Project Portfolio Milestone

Mikayla Jayne Scott

Syracuse University

September 7, 2023

Table of Contents

Introduction

As a student at Syracuse University School of Information Studies, in collaboration with the Whitman School of Management it is with great pleasure that I present my final portfolio for the Master of Science in Applied Data Science program. During my time at Syracuse I have learned, reflected and evolved as a student and member of the Data Science, Computing and Mathematics fields.

This document describes the program learning outcomes for the degree, provides specific evidence of learning through final projects of the completed courses, and delivers a metacognitive approach to the program.

Program Learning Outcomes

There are several learning outcomes that were meant to be achieved by the end of the course load at Syracuse. Those outcomes are outlined in the list below:

1. Collect, store, and access data by identifying and leveraging applicable technologies.
2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle.
3. Apply visualization and predictive models to help generate actionable insight.
4. Use programming languages such as R and Python to support the generation of actionable insight
5. Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads
6. Apply ethics in the development, use and evaluation of data and predictive models. (e.g., fairness, bias, transparency, privacy)

On the list there are several learning outcomes that are technical. The technical learning outcomes are important; however, the other learning outcomes that focus on real-life application and effective communication are where the heart of the program falls. The technicalities are only just a portion of what it means to be a data scientist, a person can be the greatest data scientist in the world, but if they are unable to communicate the results of their work to others who are not experts in the field, then the work is all for not. This program emphasizes the importance of communication and result delivery. This is achieved through large final projects in each of the courses where students are challenged with data and problems from real world scenarios and applications.

Evidence of Learning

This section provides evidence of learning through some of the final projects that I completed during my time at Syracuse. Each one of these projects helped me to achieve the final program learning outcomes that were listed above. Each of the project subsections will provide the course description and learning outcomes directly from the course's syllabi, describe the project in general, explain my contributions and discuss the final project findings.

Big Data Analytics

Course Description

This course is a broad introduction to analytical processing tools and techniques for information professionals. The overall goal of this course is to develop a portfolio of resources, demonstrations, recipes, and provide examples of various analytical techniques. The course is focused in applying existing knowledge of Python skills to datasets to emphasize experimental, and collaborative learning.

Learning Outcomes

The learning outcomes for Big Data Analytics as listed in the syllabus are listed below:

1. Obtain data and explain data structures and data elements.
2. Scrub data by applying scripting methods, to include debugging, for data manipulation in Python, R or other languages.
3. Explore data by analyzing using qualitative techniques including descriptive statistics, summarization, and visualizations.
4. Model relationships between data using the appropriate analytical methodologies matched to the information and the needs of clients and users.
5. Interpret the data, model, analysis, and findings. Communicate the results in a meaningful way.
6. Select an applicable analytical methodology for real problems in areas such as business, science, and engineering.

Project Description

The project that was completed for Big Data Analytics revolved around the Rolling Stone's "500 Greatest Song's of All Time List". The Rolling Stone is a magazine that is best known around the world for their reporting on music and pop culture. With the first issue being published in 1967, Rolling Stone has stood the test of time, and has reported on music legends from John Lennon to rising stars like Lizzo and Drake. In 2004 Rolling Stone published the first ever "500 Greatest Songs of All Time" list, they followed that list with 26 updated songs in 2010, and most recently released the 2021 version of the list. The new version of the list was created by polling more than 250 musicians, artists, producers, writers in the music industry on their greatest songs of all time. Each of the participants ranked their top 50 songs and Rolling Stone calculated the results. Of the resulting songs there were more than 4,000 songs suggested to be on the list.

As the music industry faces new challenges with the growth of artificial intelligence, music producers and musicians are turning to analytics to make decisions. The purpose of this project was to find patterns and relationships that would help to advise artists and music producers on musical choices that will maximize streams, revenue and popularity.

There were several business questions that the project aimed to answer, those questions are outlined in the list below:

1. Are there any key indicators that would demonstrate to a producer or record company that a song has a greater chance at becoming a hit?

2. What are the relationships between the songs, specifically their lyrics, on the Greatest List? What songs are most closely related?
3. What are the most popular topics of songs in the list?
4. What song characteristics influence the popularity of a song?

