SYRACUSE UNIVERSITY
MS Applied Data Science
Portfolio Milestone Winter '22

Jiebin Zhu

Email: jzhu11@syr.edu
SUID: 541649864

Table of Contents

About Me

My name is Jiebin Zhu, but my preferred name is Alex, so some of the works that will be submitted here might contain those two names interchangeably. At the time of writing this portfolio (March 2022), I am currently a Product Associate at Wayfair. I was a recent undergraduate of Syracuse University (May 2020) that obtained a BS in Information Management & Technology and in Finance. I may be early in my career, but I've developed an interest within the data science field while being in college over the past few years and understood what data scientists could bring to the table to help revolutionize the current industry. This ultimately led me down to the path of pursuing this degree to also seek out opportunities in this area as well.

Overview

Throughout the course of the Applied Data Science program at Syracuse University there are learning goals that are achieved by students. I, being one of the students in this program, would like to reflect on how I have incorporated my studies with the following learning goals and learned these concepts:

1. Describe a broad overview of the major practice areas in data science.
    a. IST 707 - **Housing Value Analysis within Staten Island**
2. Collect and organize data.
    a. IST 652 – **BOTW Assignment** / **Health Awareness Analysis**
    b. IST 718 – **College Football Coach Salary Analysis**
3. Identify patterns in data via visualization, statistical analysis, and data mining.
    a. IST 652 – **Health Awareness Analysis**
    b. IST 718 – **College Football Coach Salary Analysis**
    c. MBC 638 – **Operation Sleep Enough!**
4. Develop alternative strategies based on the data.
    a. IST 718 - **College Football Coach Salary Analysis**
    b. IST 707 - **Housing Value Analysis within Staten Island**
5. Develop a plan of action to implement the business decisions derived from the analyses.

   a. IST 652 - **Health Awareness Analysis**
   b. MBC 638 – **Operation Sleep Enough!**

6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
   a. IST 664 - **Twitter Sentiment Analysis Experiment**
   b. IST 652 - **Health Awareness Analysis**
7. Synthesis the ethical dimensions of data science practice (e.g., privacy)
   a. IST 618 - **Reflection Essay: Privacy and Security**

The link to GitHub here contains all the files for this portfolio milestone:
https://github.com/jzhu11/portfolio-milestone

IST 618 – Information Policy

Reflection Essay: Privacy and Security [Link]

-Synthesis the ethical dimensions of data science practice (e.g., privacy)

When learning about information policies, there's a lot that go behind it. People are always trying to find that perfect balance between high regulation/security with low privacy and low regulation/security with high privacy. Both are aspects are important and were reflected upon in the essay.

Within the class itself we were always providing our opinion in these types of matter. The ability to identify the pros and cons when encountering any type of ethical questions in data science and getting everyone's opinion on it, is very important. This soft skill is applicable practically everywhere, and this assignment allowed me to have that practical approach of looking at pros and cons from multiple angles.

IST 652 – Scripting for Data Analysis

The two assignments that really helped showcase these learning goals are the ones listed below:

- Homework 2: An analysis of "The Legend of Zelda: Breath of the Wild" Compendium (will be referred to as ***BOTW Assignment*** later on)
- Final Project: Health Awareness Analysis

BOTW Assignment [Link]

-Collect and organize data

The instructions of this assignment were to try and scrape semi-structured data from the internet and organize into a way for answering a few questions we may have in a report format. It also noted that we were able to obtain different data sets apart from the common Twitter and Facebook data, so I spun this approach into a more applied manner. The format I decided on with this is to obtain some semi-structured data from a game that I love a lot via an API. This assignment is dedicated on the game called "The Legends of Zelda: Breath of the Wild" aka BOTW. The API returned the information in a JSON format for me to scrape and turn into organized data frames with functions I created for further analysis. I was able to identify what types of enemies I need to defeat and where some of the common locations were. This helped lead me into ways of understanding the game that I never would have and provided some guidance into how to 100% complete the game.

Health Awareness Analysis [Link]

The purpose final project for this class was to test our abilities to demonstrate our ability level with Python to create scripts that collect data from either structured, semi-structured, and/or unstructured data from the web to produce data summaries. It also must surpass our homework assignments in complexity such as including additional datasets from different sources, hence the BOTW Assignment being provided as a reference. This project included a

real-life problem because I knew someone that has had a stroke at the beginning of the pandemic, so I wanted to investigate further into strokes for that reason, but I uncovered something more from that initial problem statement. I saw some potential worrisome patterns across illnesses and death counts. The learning goals that were achieved from this project below:

-Collect and organize data

Three data frames were generated from this approach. The first set of data was obtained via Kaggle that contained patient data that's used for educational purposes only without any means of directly tying it back to a specific individual and had a confidential source. It included information on whether the patient has had a stroke and if they have any heart related illnesses. The second set of data was directly exported from the CDC on the states' stroke mortality rate. Since the base dataset that contained some patient data on stroke and heart diseases, I decided to create my own by scraping the CDC site for the rank of the leading cause of death in stroke and heart diseases. Afterwards, all the data was cleaned as explained in the project report and imported as pandas data frames.

