**Emma Gillen**

**ekgillen@syr.edu**

**544448449**

**Applied Data Science Portfolio**

**Spring 2021**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **0 Introduction** |  | 3 |
| **1 Project Overviews** |  |  |
| 1.1 IST 623 - Information Security |  | 4 |
| 1.2 IST 719 - Information Visualization |  | 6 |
| 1.3 IST 736 - Text Mining |  | 9 |
| 1.4 MBC 638 - Data Analysis and Problem Solving |  | 13 |
| **2 Learning Objectives** |  |  |
| 2.1 Data Science Overview |  | 16 |
| 2.2 Collection and Organization |  | 17 |
| 2.3 Pattern Identification |  | 18 |
| 2.4 Alternative Strategy Development |  | 18 |
| 2.5 Plan of Action Development |  | 19 |
| 2.6 Communication |  | 20 |
| 2.7 Ethics of Data Science |  | 20 |
| **3 Works Cited** |  | 22 |
|  |  |  |

**Introduction**

This paper will be divided into two parts. In the first part, a high level overview of a selection of projects will be explained. In order to complete the Applied Data Science masters program through Syracuse University, 12 courses were required. A selection of five of those courses will be discussed in this paper. In the second part, these selected projects will be used to illustrate the completion of the seven goals of the program. The overall goal is that a student completing the program will be able to meet the following objectives:

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).

**1**

**Project Overviews**

**1.1 IST 623 - Information Security**

The final output of this course was a case study exploring the Atlanta ransomware attack in March 2018. This attack was orchestrated by the SamSam group who demanded a Bitcoin ransom in exchange for a key to decrypt approximately 3,789 computers. The attack temporarily plunged the city into cyber darkness and destabilized the municipal operation.

The attack gained access to these computers using vulnerabilities in the Java-based web applications and servers, remote desktop protocols (RDP), and file transfer protocol (FTP) servers to gain access to the victims’ machines. After gaining access the hackers were able to escalate their own privileges to those of an administrator. They then dropped malware on the server. Analysis of the victims’ networks indicated that the cyber actors had purchased several stolen RDP credentials from darknet marketplaces. In contrast to many ransomware attacks, this did not rely on a victim to complete an action like opening an email. By exploiting RDP, they were able to infect victims with minimal detection.

If the victims chose to pay the ransom, they were able to contact SamSam through a domain that was only accessible using a TOR browser. After paying, they would receive a link to download the decryption keys. It was unclear in the Atlanta attack if the ransom was paid. The United States Government does not encourage paying ransoms. Paying a ransom does not guarantee that an organization will regain access to their data and many victims have been targeted again. However, the decision whether or not to pay a ransom requires evaluation of all options to protect shareholders, employees, and customers.

The attack highlighted more than 2,000 major vulnerabilities in the Atlanta municipal system. An audit performed three months prior to the attack in January 2018 showed that there were known issues before the attack. A number of recommendations were made in the audit that had not been put in place. These vulnerabilities included missing or outdated policies, inconsistencies in definitions of scope, and lack of formal risk management when using third party services. Following the attack, Atlanta hired a new CIO that planned to establish a new cybersecurity framework and put a renewed focus on awareness. In addition, many of the city’s critical applications were moved to hybrid cloud services which improved security.

The attack caused millions of dollars in damages for the city of Atlanta. Estimates ranged from $7 million to $17 million. In addition to monetary damages, the attack also led to lost productivity when employees of the city were unable to use their computers for five days. The attack on Atlanta is considered the largest, most expensive cyber disruption in city government to date. The SamSam group has targeted many other cities and health care entities, collecting roughly $6 million in ransom over a three year period. They caused harm to more than 200 victims across the country and caused more than $30 million in losses.

**1.2 IST 719 - Information Visualization**

This course focused on creating visualizations to better convey data. The final output of this course was an information visualization poster featuring data from Goodreads. The Goodreads website contains entries on thousands of books as well as user reviews and ratings. Based on this user feedback, Goodreads provides recommendations. The motivation in creating this poster was to explore the relationship between popularity and quality. The plots were created in R and the final poster was completed in Adobe Illustrator. The final poster is included as figure 1.2.1. The individual visualizations will be explained in the figures that follow.