Contributions

For this project I was involved in several pieces that were integral to the completion of the project. Starting with data mining and preparation, I found the initial dataset, and then wrote the algorithm to mine the lyrics for each of the songs. Once the lyrics were mined, the dataset was in a place where it was able to move on to the cleaning and preprocessing stage. I cleaned the lyrics and merged (left joined) the datasets so that there was one large cohesive set for the rest of the team to use for the duration of the project. Moving on to the analysis portion, I completed the initial EDA of the dataset, providing a baseline analysis of the data that was collected in the previous stage. This included general metrics and graphics such as the distribution of numeric variables across the dataset, in addition to the sentiment analysis of the lyrics.

For analysis portion of the project itself, my main contribution was focused on answering question number two outlined above. To answer this question, I implemented semantic search technologies. Using cosine similarity to measure the similarity between the lyrics of each of the songs, relationships were defined and then mapped using network graphs to show the strength of connections between each of the songs.

Project Findings

Where the original hope was to provide answers to questions through exploratory data analysis, and different modeling techniques, the results provided insights, but were lacking in evidence that could support full conclusions for each of the questions. The first and fourth questions were addressed through the logistic regression model. The results showed that 'instrumentalness', 'liveness', and 'duration_ms' features are statistically significant predictors of song popularity. Question two was answered via the semantic search portion of the study. There were relationships that were defined between each one of the songs using the embedded versions of the lyrics. The songs that were most closely related often came from similar genres of music, or similar time periods. Question three, was addressed through topic modeling. This portion of the study confirmed that Topic Modeling may be a good tool for further development. It showed various topics such as Wishful Thinking, Raw Emotions and Emotional Resilience among the top topics across the Top 500 list.

Additional Information

Final Project Grade: 97%
Link in Portfolio: https://github.com/mikayla-j-scott/scottMikaylaMADSSyracuse/blob/main/IST718/Final_Paper_IST718%20(1).pdf
Link to Syllabus: https://github.com/mikayla-j-scott/scottMikaylaMADSSyracuse/blob/main/IST718/ist_718_syllabus.pdf

Cloud Management

Course Description

The focus for this course is on cloud computing from a user perspective, that is, organizations that utilize cloud computing for their information infrastructure, application development, data analytics, customer service, and other processes. Students who complete the course will be able to contribute to cloud decision-making, deployment, and management in any organization. The course will utilize readings, lectures, case studies, labs with widely used cloud platforms, and a final project.

Learning Outcomes

The learning outcomes for Cloud Management as shown in the syllabus are listed below:

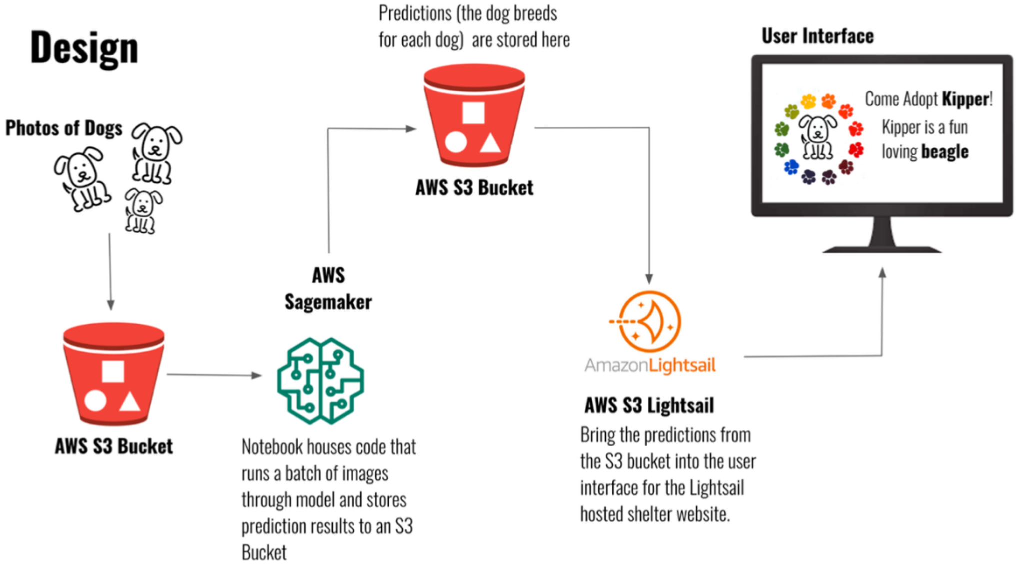
1. Understand and apply cloud concepts and terminology
    a. Understand the economics of cloud computing and the financial implications of cloud computing decisions
    b. Identify fundamental technologies that support and enable cloud computing
2. Make management-level decisions about cloud adoption
    a. Analyze the benefits and risks of cloud adoption in general and as they relate to a specific organizational context
    b. Assess accurately how cloud management approaches, technology, and policy may affect organizations, users, and information systems
    c. Decide whether to adopt the cloud, and what deployment and service models are most appropriate
    d. Critically assess potential cloud services and cloud providers
3. Effectively implement the cloud deployment process
    a. Map the critical steps and decisions involved in cloud deployment and migration
    b. Identify and manage privacy and security risks associated with the cloud
    c. Identify and analyze policy and sustainability issues around cloud computing and how they affect providers, users, and other stakeholders
4. Create and deploy cloud-based applications on major platforms

