|  |
| --- |
| Syracuse University |
| Portfolio Milestone |
| M.S. Applied Data Science |

|  |
| --- |
| Kelsey Johnson  Winter 2022  SUID: 768295784  [GitHub - kelsajohnson/Syracuse-Applied-Data-Science-Portfolio](https://github.com/kelsajohnson/Syracuse-Applied-Data-Science-Portfolio) |
|  |

Table of Contents

**Introduction2**

**Why Data Science2-3**

**Project Background3-5**

IST 659 Database Administration and Management3

IST 718 Big Data3

IST 707 Data Analytics3-4

IST 772 Quantitative Reasoning4

MBC 638 Data Analysis and Decision Making4-5

**Program Objectives5-**

Problem Identification (Objective 1)5

Collecting and Organizing (Objective 2)5-7

Exploring (Objective 3)7-9

Data Analysis (Objective 4)9-11

Interpretation and Recommendations (Objectives 5 & 6)11-12

Ethical Dimensions of Data Science (Objective 7)12-13

**What’s Next13**

**References14**

IST 687 Intro to Data Science Figure 1 Only References14

IST 707 References14

IST 718 References14

GitHub Repository14

Introduction

Data Science has become popular in recent years and many people use it to describe data analytics, but they are different. The main difference being that data scientists are well versed in computer programming and machine learning. Data Science is a field made of a combination of 3 areas of expertise: computer programming, statistics, and business knowledge. A popular approach to performing data analysis is OSEMiN (O’Neil & Schutt 2013). This stands for Obtain, Scrub, Explore, Model – Interpret. Many of the classes followed this same pattern and often expanded on it by formulating a problem in the beginning and providing recommendations instead of just interpreting the data.

Diagram

Description automatically generated

Why Data Science

In 2019 I was working as a Fraud Strategy Analyst and had the opportunity to help our model owners with documentation and to learn about our vendor models that we use to help predict and prevent fraud. As much as I was learning from work, I decided to go back to school and learn the process behind creating models. This Applied Data Science Graduate Program is formulated around 7 main objectives which help prepare students to become data scientists. These objectives are:

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

Project Background

IST 659 Database Administration and Management

In IST 659 we had an overall project in which students were tasked with designing and implementing a database to solve a data management problem of our choice. The project I chose was to create a database for Cass County Special Olympics of North Dakota to gain efficiencies by storing and maintaining data. This gives users the ability to compare new data to historic information. By creating this database, the Area Director or other users could quickly update, maintain, and delete data related to athletes, coaches, sports, and state games. This database was created to be used for swimming but was also built knowing that in the future it could be used for additional sports. The project documentation and files can be found in the IST 659 folder under Syracuse Applied Data Science Portfolio GitHub repository.

IST 718 Big Data

In IST 718 the final deliverable was a group project. We were tasked with specifying our own research problem or problems and developing the answer to each. Our project focused on whiskey, and we identified 4 problems/questions to answer.

1. Can cost be predicted based on attribute scores of whiskeys?
2. Can ratings be predicted based on attributes such as alcohol percentage, brand, and/or country (among other things)?
3. What are customer reviews telling us about whiskey preferences?
4. What recommendations can be made to consumers (based on specific preferences, attributes, etc.)?

The project documentation and files can be found in the folder titled IST 718 Big Data in the GitHub repository.

IST 707 Data Analytics

In IST 707 we again had a final group project. This project also followed the data science steps - cleaning, prep, exploratory data analysis (EDA), and visual EDA. It also required using Association Rule Mining, Clustering, Decision Tree, Naïve Bayes, Support Vector Machines (SVM) and Text Mining. The outcome of the project demonstrated our ability to create visuals as well as apply the different modeling methods we had learned that quarter.