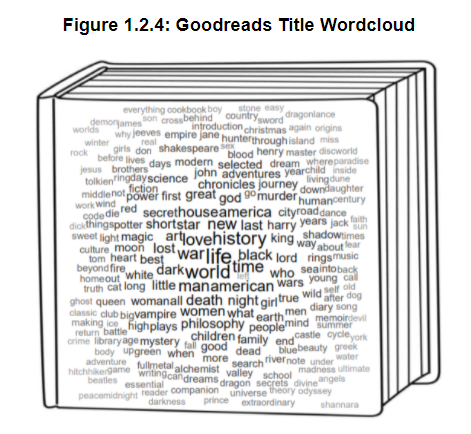
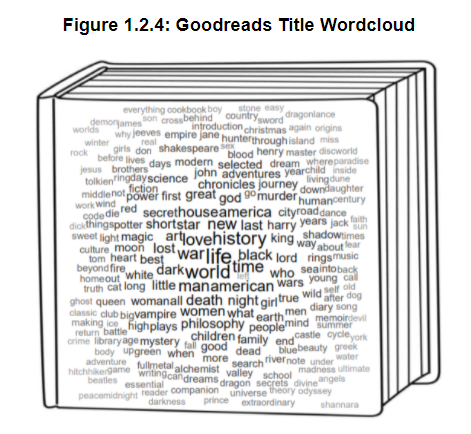
|  |
| --- |
| **Figure 1.2.1: Goodreads Complete Poster** |
|  |

A scatterplot was created with the number of ratings in thousands on the y-axis and the number of text reviews in thousands on the x-axis. This plot was the key visual in this poster and as such was placed in the upper left corner where the reader's eye would first be drawn. This plot is included as figure 1.2.2. This plot visualizes the relationship between the two forms of user feedback. In addition to showing a general upward trend, there are six points called out with their specific book information. These serve to highlight significant outliers as well as providing context to the plot. In addition to the basic scatterplot, a color gradient from red to blue was used to indicate the average book rating.

|  |
| --- |
| **Figure 1.2.2: Goodreads Review & Rating Counts by Score** |
|  |

Two bar plots were generated showing the top fifteen most rated authors and titles. These plots are displayed in the upper righthand corner of the poster and included in this paper as figure 1.2.3. In the plot to the right, the length of the bars indicate the average ratings for each title. In the plot to the left, the length of the bars indicate the average ratings for titles aggregated by author. A dashed line is used to indicate the median rating in both plots. The colors of the bars are used to associate books with the same author or book-to-author relationships.

|  |  |
| --- | --- |
| **Figure 1.2.3: Goodreads Top Authors and Titles** | |
|  |  |

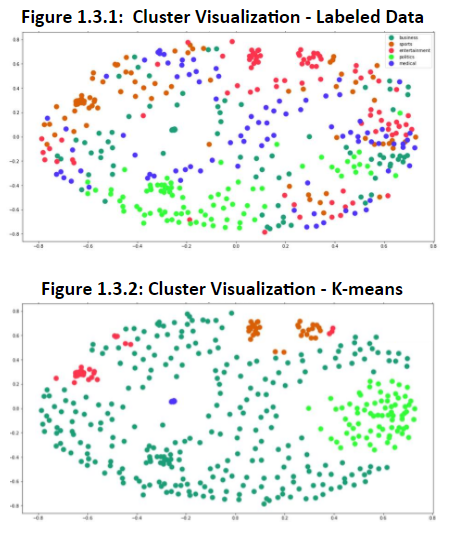
A word cloud was generated using the most commonly used words in the book titles. This plot was displayed in the lower left hand corner of the original poster and is included as figure 1.2.4 to the right. The word cloud was used to give a high level overview of the selection of books in the Goodreads dataset. The words displayed are filtered of any function words that do not add context to the titles. 

The lower right corner of the poster features three separate plots that are included in this paper as figure 1.2.5. The larger visualization to the left is a scatterplot, displaying the number of pages in a book on the y-axis and the average rating on the x-axis. The median lines for both of these dimensions are displayed using dashed lines. The median number of book pages is closer to the x-axis. To help understand this dimension, a histogram of book lengths is displayed to highlight the outliers toward longer books. On the scatterplot, red points are used to call out titles that are boxed sets, collections, or completed works. These tended to be longer in length and higher in rating. To better understand this, a density plot of the average rating is also displayed. In addition to showing the median for the entire dataset, two dashed lines are used to illustrate the median ratings for books over 500 and 1000 pages in length.