Project Description

The project for this course was an experimental application of AWS cloud service applications that could be used by animal shelters. The proposed solution aimed to utilize an existing breed prediction machine learning model in conjunction with Amazon S3 Buckets, SageMaker, and LightSail. The idea behind this project was to demonstrate knowledge of cloud management providers, describe a scenario in which a company or organization could benefit from the implementation of a cloud based solution and then communicate the findings through a memo structured proposal.

Contributions

Since there were not any technical pieces to this project, all of the contributions came in the form of research, theory and design. From the beginning the team and I framed a problem that animal shelters face, finding the breed of the dogs within the shelter. From there we were challenged with finding a way in which a cloud provider could help this situation. I designed the implementation of the AWS system. Images from dogs inside the animal shelter would be stored in an S3 bucket. Using AWS Sagemaker a scheduled notebook would run a pretrained image recognition model on the images that were stored in the S3 Bucket, then store the prediction results (the breed of the dog) back into an S3 Bucket. The last portion of the process was to find a way to deliver the results to the end user, the person looking to adopt. This was done by utilizing AWS Lightsail, Lightsail would be used to host the shelter website and provide the breed results to the end user. This process can be seen in the image below:


Design
Photos of Dogs
AWS S3 Bucket
AWS Sagemaker
Notebook houses code that runs a batch of images through model and stores prediction results to an S3 Bucket
Predictions (the dog breeds for each dog) are stored here
AWS S3 Bucket
AmazonLightsail
AWS S3 Lightsail
Bring the predictions from the S3 bucket into the user interface for the Lightsail hosted shelter website.
User Interface
Come Adopt Kipper!
Kipper is a fun loving beagle

In addition, I also worked out the cost of implementation, and defended the business case for the proposed solution.

Project Findings
Since this project was to be submitted in the form of a proposal/memo, there was no actual implementation of the system that was designed. The final result from this project yielded all the points necessary for a decision maker to weigh the pros and cons of a cloud-based solution.

Additional Information
Final Project Grade: 98%
Link in Portfolio: https://github.com/mikayla-j-scott/scottMikaylaMADSSyracuse/blob/main/IST615/615_Final_Paper.pdf
Link to Syllabus: https://github.com/mikayla-j-scott/scottMikaylaMADSSyracuse/blob/main/IST615/ist_615_cloud_management_syllabus.pdf

Text Mining

Course Description
The main goal of this course is to increase student awareness of the power of large amounts of text data and computational methods to find patterns in large text corpora. This course is designed as a general introductory- level course for all students who are interested in text mining. Programming skill is preferred but not required in this class. This course will introduce the concepts and methods of text mining technologies rooted from machine learning, natural language processing, and statistics. This course will also showcase the applications of text

mining technologies in information organization and access, business intelligence, social behavior analysis, and digital humanities.

