

Data mining: is the important data science specialty area that focuses on finding certain patterns in large pools of otherwise loose data. Once these patterns and associated values are established, they can be further utilized in machine learning, big data, and numerous other data science venues

Predictive analytics: is used all throughout the data science sector as well as throughout many of the other data science specialty areas mentioned here

Big Data: As its name suggests, "Big Data" is the term given to extremely large sets of data. In the data science world, these particular sets of data are said to be characterized by "The 4 Vs". These four, telling attributes, all starting with the letter V, are volume, variety, velocity, and veracity.

Modeling: In data science, there is a very regular need for the use of illustrations and diagrams in looking at varying kinds of data. Through the use of these visual tools, workers can then identify valuable patterns and other markers as well as generally work with that data. An example of data modeling at work might be a graph and chart setup designed to show a company's purchase costs history

Data Consultancy: Finally, data consultancy is a sort of operational collection of the many specialties and even general practice areas of data science. In this line of expertise, the worker, in this case, called a "consultant", works with clientele from different companies to help provide advice on their various data science needs. This is an outside contracting position in which the worker provides their services to paying customers with whom they have no other affiliations. From start to finish, a basic rundown of the process includes initial consultation, assignment to work on a specific issue or issues, investigation of those issues, presentation of a report on the findings, and subsequent work with the client to fix or improve upon those issues.

Goal 2. Collect and Organize Data

Data collection is the process of gathering and measuring data, information, or any variables of interest in a standardized and established manner that enables the collector to answer or test hypotheses and evaluate outcomes of the collection. https://en.wikipedia.org/wiki/Data_collection.

I am currently a contractor technical project manager for a government agency and deal with massive amount of data day to day and through out my ADS courses I had a chance to collect and organize data from personal day to day expenses to, Data in Jason format to Data by using API

Course Alignment:

1. In DATA ANALYSIS & DECISION MAKING (MBC 638) course, I worked on a process improvement project called "DECREASING DAILY DISCRETIONARY EXPENSES BY 45%" where I collected data for 9 weeks and organized it. This project was focused on Defining what you want to get at the end, collecting data, Baseline Measuring where I used descriptive statistics such as identifying mean, median, standard deviation and range, analyze the data where I did hypothesis testing, Correlation and regression, improve the process and controlling it

2. In Big Data Analytics (IST 718), for Book Recommendation System project, collected multiple data as we used multiple sources and built 4 models which is correlation recommender, Item based content filtering by using KNN-Unsupervised Machine Learning Algorithm, User Based content filtering by using KNN-Unsupervised Machine Learning Algorithm and Content based filtering by using NLP – Unsupervised ML algorithm and finally compared the 4 models and recommended the best model.

3. In Scripting for Data Analysis IST 652, collected and organized a semi structured data for my project called Traffic Violations in Montgomery County Maryland where I prepared the data, extracted information for the Json file, analyzed it and answered questions

4. In Business Analytics SCM 651, for Housing Prices project, Organized data and run correlations and regressions and identified and explained why each variables is intuitive/non intuitive and answered multiple questions that was asked.

Goal 3. Identify patterns in data via visualization, statistical analysis, and data mining

Pattern recognition is the process of recognizing patterns by using machine learning algorithm. Pattern recognition can be defined as the classification of data based on knowledge already gained or on statistical information extracted from patterns and/or their representation.

Project alignment

1. In DATA ANALYSIS & DECISION MAKING (MBC 638), for my project, DECREASING DAILY DISCRETIONARY EXPENSES BY 45%" multiple visualizations were done to show the control phase of the project. I have also used statistical methods such as correlation and regression. In doing so I have also observed patters in the data as to where change should be made to improve process.

2. Big Data Analytics (IST 718), this is a big data project which uses machine learning algorithms to recommend books to users and identify data trends and patterns of attributes. For example, while doing the EDA, I realized that most books were written in English and the number of books written in other languages were insignificant, also most people tend to give high ratings, very surprisingly there were individuals who gave around 200 reviews in the data set used

3. In Scripting for Data Analysis IST 652, for my Traffic Violations in Montgomery County Maryland project, I have used Visualizations to show patterns of data as to what days have the most traffic violation, which gender had the most violations and the type of violation etc.

Goal 4. Develop alternative strategies based on the data

1. In Business Analytics SCM 651 the Project Recruiting Advertising Strategy is a good example of developing an alternative strategy based on the data (This project is not completed as I am taking the course and working on it now. I am halfway through it, but it should be ready when I submit my final portfolio)

2. Big Data Analytics (IST 718), this is big data project, when we started it what we planned to do was create different models and test the model to get the best model but we also planned to include if for example we can include if what our book recommender recommended as best book is also the best book for the general public by using twitter hashtags but when we look at our data set, the recommender we can create was more based on the users experience in the data set by using content filtering and collaborative filtering and to try to see what the public think about those books wouldn't make sense and so we didn't include any sentiment analysis from

twitter for the specific book recommended (to be expanded fot final Portfolio submission)

Goal 5. Develop a **plan of action** to implement the business decisions derived from the analyses.