|  |
| --- |
| **Figure 1.2.5: Goodreads Book Length** |
|

**1.3 IST 736 - Text Mining**

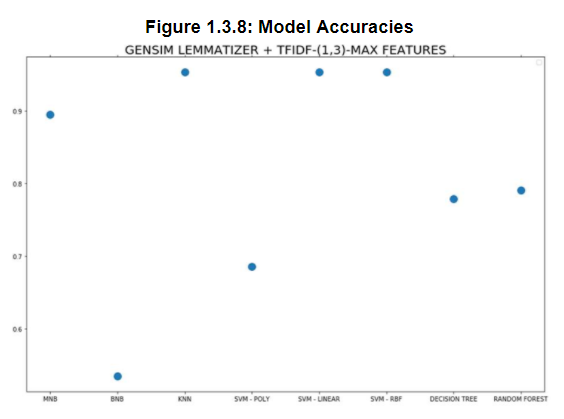
The final output of this course was an exploration of news articles. The goal was to create a process to consume media devoid of bias. This would be done by gathering articles from various sources with varying affiliations and perspectives and producing only a high level overview of the contents of all the articles. A total of 429 articles were collected and manually labeled as belonging to one of five categories: medical, politics, business, sports, and entertainment. When collecting news articles, it was apparent that the data was heavily skewed toward the United States presidential election and COVID-19. To account for these issues, additional articles were collected from specialized media sources.

The text for each article was scraped using Python. The texts were then filtered of punctuation as well as a selection of words that did not provide context. The dimensionality was further reduced using either the Gensim lemmatizer or the Porter stemmer. The stemming technique reduces words to their root form while the lemming technique combines words with similar meanings. With the texts cleaned, they were vectorized for modeling. Vectorization translates a text corpus into a dataframe where each row represents an article and each entry represents a word. Each entry represents the usage of each word in each document. In this experiment, binary and TFIDF vectorization were used. The binary approach indicates if each word appears or does not appear in a document. TFIDF computes the term frequency-inverse document frequency which accounts for the importance of each word to each document. 

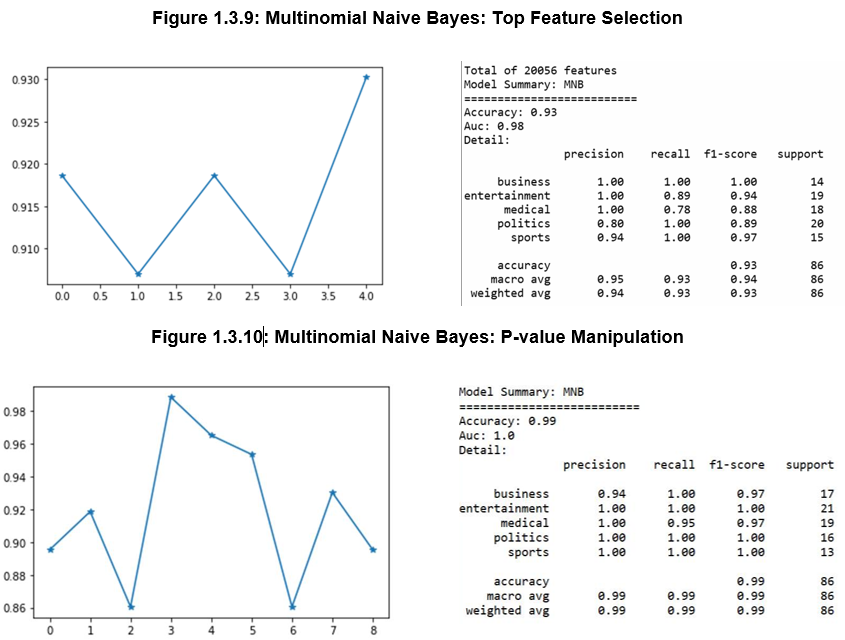
To confirm the natural separation of the dataset, k-means clustering was performed. In figure 1.3.1, a visualization of the clustering of the labeled data is displayed. In figure 1.3.2, a visualization of the k-means clustering is displayed.