Learning Outcomes

The learning outcomes for Text Mining as shown in the syllabus are listed below:

1. Describe basic concepts and methods in text mining, for example document representation, information extraction, text classification and clustering, and topic modeling
2. Use benchmark corpora, commercial and open-source text analysis and visualization tools to explore interesting patterns
3. Understand conceptually the mechanism of advanced text mining algorithms for information extraction, text classification and clustering, opinion mining, and their applications in real-world problems; and
4. Choose appropriate technologies for specific text analysis tasks and evaluate the benefit and challenges of the chosen technical solution.

Project Description

My final project for Text Mining was a script analysis across all "*The Avengers*" movies. Marvel's "*The Avengers"* is a movie franchise based on the Marvel comics, "*The Avengers"*. The comics, first debuted in 1963, written by Stan Lee are a collection of stories revolving around "Earths Mightiest Heroes" (Lee, 1963). The movie franchise is composed of four movies. The first in the series, "*The Avengers*", released on May 4, 2012, focuses on the origin stories of the team, and follows the team as they work to defeat Thor's brother Loki and regain control of the tesseract. The second movie, "*Age of Ultron*", released on May 1, 2015, the Avengers work to take down Ultron, a sentient artificial intelligence (AI) program written by Tony Stark and accidentally brought to life. The third and fourth movies, "*Infinity War*" and "*End Game*", the Avengers find themselves going head-to-head with Thanos, collecting infinity stones and battling complicated timelines. The Avengers films are among the highest rated movie franchises with an average Rotten Tomatoes score of 89% and grossing over 7.772 billion US dollars in box office sales.

This study focused on the scripts for each of the Avenger movies. In all the movies combined there are 112 unique characters that have lines. There were 4,100 total lines that were used for exploratory data analysis and in the comparison of different classification model approaches. The models for this study attempt to classify both characters and movie titles by using lines from the movie scripts.

Contributions

This project was one that I chose to do by myself, all the efforts made were completed solei by me. For the collection of the scripts, I had to get creative, there was no existing dataset, or corpus in this case, for each of the movies. To get the movie scripts I copied the text from a free movie script website, pasted it into a text document, and from there converted the text from the file into a csv using a semicolon as the delimitator. It worked out that way because the name of the character saying the line was always proceeded by a semicolon and then the line said after that. Once the each of the movies was converted into a csv format, the lines were cleaned for text analysis using standard NLP techniques, removing punctuation, stop words, etc. With the conclusion of the cleaning stage, the text was ready for exploratory data analysis. For this specific case the exploratory data analysis was extensive, there was analysis done on the frequency of character lines, word statistics, sentiment, and the relationships between characters, and the movies in which they appear.

Moving into the modeling phase, the goal was to determine the best classification model to classify lines by the character and movie that they originated from. There were four different kind of classification models that were compared for this study, multinomial naïve bayes (MNB), support vector machine (SVM), decision tree classifier (DT) and K-Means. Each model was trained to either predict a movie, or to predict a character based on the lines that were fed into it. In total there were 29 models trained and tested for this project on top of the original analysis of the text collected.

Project Findings

The accuracy across the models were considerably low; however, after further investigation there were a few things that contributed to the low scores across each of the models. In the movies there are characters that have roughly triple the number of lines that other characters do, even when the dataset is subset into just the main characters. In result, the models misclassify a lot of the lines for lines that belong to the majority class, in a lot of instances this is Tony Stark. Another point of further analysis was to do with the sentiment of each of the characters over time. The movies take place over a seven-year span and fall within the plot line of the first three phases of the Marvel universe. The main characters have their own spin off films and appear in other movies that contribute to the Marvel universe timeline. Those movies ultimately affect the plot line of the Avenger's movies. For better accuracy and reflection of the true sentiment of each of the characters, the dataset should be expanded to include those movies. This could also help with the low accuracy of the classification models, because it would provide more lines per each character.