The hardest part about this data collection process was figuring out how to specifically scrape a portion of the web. Since the information was static on the CDC website at the time of completing the project, it took some figuring out to do it via scraping the HTML. However, this is a skill that I can carry along to my future since I might be able to do some passion project with web scraping using this technique and solve some other problems I may have encountered in life.

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

Some basic statistical analyses were conducted to the different data frames to obtain a better understanding of them. In this case it was mainly looking into the death counts and certain attributes that call out to me, such as variables that could possibly have a factor in increasing

the risk of an individual getting heart diseases and/or strokes along with some other different traits. Overall, I was able to identify some patterns from some basic statistical analysis here with the tables that I curated.

The biggest challenge in this aspect was the lack of visualization here in the report, but based off the tables that were created, it worked pretty well for me intuitively to grasp that basic understanding. However, if there was another chance to improve this project, more charts will be added within the report in the future. Despite the lack of charts in the report, the final presentation did include a chart for visualization for storytelling. That chart displayed the pattern in a more visual manner that was called out within the report.

This was a great learning experiencing and beneficial to my future career in terms of having the ability to quickly identify any types of trends to be wary about without doing a hyper deep dive that could take up more time. It also taught me that sometimes I could just use the more charts in reports rather than tables since it speaks more volume.

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

As mentioned earlier, another deliverable of this project was a presentation given at the end to the class. This was a quick presentation that I had to complete in just a few minutes. Within this short amount of time, I was about to explain the context of this entire project which included the problem statement, why this issue was important to me, and how it could be applied to everyone else who was listening in. In addition to the context, the major call outs to the report and recommendations I made were clear to the audience including how a major part of the program runs.

In terms of how this learning activity from this project would apply in my future career is to really remain clear and concise when trying to present a story and recommendation. The real

challenge of this presentation was essentially the time limit. We had about 3-4 minutes to present our project here, which challenged us to be nimble, confident, and prepared.

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

Based on the analyses, I summarized and called out some recommendations and alerted people to watch out for their cardiovascular health based on some of the factors that were identified. The high correlation between heart diseases and people working in the private sector was a big finding and can open more doors for further investigation in the future, as noted in the report.

I learned that a certain analysis would only answer some questions, but not all questions. However, it will open opportunities to do further research for answering some of the newer questions that may arise. An example here would be trying to identify the relationship between "stress" and job types. That's something I had noted in the report already, but it's an important skill to have when conducting an analysis as a data scientist in the future. I can apply this knowledge to identify things like "if I recommend X item, would a customer buy Y item more" in association rules.

IST 664 – Natural Language Processing

Twitter Sentiment Analysis Experiment [Link]

This project was meant for students to sharpen their NLP skills, from the initial data cleaning steps to building complex models and providing recommendations towards the end.

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

The final report for this was tailored to audience who don't have a high understanding in Natural Language Processing. In other words, if they know a little bit about NLP, they should be able to understand this report. The program was explained clearly and what specific functions do. Readers were guided along to ensure that most of the processes made sense and reasonable. When pieces of code snippets were introduced, an explanation was provided on what it does to provide enough visibility to the readers. It also provided means of replicating the project if needed. This report is very streamlined, meaning it takes the users from the start of data processing all to way to building and understanding some complex models at the end.

Essentially because this project was meant as an experimental task with text classification, it's really an amazing skill to have for the future. NLP researchers are always looking for ways to improve until a machine can understand human speeches without failure. In the case of this project, I've already made recommendations for improving this activity in the future, and that skill can be applied to future opportunities in this field.

IST 707 – Data Analytics

Housing Value Analysis within Staten Island [Link]

This project focuses on another real-life situation that I've encountered and want to use the power of data science to help guide myself through. This project allowed us the freedom to conduct some data analyses using different models and algorithms. I chose to focus this on a particularly niche housing market within Staten Island, NY because my family wanted to purchase a home there. This project will create a few models that essentially will help us determine whether or not a house that we are viewing on the market currently is worth its actual price when comparing to houses sold in its area with similar attributes. If it's a house that should be within our budget based on the attributes, then it would be worth looking into why the owner want to sell at a higher price than other similar houses in the area. In other words, it will provide some sort of support with the negotiation of the closing price if needed.

-Describe a broad overview of the major practice areas in data science

A problem was first identified since that's what would help shape and guide the way to making this work. I had a real problem I wanted to solve with assessing whether houses listed are within our family's budget or not so we don't need to spend extra time into a house that would eventually close at a price range higher than what we want to pay. The specific problem statement here helped shape the data collection and onwards. This is a skill that I currently have while working as a product associate since I would need to help identify specific problems and pain points that stakeholder are facing and assist them. This project helped me transfer this knowledge to another type of problem and was the eureka moment for me that many things can be solved using data science no matter how niche if the problem is clear enough.

The part after that was data collection and data cleaning. Those two are part of some major practice areas in data science here, because nothing would be possible without this. I was able to identify a base dataset from Kaggle, but it didn't serve as something that's applicable for my very niche targeted problem. Therefore, I took on the challenge to find a better dataset to

make it more applicable to my problem by utilizing the source of that base Kaggle dataset to lead me to the NYC government housing website. I was able to also filter to houses sold during the pandemic as well. The main learning activity from this was to really understand my problem and the type of data that's needed. Essentially it would mean garbage in, garbage out, forcing me to have a stellar and useful data set.