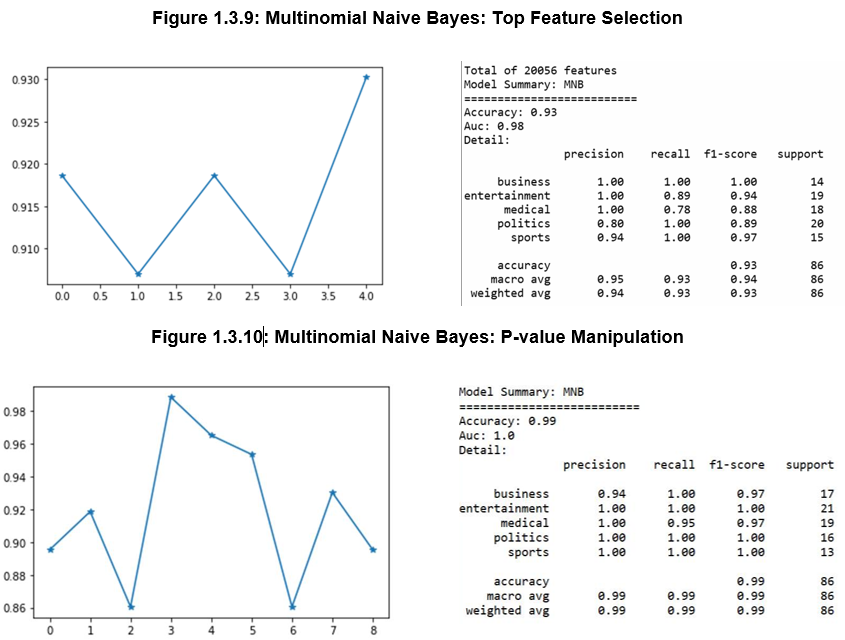
To get a better idea of the contents of each manually labeled cluster, a group of word clouds were generated (figures 1.3.3 - 1.3.7). In addition to describing the contents of the training corpus, these word clouds give an idea of the type of results that can be generated for future applications. When an accurate model is identified, new articles may be gathered and labeled programmatically. Then, word clouds such as these may be produced using the new, machine-labeled corpora.

|  |  |
| --- | --- |
| **Figure 1.3.3: Medical Word Cloud** | **Figure 1.3.4: Entertainment Word Cloud** |
|
| **Figure 1.3.5: Politics Word Cloud** | **Figure 1.3.6: Business Word Cloud** |
|
| **Figure 1.3.7: Sports Word Cloud** | |
|

To predict the topic of each article, train test splits of the labeled corpora are created. The training sets are used to produce a series of models. For this research, Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), k-Nearest Neighbors (kNN), Support Vector Machines (SVM) with various kernels, decision trees, and random forests are produced.

In total, 32 combinations of models and corpora were produced. These were made up of combinations of the eight model types and the four vectorized text combinations. The best results were found when using the Gensim lemmatizer over the Porter stemmer. In addition, the TFIDF vectorizer performed better than the binary vectorization. The accuracies for the eight models produced using the combination of Gensim lemmatization and TFIDF vectorization are summarized in figure 1.3.8.

The best results were found when using the linear kernel SVM, radial kernel SVM, or kNN. The Multinomial Naive Bayes model followed closely behind in both cases. While the SVM and kNN models had great results, the Multinomial Naive Bayes model is better suited to this task for two reasons: runtime complexity and feature ranking. The goal of this experiment was to produce a model to generalize future, unlabeled data. The time taken to run the SVM and kNN models many times would be prohibitive. In addition, the feature ranking provided by Multinomial Naive Bayes is better suited for this task. The model will highlight which tokens are most indicative of each news topic. As the goal of this task is to understand the contents of each topic, feature ranks will be very useful.

After selecting MNB as the best algorithm for this task, feature selection was explored. A series of MNB models were trained on different subsections of the original data. The top 1000, 2000, 3000, 4000, and 5000 features for each label were used. With this process, the highest accuracy was 93% when using a total of 20,000 features (figure 1.3.9). For a second attempt at optimization, chi square feature selection was utilized. With this method, the highest accuracy was 99% when using a p value of 0.4 (figure 1.3.10). The Naive Bayes model fit to a 99% accuracy is likely overfitted to our data. However, the hyperfit feature selection is very indicative of the high level summary of each topic. Below, the top features from the improved Multinomial Naive Bayes model are displayed for each topic. 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Medical** | **Politics** | **Entertainment** | **Business** | **Sports** |
| vaccine  health  care  study  testing  disease  medicine  test  drug  blood | trump  election  president  republican  senate  fraud  campaign  court  presidential  trump campaign | entertainment  actor  movie  music  perry  check fashion  check fashion entertainment  division re  division re serve  entertainment gossip | stock  company  investment  dividend  oracle  business  analyst  homology  price target  income | coach  season  fantasy  team  football  sport  fantasy football  game  player  fantasy football expert |
|
|
|
|
|
|
|
|
|

Overall, this experiment produced extremely promising results. This method was able to correctly determine a topic in over 90% of cases and was able to generate high level key words surrounding the topics. The table below shows the words that were determined as important to each topic for the articles used in the experiment.

