|  |
| --- |
| Master of Science in Applied Data Science |
| Portfolio Summary |
| Syracuse University |

|  |
| --- |
| Scott Snow  6-18-2019 |

The field of data science can be constructed as a ladder. This ladder had a very wide base but gets more, narrow as one goes up. Data scientists can either be professionals who climb and work at varying points on the ladder, or they can work on a very specific part. The top of the ladder is the action that is taken based on everything done previously. The goal of data science is to act based off all the other aspects.

In this domain, taking no action can be also considered an action but that decision is made based on the work that was done. The first step is to gather data. In certain workplaces, this is done manually with a constant input of data from workers and managers. Certain data that eventually is analyzed starts as blanks on a form. For my initial projects in my first term, this was how I acquired my data. I didn’t really know much yet and I didn’t realize both how difficult it is to effectively communicate data collection needs to other parties and how big of an issue missing data could cause in analysis. It was however a great learning experience as I gained valuable tools in SQL and Excel. For the Excel work, I was required to create a storyboard to present my data acquisition and the analysis that I did.

Other forms of data collection come from machine input such as scanners, sensors, and automated cameras. Most recently I used battery data to perform a time series visualization. The data was from a full lifespan test on a battery showing its charge and drain cycles. Outside of that, I have only done basic SQL work with grocery store data.

The most common form of data collection is from the internet. This is data that comes from websites or IoT devices. My final project for IST 652 involved web-scraping multiple Wikipedia pages to put together a list of songs. My most recent project for Visualization involved getting real time data on video Streams from the website Twitch.tv using their API. Other projects

Ultimately, in an education setting, these data collection processes resulted in only subset of the data that I cleaned, analyzed and used for data mining. The remaining data that I used during my education came from websites such as Kaggle.com which included ready to be imported data frames with either numerical data or already cleaned unstructured data that only required a small amount of effort to scrub and transform into structured data.

My first project and homework in MBC 638 refreshed much of my knowledge on statistical analysis including T-Test, normality, Correlation, and Chi-Squared test for normality. This knowledge was utilized again in preparing my final group project for IST 687 where we used a T-Test to compare non-finalist teams with a player replaced with the MVP or a finalist player had an aggregate measure for key metrics outside of the confidence interval or inside. Inside 95% we were confident that the given player could have been changed out without a drastic change in overall score. Outside 95% we concluded that that specific player was essential to their team’s finals appearance and that said team would not have played in the finals with an average replacement player.

Overall, much of my statistical work has been done in my business classes while much of my IST coursework has been devoted to data cleanliness and learning models. An example of some of my early work can be seen in my project for MBC 638. For that project, students were individually tasked with developing a business or business-like problem. At the time, I was interested in using new skills to improve efficiency at my current place of employment. For that purpose, I formulated a process improvement problem where we took our current process for responding to floor defects, of which there were many undefined decision points, and attempted to improve them. First the response time was measured with no change, and then definitions for some of the decision points were implemented and the response times were measured for subsequent defects. Due to the current efficiency of the manufacturing process, defects were uncommon. As such, the data set only contains around 13 data points accumulated over a three month span.

**In progress notes**

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

**Collect and organize data**

The final project I did for IST 652

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

Text Mining

IST 707

IST 719

**Develop alternative strategies based on the data**

MAR 653

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

IST 719, MAR 653, IST 657, MBC 638

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

Multiple final projects

**Synthesize the ethical dimensions of data science practice**

Twitter data. Quora data, IST 657

IST 657 – SQL

Final project – Workplace Database storing production records(work)

MBC 638 – Data Analysis and Decision Making

Process improvement analysis on Defect response process(work)

IST 687 – Intro to Data Science

NBA player data analysis Predicting MVP or Played in Finals

SCM 651 – Business Analytics

No final project but four smaller HW Assignments

IST 707 Data Analytics

Quora Insincerity Challenge. Text mining classification

MAR 653 Marketing Analytics

Linear Regression – Housing Market, data set for Ames, IA

IST 652 – Scripting for Data Science

Scraping song lyrics and Wikipedia articles. Text mining classification

IST 736 – Text Mining

Tweets regarding music festivals. Text mining. no machine learning.

IST 719 – Information Visualization

E-sports. Twitch API connection and pre-cleaned. tsv’s. Visualization only no ML

IST 718 – Big Data

Tweets regarding 2016 election. Classification problem, real twitter user or Troll twitter user. Text Mining