My group decided to use the Google Hash Code Delivery Project, which was the problem statement for the online qualification round in 2016 (also referred to as the Kaggle Drone Delivery Problem). Due to the limitations of the specific problem, we were granted approval to limit our models and instead use basic algorithms to obtain the optimization of the drones. Research was conducted by reading articles on similar scenarios our professor helped identify. These were articles about the multiple traveling salesman problem and vehicle routing problem.

The Kaggle Drone Delivery Problem required optimizing the product delivery of customer orders from a warehouse in the most efficient time. The goal was to meet the scenario completion deadline as efficiently as possible to ensure a high score. Solving such a scenario would, in the real world, increase profits and decrease operation costs. Considerations for this project included the weights of the individual and collective products per delivery, turn-based time requirements, product availability, available drone commands, and entity coordinates, which are the locations of warehouses and orders. The project documentation and files can be found in the folder titled IST 707 Data Analytics Group Project in the GitHub repository.

IST 772 Quantitative Reasoning

For IST 772 there was not a final project, but rather a final exam. This wasn’t a test in the formal sense but allowed us to take the data provided and demonstrate our knowledge of the material we had learned by answering specific questions about the data. The goal was to conduct the analyses and write up a technical report for a scientifically knowledgeable staff member in a state legislator’s office. This included providing sufficient numeric and graphical detail assuming the staff member understood the concepts.

Dr. Block provided 3 datasets for use, and all pertained to vaccinations. The questions to be answered were split into two separate sections: Introductory/Descriptive Reports and Predictive and Prescriptive analyses. This was completed in R and a markdown file submitted. The project documentation can be found in the folder titled IST 772 Quantitative Reasoning in the GitHub repository.

MBC 638 Data Analysis and Decision Making

In MBC 638 the final individual project involved selecting an issue or opportunity that could be written as a problem statement. It had to involve accessible data that could be collected, measured (current and future state), and analyzed. Next, a solution had to be implemented and information controlled to take it to the next level of improvement. This was done by developing a business case to support the working issue.

I chose to use this opportunity to analyze the problem of not exercising enough during the day. Data was collected on myself, and success parameters were established along with my baseline. Several techniques were used in this process including a data stratification tree, data measurement plan, sigma quality level, normal distribution and hypothesis testing, confidence interval, chi square, and multiple regression to name a few. Both a Storyboard and PowerPoint were used for the final submission. The Storyboard was a fun way to show all the different work that went into the entire process on 1 slide. The project documentation can be found in the folder titled MBC 638 Data Analysis and Decision Making in the GitHub repository.

Program Objectives

Problem Identification (Objective 1)

Problem Identification happens often in the workplace and in our own personal lives. This happens when we notice a difference between what should be happening and what is happening. The difficult part is formulating the overall problem into a specific problem statement. This is the first step in applying data science.

For IST 659, the project was to create a database for Cass County Special Olympics of North Dakota to gain efficiencies by storing and maintaining data. The problem of lack of stored data had been occurring for years but had never been resolved. As a coach I knew there was no way to track athletes’ performance from year to year except for reviewing handwritten results from different meets and other information forms submitted. Meeting with other coaches helped develop the information to be included. It was found best to track not only athletes’ performance but coach and athlete information including medical conditions.

This project taught me that it is critical to work with other subject matter experts (SMEs) to identify a solution and formulate a plan to implement the solution. In most workplaces, the identification of the problem may involve more than just one business line. This shows that communication skills and being able to work as a group can be very important even if daily work is done individually.

Collecting and Organizing (Objective 2)

Collecting and Organizing data is a critical step in the data science process. It is impossible to solve a problem if the data is not cleaned and organized appropriately. How data is collected depends on every individual problem. It can be collected by hand, which requires it then be stored digitally or it could already exist digitally. If it already exists digitally, it must be formatted to fit the needs of the project which could require joining datasets, removing, and reformatting data, and transposing data.