**1.4 MBC 638 - Data Analysis and Problem Solving**

One of the focuses of this course was understanding the process of DMAIC. DMAIC is made up of the phases define, measure, analyze, improve, and control. The final output of this course was a process improvement project.

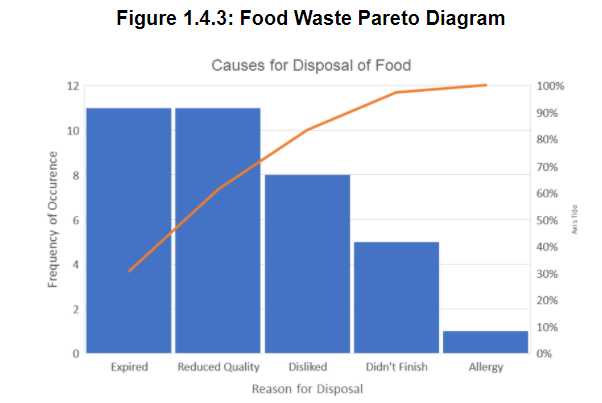
Throughout the term, I collected data about my own household’s food waste. My goal for this process was to reduce the dollar value of what is wasted. By improving the process, I wanted to cut the amount of money wasted in half. I estimated that I throw away about 5% of the food I purchase (including groceries, carry in, dining out, coffee, alcohol, etc.). In addition, I estimated that my household spends around $200 a week on food. With these estimates, I was losing approximately $10 a week or $520 a year.

|  |
| --- |
| **Figure 1.4.1: Process Improvement Poster** |
|  |

Above, the final poster is displayed. Through this process, I was able to identify and implement two changes to my process. The original process and the process following this improvement are both displayed using process flow charts. These are shown below in figure 1.4.2. By implementing this change, the sigma quality level (SQL) improved from 3.35 to 3.75. The SQL is a quantitative measure of the capability of the process.

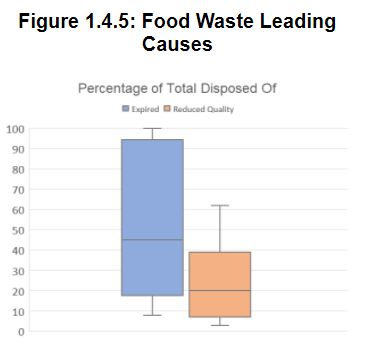
|  |  |
| --- | --- |
| **Figure 1.4.2: Process Flow Chart** | |
| **Before** | **After Improvement Process** |
|  |  |

As shown in the process flow charts above, the two changes implemented were: creating a household rule to not purchase an item if we have more than 20% of it remaining, and increasing efforts to preserve foods before they expire. These areas of improvement were identified using a variety of statistical methods and visualizations.

In order to understand the reason why food was being disposed of, the reason was recorded for each observation as belonging to one of five predetermined scenarios. In figure 1.4.3 to the right, a pareto chart displaying these causes for food disposal are displayed. The Pareto chart indicated that 80% of the food being disposed of was caused by just two of the five scenarios. These two scenarios are foods that expire or reduce in quality before they are able to be consumed. 

Next, the relationship between the cycle time and the dollar value lost was explored. The cycle length was defined as the amount of time between an items purchase and its disposal. The dollar value lost was calculated using the original cost of the product, multiplied by the percentage disposed of. A scatterplot was created for these variables. Visually, there was indication that there may be a positive relationship between the two variables. However, when a linear regression was run, the model was lacking. The first regression had an R value of 0.648. As a general rule of thumb, an R value needs to be more extreme than 0.7 to be significant. For this reason, the focus remained on the reason for disposal as an area of improvement.

|  |  |
| --- | --- |
| **Figure 1.4.4: Food Waste Regression** | |
|  |  |

The focus returned to investigating the causes of food waste. According to the pareto chart, the majority of food waste was caused by expired or reduced quality food. Boxplots showing the percentage of food disposed of for each of these causes was created. These plots indicated that reduced quality goods tend to be mostly used up by the time they are thrown away. In many cases, the reduced quality goods are replaced before they fully expire but are otherwise lacking in some way. For this reason, the household rule was implemented that a product would not be replaced if greater than 20% reamined. For expired goods, the percentage thrown away has a much larger spread. To address this, there were increased efforts to preserve foods before their expiration. These two recommendations were illustrated in the process flow chart in figure 1.4.2. 