Additional Information

Final Project Grade: 100%
Link in Portfolio: https://github.com/mikayla-j-scott/scottMikaylaMADSSyracuse/blob/main/IST736/Scott_Mikayla_Final_736%20(1).pdf
Link to Syllabus: https://github.com/mikayla-j-scott/scottMikaylaMADSSyracuse/blob/main/IST736/ist_736_syllabus.pdf

Applied Machine Learning (Data Analytics)

Course Description

This course will introduce popular data mining methods for extracting knowledge from data. The principles and theories of data mining methods will be discussed and will be related to the issues in applying data mining to problems. Students will also acquire hands-on experience using state-of-the-art software to develop data mining solutions to scientific and business problems. The focus of this course is in understanding of data and how to formulate data mining tasks in order to solve problems using the data.

The topics of the course will include the key tasks of data mining, including data preparation, concept description, association rule mining, classification, clustering, evaluation and analysis. Through the exploration of the concepts and techniques of data mining and practical exercises, students will develop skills that can be applied to business, science or other organizational problems.

Learning Outcomes

The learning outcomes for Data Analytics as shown in the syllabus are listed below:

1. Document, analyze, and translate data mining needs into technical designs and solutions
2. Apply data mining concepts, algorithms, and evaluation methods to real-world problems
3. Employ data storytelling and dive into the data, find useful patterns, and articulate what patterns have been found, how they are found, and why they are valuable and trustworthy

Project Description

For the final project in Machine Learning I worked on a team of four. The four of us decided to focus on AirBnB data from New York City to answer a series of questions. Airbnb provides an alternative to hotels, making business travel and vacations affordable and comfortable for almost anyone. We aimed to answer the following questions: is there a way to help tourists make confident decisions for those who want to take a vacation in New York City? Is there a way to ensure business-minded people make calculated decisions that are on par with their own commitment to business decisiveness? Is there a way to summarize the five boroughs to find if certain things affect Airbnb pricing?

The data for this project was provided free directly from AirBnB. There was a total of 6 datasets that related to each of the five boroughs in New York, 5 of which were provided by AirBnB and the sixth was created by the team. The data was explored through traditional exploratory data analysis, the reviews were analyzed via NLP techniques and then data was used to train three different types of classification models. A K-Means model that predicted the pricing group of a specific AirBnB, Decision Tree models that predicted the sentiment of a review based on the NLTK generated results, and decision trees that predicted the price of a listing depending on which borough the listing was in, and finally a logistic regression model to predict listing availability.

Contributions

For this specific project because the original five csv files were already composed of clean data there was not a lot of data cleaning to do. However, I was interested in the correlation between major tourist attractions and the price of the AirBnB listings, hence the reason that there was a sixth dataset added to the mix. I was responsible for collecting the latitude and longitude of the major tourist attractions and putting that dataset together for use across the team. For the exploratory data analysis portion of the project, I focused most of my time on different numeric variables in relation to where they were located geographically. Then mapping the results on interactive scatter plots overlayed on interactive google maps. I also spent some time looking at the reviews of each of the listings, this included looking at the frequency of words, common word pairings and mapping the sentiment of reviews across each of the five boroughs, then plotting it on an interactive map. For the machine learning aspect of the project, I worked on the decision trees with the sentiment analysis data set, as well as the logistic regression model for predicting the availability of a listing.

Project Findings
The initial data analysis showed that, as suspected, shared space is less expensive than non-shared space in NYC. The more private the area or the building, the higher the price will be. Amenities tend to add to the overall cost, in addition to popular locations and being in the more tourist-oriented boroughs. Manhattan is the highest-cost borough followed by Brooklyn and Queens. Through multiple types of data visualization, analyses, and machine learning algorithms many different components make up the price of an Airbnb rental. It is not just the borough in which it resides, but also the amenities and location in relation to different tourist destinations that can affect the price at any given time of rental.

Additional Information
Final Project Grade: 99%
Link in Portfolio: https://github.com/mikayla-j-scott/scottMikaylaMADSSyracuse/blob/main/IST707/Final_Project_Demming_Mercado_McCambridge_Scott.docx
Link to Syllabus: https://github.com/mikayla-j-scott/scottMikaylaMADSSyracuse/blob/main/IST707/ist-707_syllabus.pdf

Natural Language Processing

Course Description
Linguistic and computational aspects of natural language processing technologies. Lectures, readings, and projects in the computational techniques required to perform all levels of linguistic processing of text. This course is designed to develop an understanding of how natural language processing (NLP) can process written text and produce a linguistic analysis that can be used in other applications. The course primarily covers the techniques of NLP in the levels of linguistic analysis, going through tokenization, wordlevel semantics, partofspeech tagging, syntax, semantics, and on up to the discourse level. It also includes the use of the NLP techniques, such as information retrieval, question answering, sentiment analysis, summarization, and dialogue systems, in applications.