In IST 659, to collect data I met with the other coaches and developed together what needed to be included in our database. It was decided to track not only athletes’ performance but coach and athlete information including medical conditions. This allowed me to apply firsthand knowledge in combination with ideas from other coaches (SMEs) from the swim team. Because this task was to create a database, the data needed to first be collected by hand from the coaches and then organized. Organization began by creating a data dictionary with all the fields required and then creating the Entity Relationship and Logical Model Diagrams (Figure 2). Multiple user interface options were designed which allowed the SMEs to pick the one they felt as a group was the easiest to use (Figure 3).

Logical Model Diagram (Figure 2)

Diagram, schematic

Description automatically generated

User Interface Example (Figure 3)

Graphical user interface

Description automatically generated

In IST 718 for the final project, we developed our own research questions. There were 4 problems/questions identified involving whiskey. First, data needed to be collected and organized. This was accomplished by searching the internet and sites such as Kaggle to identify datasets containing information such as price, distilleries, and ingredients. Significant data cleansing was required as 3 datasets were used (Figures 4 & 5). Once that was complete, we identified and formulated our official research questions.

These projects taught me that there are multiple ways to collect data. Collecting and organizing data is dependent upon the problem you are trying to solve and how the organization is built. In the first project I had no stored information to work with, so I had to review paper copies of information and then create digital copies of the data I was going to use. This was done by organizing the data and then determining how to store the data that would be used in the database. For the second project I needed to search the internet to identify datasets that could be used, clean and analyze the data and then clearly define and formulate the questions that were to be solved.

Before Data Preparation (Figure 4)

Table

Description automatically generated

After Data Preparation (Figure 5)

Graphical user interface, table

Description automatically generated

Exploring (Objective 3)

Data Exploration can only occur after the data has been cleaned and prepped. This stage involves creating visualizations to see how data is related and what is happening with the data. It also involves using statistical analysis to uncover any patterns or relations between the data. There are several techniques that can be applied to create visualizations and perform statistical analysis. The ones that will be applied will vary in every situation.

For IST 718, there was a lot of data to explore. It began with analyzing the distribution of scotch whiskey prices, distribution of scotch whiskey review points, plotting the price and price by review point, plotting locations of distilleries, and conducting a sentiment analysis on reviews which was shown by a word cloud (Figures 6 & 7). Several other data explorations were conducted as well. The ones mentioned were created by using boxplots, calculating descriptive statistics (mean, std, etc.), using histograms, and word clouds.

Distillery locations (Figure 6) and word cloud (Figure 7)

Map

Description automatically generatedA picture containing table

Description automatically generated

For IST 707 to perform any data exploration we first had to plot our data. The customer orders were plotted as orange circles, warehouse as green, and the original warehouse where all the drones began as blue (Figure 8). The next step was to analyze products by weights so we could start identifying how to load the drones (Figure 9). Between the two graphs it is easy to see that there are numerous possible combinations of products that can be carried by the drone and how much time is required to deliver orders centrally located and to customers who are located on the edge of the grid.

Location of warehouses and orders (Figure 8) and Product by Weights (Figure 9)

Chart, scatter chart

Description automatically generatedChart, histogram

Description automatically generated

These projects showed that exploration is always dependent on the data you are using. The first project included descriptive statistics as part of the exploration while that was not possible for the second project (nor would it make sense). Exploration helps you get a feel for what your data is showing you and what you can use it for. Being able to use graphs helps to visually explain what the data is telling you and provide insights. Two examples of insights provided by graphs for these projects were customers are willing to pay more for whiskey if it contains notes of honey and that due to locations, warehouses located on the edge of the grid will require more time to deliver products for centrally located orders and vice versa.