**2**

**Learning Objectives**

**2.1 Data Science Overview**

**Learning Objective: Describe a broad overview of the major practice areas in data science.**

The field of data science can be applied to endless areas of application. The connecting focus between these areas is the collection, cleaning, and processing of data. Once the data is gathered, it may be analyzed through techniques of data visualization, data modeling, and machine learning. Data visualization focuses on the presentation of data in a clear and understandable visual format. Information Visualization (IST 719) in particular focused on this component of data science. Plots can be produced directly from the programming languages, but often other visualization platforms may be utilized like Tableau or Google Analytics. In IST 719, plots from R were improved upon with Adobe Illustrator. The focus went beyond simple plots and focused on creating dynamic and usable visuals.

Data modeling and machine learning are a closely linked aspect of data science. Data modeling aims to create a statistical model which can replicate or predict the behavior of the data. Machine learning takes data modeling a step further, by training a computer to improve the model automatically. This area of focus attempts to replicate human understanding, but at a scale infeasible for a human to compute. Models can be either supervised or unsupervised. Supervised models aim to classify data based on existing labels. Examples of supervised models include linear or logistic regression, support vector machines, and decision trees. Unsupervised models cluster the data without known labels. Examples of unsupervised algorithms are k-means clustering and association rule mining.

As part Text Mining (IST 736), a variety of models were produced. These included both unsupervised models to find the underlying clustering of the training set and supervised models to predict the newspaper categories. In this case, the best choice of supervised model was selected by considering the accuracy, the time complexity, and the type of output produced. These factors all need to be taken into consideration by a data scientist in each application.

**2.2 Collection and Organization**

**Learning Objective: Collect and organize data.**

Before any analysis can take place, data needs to be collected and prepared. Data can be gathered from a variety of sources including transaction logs, records, raw text, etc. Data can be quantitative or qualitative. Quantitative data involves numerical values and can be either discrete or continuous. Discrete data is limited to whole numbers such as a count whereas continuous data can be manipulated with fractionally. Qualitative data deals nonnumeric characteristics like gender, color, or rating. Qualitative data can be nominal, meaning it has no natural ordering, or ordinal, meaning it does have a natural order. Some additional data cleaning steps include removing or otherwise handling missing data, removing any unnecessary data, and correcting data structuring. In R and Python, data is most often utilized in the form of a dataframe. A dataframe contains data in the form of rows and columns. When handling data, it is important to ensure that the correct data types are used and the correct models are utilized. Different models may be limited to different data types.

Additionally, the collection and processing steps necessary vary by task. In Text Mining (IST 736), APIs were utilized to gather text data from various websites. Text data needs to be stripped of various characters before it can be used. Text also needs to be transformed into a computer understandable format before it can be used in modeling. This process is called vectorization, where a vector represents a single text, and each entry represents the occurrence of a word in the text. In some cases, data needs to be collected. This was part of the process improvement project in Data Analysis and Problem Solving (MBC 638). When collecting data, it is important to be very rigorous in order to ensure accuracy. It is also important to clearly define data definitions beforehand to ensure clarity. Collecting data has the advantage of not being limited by the original data collector’s methods. Once data is collected, there are limited ways to increase the rigor of data. For example, if data is measured on an hourly basis, only so much can be done to interpolate between samples.

