Name: Teng Siong (T.S) Yeap
Project Portfolio Milestone

Overview of the Major Practice Areas in Data Science

According to Wikipedia, Data Science is "a multidisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured." In other words, to gain insights to solve a business problem, one needs to master the skills required to extract the data. Often time, Data Science and Data Analytics terms are used intertangle. The workflow for both is about the same. According to Big Data Analytics, to extract information from data, we need to utilize several steps include data collection, preparation, analysis, visualization, management, and preservation. The process is also known as OSEMiN (O'Neil & Schutt (2013), the abbreviation of Obtain, Scrub, Explore, Model and Interpret. I used this approach in most of my projects or assignments in this program. There are 4 main learning goals in this program:

1. Data collection: using tools to collect and organize data
2. Data analysis: identify patterns in the data via visualization, statistical analysis, and data mining
3. Strategy and decisions: develop alternative strategies based on the data
4. Implementation: develop a plan of action to implement the business decisions

In this paper, I will demonstrate how I apply each learning into my projects/assignment.

Deliverable 1: 2008 Arrival Delay Analysis: American Airlines vs. Everyone Else

Deliverable Overview:

This project was done in the IST 687: Introduction to Data Science. The project was about Airline Delay in 2018. My other teammates were Brandon Croarkin and Michelle Mak. We had a pretty good hands-on experience in this very first data science project, under the supervision of our instructor, Professor Corey Jackson.

Name: Teng Siong (T.S) Yeap
Project Portfolio Milestone

Data collection:

The dataset was obtained from Kaggle. The link was included in the GitHub page. Some major metadata found in the dataset were Month, Day of Month, Day of Week, Unique Carrier. Arrival Delay and others.

Data Analysis:

In order to give our analysis some context and a goal, we framed our questions through the perspective of American Airlines(AA). We used AA's data and compared it to the industry average in order to find out where and how a specific airline could improve. Some retrospective questions that we had were:

1. How many delays did each airline experience in 2008?/ Which airlines experience the most delays?
2. Which airline experiences the least delays?
3. What is the average length of a delay?
4. Which airport has the most delays?

The main research questions that we had were:

1. What factors can we use to predict delays?
2. Can we predict the length of a delay?

We applied linear and logistic regression to build our predictive model.

Strategy and decisions:

Unfortunately, we were unable to predict airline delays. Airline delays have a large degree of uncertainty applied to them and without additional data such as weather or diagnostic reports of the plane, it is hard to make accurate predictions -- especially since human behavior is such a

large component of delays as well. There is not much we can recommend based on our analysis since we do not have other dataset to support the prediction of a delay. We, however, do have some insights to share. For example, Southwest Airline is one of the airlines that has the least delays compared to other like airlines. Additionally, the data shows that a delay is very likely to happen when traveling on Friday, so maybe plan accordingly for your next long weekend. Furthermore, plan for more travel time in December as delays often occurs due to holiday/vacation season. You do not want to miss celebrating a holiday with your family due to your flight delays!

Implementation:

For future work, we would like to add some datasets, for example weather dataset, that may be helpful in the analysis. Besides, our instructor, Prof. Corey Jackson pointed out that we should take the size and the traffic of the airport into considerations when showing which airport had the most delay.

Deliverable 2: Investigating the Case of Jack the Ripper with Data Mining

This project was done in the IST 565: Data Mining. The project was about Jack the Ripper. My other teammate was Audrey Crockett.

Data Collection:

We began our data collection by acquiring the texts from the original Jack the Ripper. Then, we converted the original Jack the Ripper letters into individual text files. Next, we researched prominent suspects in the Jack the Ripper case. We had to rule out some suspects due to lack of accessible writings. The suspect pool for our experiment included six suspects; Joe Barnett, Lewis Carroll, Prince Albert, Carl Feigenbaum, Mary Pearcey, and Walter Richard Sickert

(Ryder, S. P., Johno, & Schachner, T., 2013). All the suspects have at one time or another in the 130 years since Jack the Ripper slayings been implicated as the famous murderer. Once we identified the suspect pool, we then acquired writing and quotes by these suspects. Our data set includes writings, testimonies, or quotes from each of the six suspects. All suspect primary source documents were divided into individual text files. R was the employed to aid in the data wrangling of the text files. This required a few packages in R, "tidytext", "readtext", and "tidyverse". Using "readtext", text files can be read in and formatted. Then using "tidytext" and the "tidyverse", this allowed for the manipulation of the data into word frequencies (Seigel, J. & Robinson, D., 2017). Once the data frame was set into a useable format, the data was transformed using the min/max transformation.

Data Analysis:

Wordcloud analysis - The most frequently used word is "ha". This makes sense with what we know about Jack the Ripper, who frequently like to taunt police over not being able to catch him (Ryder, S. P., Johno, & Schachner, T., 2013).

We also used K-means clustering, decision tree and SVM to answer our research question: Who is Jack the Ripper from the suspect pool.

Strategy and decisions:

Since all the models did not come to a consensus, it is difficult to say beyond a reasonable doubt who Jack the Ripper really was. This is because we had not much time and resources. It was challenging to find the letters of the suspects. There were 30 suspects at the time. However, we only managed to locate the letters of 6 out of 30 suspects.

Name: Teng Siong (T.S) Yeap
Project Portfolio Milestone

Implementation:

I would love to collect any text related documents on all 30 suspects at the time in the future work.

Deliverable 3: Case Study - Future Coach Selection Strategy

This assignment was done in the IST 718: Big Data Analytics. This assignment used several datasets from different sources and predict the future coach selection strategy.

Data Collection:

There were 4 datasets in this assignment: Coaches, Stadium Size, Graduation Rate and stadium2017(win loss ratio), which was the win loss ratio. The datasets were provided by the instructor on GitHub and the https://www.ncaa.com/ website. The main challenge in this assignment was how to merge all datasets into a single data frame.

Data Analysis:

The research question in this assignment were

1. What is the recommended salary for the Syracuse football coach?
2. What would his salary be if we were still in the Big East? What if we went to the Big Ten?
3. What schools did we drop from our data, and why?
4. What effect does graduation rate have on the projected salary?

To answer these questions, I used OLS regression and answered the research questions based on formula from the regression, the p-value of each variables, and coefficients.

Name: Teng Siong (T.S) Yeap
Project Portfolio Milestone

Strategy and decisions:

It seemed like coach salary was positively related by stadium capacity.

Implementation:

I recommended 2 million dollars for the next Syracuse football coach. Syracuse University might not have the biggest stadium, but it had high graduation rate.

Summary:

I feel comfortable to demonstrate the main learning goals stated above. I also included two assignments to show my competency in information visualization using "gganimate" and "shiny" package in R. There is one learning goal which is not listed above: the ethical dimension of data science practice. I did not have any experience in handling datasets which contained confidential information. Most of the datasets were obtained from Kaggle. The confidential information had been hidden or replaced by generic labels before they were shared to the public.

My GitHub link which include my deliverables above:

https://github.com/tsyeap88/Portfolio

Reference:

1. https://en.wikipedia.org/wiki/Data_science