Data Analysis (Objective 4)

Data Analysis is performed after initial exploration has been completed. There are several approaches and techniques that can be applied. 3 categories of data analysis exist: Descriptive, Predictive, and Prescriptive. Descriptive involves using statistical methods to summarize data which leads to an understanding of the data. Predictive involves using past data to predict future outcomes used methods such as regression, modeling, and forecasting. Prescriptive involves optimization and forming solutions. These can include using one or both supervised and unsupervised machine learning techniques. Some common examples of supervised learning are linear and logistic regression and decision trees. Examples of unsupervised learning are clustering and association rule mining. Neural networks can fall under both categories. All of these are used to help the users identify solutions or trends as needed.

In IST 707 for the Google Hash Code Delivery Project there wasn’t numerical data to analyze. Instead, the locations and products going to those locations were analyzed. A cluster plot was based on a simple K-Means for 7 centers (Figure 10). This provided insight for us to consider how to ship products together for multiple orders. When optimizing the best route, clusters could be utilized for designating products to be shipped together for maximum efficiency.

Association Rule Mining was also used to determine if relevant associations existed between various products. This was done to further understand product relationships in customer orders that were partially identified during clustering. The Apriori algorithm was helpful in this situation to show us that there were never orders with more than 4 products. The algorithm utilizes confidence, support, and lift (Figures 11 and 12). For our scenario we focused on lift. It indicated the relationships between products are relatively weak or non-relevant due to the lack of orders they applied to.

Cluster Plot (Figure 10), Top 10 Rules by Lift (Figure 11)

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Top 100 Rules by Lift (Figure 12)

A picture containing diagram

Description automatically generated

To solve this problem optimization was used. A variety of ways of assigning orders to drones was utilized in attempt to increase the scores of drones. These assignment variations could be basic algorithms for optimizing the delivery scenario, such as assigning orders to drones according to warehouses or by the drones themselves. More complex algorithms such as genetic sweep, dynamic evolution, or gravitational search algorithms could also be applied. Unfortunately, most of these algorithms are not available for use in an R package which required this algorithm be built by hand, so only basic algorithms for assigning orders were utilized.

IST 772 had two separate sections on the final. Section 1 was introductory/Descriptive reports and section 2 was Predictive and Prescriptive analyses. Section 1 used basic descriptive statistics such as Min, Max, Median, Correlation and ACF on each vaccine. Several plots were created for comparing California to the US by vaccination rate (Figure 13). Section 2 contained both linear and logistic regression models as well as used Bayes Factor.

Descriptive (Figure 13)

Text

Description automatically generated

Predictive (Figure 14)

Text, table

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

A main takeaway from these two projects is that every project requires different tools and types of analysis. For the first project it could not be solved using the more common linear or logistic regression. To help determine how to build the algorithm Clustering (K-Means) and Association Rule Mining (Apriori) were used to identify how the products and deliveries should be set up. Then an algorithm was built to optimize the drone deliveries. For the second project; descriptive, predictive and prescriptive analysis was used. This allowed standard statistics to be calculated as well as formulating linear, and logistic regression and Bayes Factor Analysis. What I applied in both these projects I have been able to apply in my day to day work.

Interpretation and Recommendations (Objectives 5 & 6)

Interpretation and Recommendations require communication skills to accurately pass the information on to the end users. It’s important to know the audience you are presenting to before determining how you will interpret the information. Technical users will want to know more details and will require less explanation of the methods. The opposite could occur when presenting to customers or other users who do not need to know how the recommendation was reached.

IST 772 involved writing a technical analysis on the information that was found and providing a recommendation based on that information. Since this specifically requested it be technical, graphs and details on the analysis were provided. However, if this would have been requested to be a non-technical analysis, I would have provided a summary of the analysis and possibly provided definitions for any important information that the project stakeholders needed to review.

MBC 638 also helped to show another way to communicate to project stakeholders by creating a Storyboard to show the entire project process (Figure 15). The Storyboard was one slide of a PowerPoint presentation developed. A Storyboard is a great way to show in one place what was done to identify the solution. However, a supplemental report or PowerPoint may need to be used with it.