**2.3 Pattern Identification**

**Learning Objective: Identify patterns in data via visualization, statistical analysis, and data mining.**

One of the primary goals of analyzing data is to identify the underlying patterns. Summary statistics are often the first step in this process. These statistics include measures of central tendency (mean, median, mode) and measures of distribution (range, standard deviation, variance). Plots are often generated to visually understand these statistics. Box plots, histograms, and density plots can be used to visualize skew and kurtosis in the data. These characteristics can indicate if the data has a regular distribution or identify potential outliers. For example, in Information Visualization (IST 719), a barplot of the ratings was created. The barplot and the kurtosis showed that although the data ranged from 1 to 5, the large majority of data points fell between 3.5 and 4.5. Scatterplots may be utilized to identify trends or signify outliers. They can also be used to visualize clustering and identify separate populations within the data. In Information Visualization (IST 719), two scatterplots were created that utilize not only the x-axis and y-axis but also colors and median lines to further illustrate the underlying patterns in the data. Clustering can be further understood with techniques like k-means clustering. This technique was performed in Text Mining (IST 736). Time series plots may illustrate patterns of change over time or seasonality.

**2.4 Alternative Strategy Development**

**Learning Objective: Develop alternative strategies based on the data.**

The goal of data science is not only to understand the data, but to draw actionable recommendations from it. Examples of actionable recommendations can include areas of interest to pursue, recommendations for features to reduce or remove, and proposals for changes of action. In Data Analysis and Problem Solving (MBC 638), a process improvement plan was implemented. The original process was identified and data was recorded. Following the data analysis steps, two areas in the original process were identified where the majority of errors occurred. Looking at the generated Pareto chart, 80% of errors occurred within those two situations. These errors were the disposal of expired or reduced quality food. A number of potential alternative strategies were available. A possible option to reduce all waste would be to cease purchasing food. While this would solve the issue, it is entirely infeasible. As a data scientist, it is important to understand the constraints associated with any alternative strategy and how they apply to the application at hand. Other areas of exploration were purchasing smaller portions, meal prep, or only purchasing prepared food as required. After considering the alternative strategies, the best choices strategies were identified as purchasing less food and making efforts to extend the time available before expiration.

**2.5 Plan of Action Development**

**Learning Objective: Develop a plan of action to implement the business decisions derived from the analyses.**

Stemming from the identification of alternative strategies, a plan of action must be created to implement changes. These changes not only need to be proposed on a hypothetical level, but enacted. While strategies may be identified that would greatly improve the process, factors like budget, time, and ethics need to be considered in the plan of action. In the case of the process improvement plan in Data Analysis and Problem Solving (MBC 638), the identified strategies were buying less food and throwing away less spoiled food. In order to enact these strategies, two rules were added to the original process. These rules were that a product would not be replaced if greater than 20% remained and increased efforts would be made to preserve foods before their expiration. These alternative strategies took place in different stages in the process. The first, aimed to reduce waste by preventing it from entering the process in the first place. The second, aimed to add preventative measures to decrease errors during the product life cycle. As part of the plan of action, the actors would have to be identified and informed at these stages in the cycle. In a real world setting, the process improvement plan may need to be completed multiple times, identifying areas of growth, making changes, and assessing the need for further changes in an iterative way.

**2.6 Communication**

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

In real world situations, data scientists will very rarely be the only members of a team. Instead, groups will include people with varying degrees of technical knowledge. In order to effectively translate findings to all members of a team, different strategies may have to be employed. As part of Text Mining (IST 736), an emphasis was placed on the writing of a non-technical introduction to the problem and conclusion. These writings needed to be readable by team members with any level of technical ability. This practice allows managers or similar roles to understand the high level business problem and proposed solutions without being overwhelmed with the technical aspects of the analysis and modeling. Another approach to better conveying information to nontechnical viewers is through visualizations. In Information Visualization (IST 619), the goal was to have interesting plots that did not require a large amount of technical knowledge to understand. Aspects like clear and legible fonts as well as effective layouts were considered. In the final plot, the most important visualization was displayed in the upper left of the poster, as most American readers will be drawn to that area first. The goal in creating the poster was to make the information as accessible as possible to any reader.

**2.7 Ethics of Data Science**

**Learning Objective: Synthesize the ethical dimensions of data science practice (e.g., privacy).**

As a data science professional, it is important to consider the ethics of any decision. Data needs to be protected for a variety of reasons depending on the specific situation. As a consumer, there are concerns that their sensitive data will fall in the wrong hands. As a company, there are concerns about sensitive industry information being given to competitors. When working with data, it is important to protect sensitive information from these risks. Using encryption is one example of a mitigation technique used to protect data. In addition, using permissions to limit access to only explicitly allowed users can prevent malicious actors from accessing data. Questions also arise about the ownership of data. The question is asked whether companies that provide services or the users of those services should own the data that is produced through their interaction.