1. For DATA ANALYSIS & DECISION MAKING (MBC 638) Project "DECREASING DAILY DISCRETIONARY EXPENSES BY 45%" I have used my families personal data for this project where I collected my data for 9 weeks and used the process improvement framework to see how I can improve my expenses, I had the following action items which I used and benefited immensely.

   Action Plan

   - Make soy latte at home
   - Amazon prime purchase restricted to not go more than 100$ a week
   - Make lunch at home
   - Restrict date with friends to be held biweekly

The goal of the project was to reduce discretionary expense by 45% but ended up reducing it by 67%

2. For my IST 659 Data Admin Concepts & Database Management Course, after I built the database tables, I presented to my team at the time how instead of using spreadsheets the company can create and use data bases

3. For IST 722 Data warehousing course, Product Review Analyze business processes of FudgeFlex and FudgeMart to capitalize on synergies of the combined company, at the end of the project the following recommendations were provided for the company to get better product reviews

   - Fudge Co should investigate the hardware department and figure out why it is having low ratings.
   - Fudge Co should target WI, OH, and DC with product improvement campaigns to increase customer satisfaction.
   - Include customer comment attributes and survey data to refine analysis on customer satisfaction.

Goal 6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization

1. For DATA ANALYSIS & DECISION MAKING (MBC 638)

   In data analysis and decision-making course, I learned process improvement where I learned how to

   - Define, measure, analyze, Improve and Control (DMAIC Process)

   I did my final project to decrease my own discretionary expenses by 45% and ended up decreasing it by 67% and I presented my process improvement for my husband who is a data scientist himself. In the process so many statistics concepts were involved, from descriptive statistics to hypothesis testing, to Correlation, to regression analysis. See GitHub Doc

2. For Big Data Analytics (IST 718)

   Big Data Analytics: Created a book recommender system.

   - Used two dataset's, Books data set with 10,000 books and ratings dataset with 1,048,575 ratings
   - We did exploratory data analysis on each data set, cleaned the data sets, removed outliers and then merged the data sets
   - Created the following 4 models: Correlation (Pearson's), Collaborative – Item, Collaborative – User, Content Filtering
   - We tested each model and finally recommended the best model
   - Advanced models like KNN-Unsupervised Machine learning Algorithms, NLP-Unsupervised Machine learning Algorithms and Correlation recommender were used. This course is where I learned how creative one can be and I also learned that data science is research, there is no one formula which can solve your problem and so you must try multiple things and that is how you learn. I loved this course so much.

Goal 7. Synthesize the ethical dimensions of data science practice

1. In IST 644 Managing Data Science Projects

   Managing Data Science Projects course in the ADS program is one of the best courses I have taken, and I got an opportunity to learn about an ethics primer in data science where the Professor addresses it beautifully and in a way one can't forget.

   In general, when it comes to ethics in data science, it's easy to make a mistake if we're not on guard, even the most kindhearted, well-intentioned data scientist can make unethical decisions at times because of this. Also, most data scientists are trained in disciplines like applied mathematics, computer science, or statistics. In fields like these, data science is used mostly for research and academic theory, rather than to inform real-world behaviors that affect people's lives

   Personal data such as passwords, photographs, and location information can fall into the wrong hands. Predictive models used for policing and sentencing can reinforce stereotypes and have adverse racial or socioeconomic implications. Economic opportunity in the form of school admissions, job hiring, and loan approval can be denied. Healthcare decisions could be made incorrectly, compromising a person's health and even their life.

   But if data scientists like us are diligent up front, particularly in terms of preserving privacy, mitigating the risk of attack, and avoiding bias in our model, ethical it can be reduced.

2. In Big Data Analytics (IST 718), when we selected our data set, we were very careful about the data and made sure we aren't using any personally identifiable information in the dataset

III. Conclusion

In my opinion, by enrolling in the ADS program, I have achieved the 7 goals listed above and a lot more than that. Especially for someone like myself who has a 10+ years of IT experience, this program is more than enough to know what Data Science is, how diverse it can be, where I want to focus and how I can expand what I have learnt. In fact, I have already started using some of the skills I got here in my current position. I now don't use the holistic approach of managing IT projects for some projects that needs research and has a data science touch to it. The data analysis techniques I learnt here are being used to present data in a better way for high level executives, my statistical knowledge grows exponentially, and I now can quickly analyze data by running quick correlations and regressions, I now know how to use the DMAIC framework to improvement processes and more.

IV. Applied Data Science Courses Included in the Portfolio

4.1 SCM 651 Business Analytics: Housing Price Analysis

4.2 MBC 638 Data Analysis & Decision Making: Process Improvement

4.3 IST 659 Data Admin Concepts & Database Management: Designing database management system (DBMS) for Oncologists

4.4 IST 722 Data Warehousing: Analyze business processes of FudgeFlex and FudgeMart to capitalize on synergies of the combined company.

4.5 IST 718 Big Data Analytics: Book Recommendation System

4.6 IST 652 Scripting for Data Analysis: Semi structured Data Analysis

4.7 IST 600 Managing Data Science Projects

4.8 IST 723 Information Security

V. References: Wikipedia, Course asynchronous, Course books…etc.