The main byproduct of this final project was the creation of several algorithms and comparing how effective they are in terms of using them for machine learning for classifying whether a certain house would be above or below our family's budget. RWeka was mainly used here to conduct these analyses for easy reproducibility and was the tool that was taught in the course. However, as an additional challenge for the future, I can see myself trying to use Python to conduct similar analyses because I am more familiar and confident with this language as I am almost done with completing the Applied Data Science Program. I can see how this is a great learning opportunity of comparing model effectiveness and parameter tweaking for other types of classification side hobbies in the future (i.e., to sell or hold a stock)

-Develop alternative strategies based on the data

Towards the end of the project, a model was ultimately selected and can be used to help resolve the initial problem. However, there were other models that had some quite accurate results. Since this solution was tailor made for my family, I commented on improvements that can be added for future reference if a similar problem or need appears. For example, if someone were to use a similar approach, they would have different steps they need to take to improve or alter the model.

I learned from this project that, in additional to having recommendations on what models to use for solving the problem, it's also important to provide alternatives with additional context so decision makers can use that information to apply it to their own situation.

IST 718 – Big Data Analytics

College Football Coach Salary Analysis [Link]

-Collect and organize data

In addition to the dataset provided, we had to go find additional college football data on the site. This was the most challenging part since the data we found on the web had different variations of its college names across different datasets (i.e., UCLA and University of California, Los Angeles). Fuzzy Matching in Python was used here to help merge these very different datasets together and organize them in a manner for building regression models.

-Develop alternative strategies based on the data.

In some ways the questions that were asked in this assignment limited to just the scope of getting the lab to be marked as "done". The most important question here and purpose of this lab was to identify the Syracuse University football coach's salary. Different salary ranges were provided based on a few different variations of the attributes such as the conference.

I learned that in the real world, it's always nice to have a range that decision makers can use to make these types of decisions. Just straight up providing the value of the coach salary itself isn't incorrect, but not a common practice that should be taken. Hence, including a range for other hypothetical scenarios (i.e., Syracuse moved out of a power 5 conference) would be good convention practice to include in reports.

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

Visualization was a big part in assisting me with building the regression models. I was able to identify key variables that would play a major role in the salary of coaches. It also helped me visualize if there were multicollinearity variables so those could be removed from the dataset even before building the model. In this case, many attributes that were included in the base file (Coaches9.csv) were removed except for TotalPay. I was able to create a dummy variable called

“power5” that denotes the power 5 conference schools, and the key variable that affects the TotalPay became clear. The most notable pattern here was that the Power 5 conference schools had higher coaches’ salary than the schools that aren’t in the power 5 conferences, which alluded that “conference” would play a huge role in predicting a coach’s salary.

MBC 638 – Data Analysis and Decision Making

Operation Sleep Enough! [Link]

This ultimate plan of this project is to help us devise a plan of action to solve a real-life problem we are encountering. The purpose of this specific project is to help me sleep earlier and try to obtain ~8 hours of sleep each night. I started off with identifying what's bothering me and quantified the impact in dollar amount. Afterward, I planned a data collection process on what data to collect and how I should collect it. Then I was off doing what I typically do for an entire month and started collecting data on when I was going to bed and when I woke up and a few extra things.

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

Within the Run Charts in the project, there was a nice visual chart that shows my spike of game time. It wasn't something that I was aware because there are days, I don't think I am playing games constantly, but with this piece of information I quickly came to the conclusion that my long hours of gaming were the main cause of me sleeping super late. Having the ability to understand a data pattern visually is helpful to many aspects of my professional career. No matter if it's related to what I am doing in the product management world or as a data scientist in the future, it's a crucial skill set that's applicable and allows me to complement my ability to build and tell and stories from it.

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

The data that was used for analysis is attached in GitHub. Based off the analysis, I was able to create an implementation plan that would allow try and help me sleep a bit earlier until I get ~8 hours of sleep. After submission of this project, I proceeded to test this implementation plan out and to my surprise it worked well initially, but things started to fall off since life is unpredictable and I started having more commitments to things after working hours. Personally, it was a huge success in terms of trying to solve a problem that's near and dear to my heart. Despite falling out of the mix there, occasionally now I sometimes would go to bed as early as 10:30PM! On a professional manner, this project helped me out within my job at

Wayfair when identifying inefficiencies within our team's workflows. I was able to find the problem at the source and saved our team a lot of time and manual effort. From the point of starting with data collection until providing an implementation plan based off the analysis, this is skillset that's applicable to my future as a data scientist.

Wrap Up

These courses and assignments/projects provide a list of well-rounded skills for me to grow and become a data scientist in the future. Considering all the different domains that data science touch, I am glad that I was exposed to many of these learning experiences and able to use this knowledge to apply it to my career. Thanks for reading this, if you have any questions, please feel free to contact me at jzhu11@syr.edu or LinkedIn.