Information Security (IST 623) emphasised the importance of understanding ethics in data science. In the completion of the final case study investigating the Atlanta ransomware attack, the huge impact of a cyber attack became clear. In this case, the attack caused millions of dollars in damages and halted many essential city operations. Professionals in Atlanta’s municipal government were aware of these issues and could have potentially prevented the attack.

**3**

**Works Cited**

2018 Atlanta CYBERATTACK. (2021, January 28). Retrieved February 08, 2021, from https://en.wikipedia.org/wiki/2018\_Atlanta\_cyberattack

Agrawal, S. (2020, June 03). What is instance-based and model-based learning? [s1e10]. Retrieved February 08, 2021, from https://medium.com/@sanidhyaagrawal08/what-is-instance-based-and-model-based-learning-s1e10-8e68364ae084

Amy Mitchell, J. (2020, August 17). Many Americans SAY made-up news is a critical problem that needs to be fixed. Retrieved February 08, 2021, from http://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/

Ben.k. (2018, June 11). Text mining preprocess and Naive Bayes Classifier (Python). Retrieved February 08, 2021, from https://medium.com/@baemaek/text-mining-preprocess-and-naive-bayes-classifier-da0000f633b2

CISA insights - Ransomware Outbreak. (n.d.). Retrieved February 08, 2021, from https://www.dhs.gov/blog/2019/08/21/cisa-insights-ransomware-outbreak

Computational complexity of machine learning algorithms. (2018, April 16). Retrieved February 08, 2021, from http://www.thekerneltrip.com/machine/learning/computational-complexity-learning-algorithms/

Cyber daily: Ransomware recovery triggers cultural change in atlanta. (2020). *WSJ Pro.Cyber Security,* Retrieved from https://libezproxy-syr-edu.libezproxy2.syr.edu/login?url=https://www-proquest-com.libezproxy2.syr.edu/trade-journals/cyber-daily-ransomware-recovery-triggers-cultural/docview/2357433761/se-2?accountid=14214

Deere, S. (2018, November 29). Feds: Iranians led CYBERATTACK against Atlanta, other U.S. entities. Retrieved February 08, 2021, from https://www.ajc.com/news/local-govt--politics/feds-iranians-led-cyberattack-against-atlanta-other-entities/xrLAyAwDroBvVGhp9bODyO/

Fruhlinger, J. (2020, February 20). Recent ransomware attacks define the malware's new age. Retrieved February 08, 2021, from https://www.csoonline.com/article/3212260/recent-ransomware-attacks-define-the-malwares-new-age.html

Gensim.utils.lemmatize()¶. (n.d.). Retrieved February 08, 2021, from https://tedboy.github.io/nlps/generated/generated/gensim.utils.lemmatize.html

Kanani, B. (2020, April 30). Cosine similarity - text similarity metric. Retrieved February 08, 2021, from https://studymachinelearning.com/cosine-similarity-text-similarity-metric/

Palmer, D. (2018, November 28). SamSam ransomware created by iranian hackers, says US DOJ. Retrieved February 08, 2021, from https://www.zdnet.com/article/samsam-ransomware-created-by-iranian-hackers-says-us-doj/

Ransomware hits Atlanta Police dashcam footage. (2018, June 07). Retrieved February 08, 2021, from https://www.bbc.com/news/technology-44397482

Ransomware prevention and response for cisos. (2016, July 14). Retrieved February 08, 2021, from https://www.fbi.gov/file-repository/ransomware-prevention-and-response-for-cisos.pdf/view

Robinson, J. (n.d.). Text mining with r: A tidy approach. Retrieved February 08, 2021, from http://www.tidytextmining.com/tfidf.html

Soumik. (2020, March 09). Goodreads-books. Retrieved February 08, 2021, from https://www.kaggle.com/jealousleopard/goodreadsbooks

Two Iranian men indicted for DEPLOYING ransomware to Extort HOSPITALS, municipalities, and public institutions, causing over $30 million in losses. (2018, November 28). Retrieved February 08, 2021, from https://www.justice.gov/opa/pr/two-iranian-men-indicted-deploying-ransomware-extort-hospitals-municipalities-and-public

What can we learn from Atlanta? (n.d.). Retrieved February 08, 2021, from https://www.govtech.com/security/What-Can-We-Learn-from-Atlanta.html