Storyboard (Figure 15)

A screenshot of a social media post

Description automatically generated

Both projects demonstrated unique different ways to communicate. A key takeaway is to know your audience. The report and recommendation supplied in IST 772 would have to be revised if it wasn’t being supplied to someone familiar with the technical knowledge. The same goes for a storyboard, it is a great visual and a way to showcase a project. However, in some cases it may need to be supplemented with a written report containing technical information.

Ethical Dimensions of Data Science (Objective 7)

Data scientists are in the unique situation where they are responsible for protecting data from being hacked, stolen, or used to commit fraud. This includes being responsible to not personally use the data they have access to in order to commit fraud themselves or sell to others. It’s important for organizations to hire the best employees by running appropriate background checks as well as purchasing or building the best software to secure their systems. Even employees who do not have access to personal data have a responsibility to learn about how data can be hacked and to use the systems appropriately.

This is an area I did not personally take a class on (such as Information Security), however working in ID Theft Analytics, my job is to develop solutions to prevent third-party fraudsters (fraudsters who do not know the customer) from accessing information. This is done in two different ways on my team; by predicting fraud applications and identifying account takeovers. Predicting fraudulent applications involves using model scores as well as writing rules to decline or queue applications. Identifying account takeovers involves writing rules as well as performing additional analysis to identify how the fraudster accessed the account. An account takeover occurs when the fraudster has enough personal information on the customer to gain access to their accounts. With what I have learned in school, I recently was lucky enough to move to another team within our ID Theft Analytics organization called ID Theft Rule Management. It has been a fun learning process where I can apply my computer programming knowledge gained in school as well as my modeling knowledge to help write rules (decision trees and neural networks).

What’s Next?

With school coming to an end, I will not stop learning about data science. I have the opportunity every day to learn more about different types of models and algorithms to help my team perform better. I also will be able to start having more time to learn how I can use python along with some vendor model scores to write rules. At work we mostly use SAS, so I plan on continuing to improve my SAS skills by taking additional training courses. I am excited to continue using what I have learned in this program and apply it to my career.

References

**IST 687 Intro to Data Science**

Figure 1 (Only) used from Corey Jackson’s Week 8 Review slides

**IST 707 References**

Christofides, N., Mingozzi, A. & Toth, P. Exact algorithms for the vehicle routing problem, based on spanning tree and shortest path relaxations. *Mathematical Programming* **20,**255–282 (1981). <https://doi.org/10.1007/BF01589353>

Rouse, Margaret. “What Is a Drone?” *IoT Agenda*, TechTarget, 15 July 2019, internetofthingsagenda.techtarget.com/definition/drone.

Tolga Bektas, The multiple traveling salesman problem: an overview of formulations and solution procedures, *Omega*, Volume 34, Issue 3, 2006, Pages 209-219, ISSN 0305-0483, https://doi.org/10.1016/j.omega.2004.10.004.

Welch, Adrienne, "A cost-benefit analysis of Amazon Prime Air" (2015). Honors Theses.

Google, #Hash Code Delivery, Problem statement for Online Qualification Round, Hash Code 2016.

**IST 718 References**

Scotch Whisky Dataset, Thatdataanalyst. (2018, March 22). *Scotch Whiskey Dataset* [data set]. Kaggle.

https://www.kaggle.com/koki25ando/scotch-whisky-dataset

Meta-Critic Whisky Database, Selfbuilt. (2021, August 27). Meta-Critic Whisky Database.

https://whiskyanalysis.com/index.php/database/

2,2k+ Scotch Whisky Reviews, Thatdataanalyst. (2018, June 13). 2,2k+ Scotch Whisky Reviews.

https://www.kaggle.com/koki25ando/22000-scotch-whisky-reviews

**Github Repository**

Github: [kelsajohnson/Syracuse-Applied-Data-Science-Portfolio (github.com)](https://github.com/kelsajohnson/Syracuse-Applied-Data-Science-Portfolio)