Learning Outcomes
The learning outcomes for Natural Language Processing as shown in the syllabus are listed below:

1. Demonstrate the levels of linguistic analysis, the computational techniques used to understand text at each level, and what the challenges are for those techniques.
2. Process text through the language levels using the resources of the Natural Language Toolkit (NLTK) and some rudimentary use of the programming language Python.
3. Describe how NLP is used in many types of real world applications.

Project Description
This project aimed to analyze sentiment in movie reviews using machine learning classification techniques. The dataset used for this analysis was the “Sentiment Analysis on Movie Reviews” dataset provided by Kaggle. This dataset contained 156,060 reviews that were categorized into

five different sentiment categories: negative, somewhat negative, neutral, somewhat positive, and positive. Bag-of-words and POS tagging were used to identify the most frequent words and parts of speech associated with each sentiment category. Different classification models were used to classify the sentiment of the movie reviews: Naive Bayes, Multinomial Naive Bayes and Support Vector Machine models. Two different Python libraries were used, from the Natural Language Toolkit (NLTK) the Naive Bayes package was utilized and from SKLearn Multinomial Naive Bayes and Support Vector Machine libraries were trained and tested on the same dataset. There were a total of eight models trained for this analysis.

Contributions

For this project I was responsible for getting the dataset ready for analysis, this included data collection and data cleaning. This was achieved through using various python libraries, and NLP techniques. Once data was in a state where it was able to be used for the exploratory data analysis, it was uploaded to a shared repository where my teammates could pull it into their notebooks without having to download it or modify the existing file. Using the data I performed all the exploratory data analysis and created all the visualizations for the EDA portion of the project. With EDA complete, the team migrated into the modeling phase. I worked on several models during this phase, using Sci-Kit Learn I wrote the support vector machine models, and the multinomial naïve bayes model for the movie review sentiment classifier models.

Project Findings

Bag-of-words and POS tagging were used to identify the most frequent words and parts of speech associated with each sentiment category. Bag-of-words analysis revealed that certain words were strongly associated with sentiment categories. For example, words like “bad” and “disappointing” were commonly associated with negative reviews, while words like “funny” and “entertaining” were associated with positive reviews. POS tagging analysis revealed.
that certain parts of speech were more commonly associated with certain sentiment categories. For example, adjectives and adverbs were commonly used in negative reviews, while nouns and verbs were more commonly used in positive reviews.

The model that performed the best out of all the models was the SVM model trained on features tokenized by a boolean count vectorizer. The accuracy of this model was 63%. There were continuous struggles across all the models, most often seen between the extremely negative (0) and negative categories (1) and the extremely positive (4) and positive (3) categories. Extreme reviews were most often misclassified as the sentiment classification below it. In result of the continuous misclassification across the models, for continuation of analysis of the dataset a possible solution could be to reduce the number of categories. Combination of the third and fourth categories and the zero and first categories could eliminate some of the misclassification, as there are a lot of similarities between each of the classes. Another thing that could increase the accuracy of the models is to increase the number of reviews for the dataset. There are around 156,000 reviews in the original dataset, however; this number contains duplicates of each of the reviews. After removal of duplicates there are only 8,529 reviews that remain in the dataset. Adding more reviews for the models to train on could also increase accuracy. If additional review data is not available for continuous research, either under sampling or oversampling techniques should be considered to evenly distribute the dataset.

This project demonstrated the potential of machine learning techniques for analyzing sentiment in movie reviews and provides a foundation and recommendations for future research in this area.

Additional Information
Final Project Grade: 98%
Link in Portfolio: https://github.com/mikayla-j-scott/scottMikaylaMADSSyracuse/blob/main/IST664/664_Final_Project%20(1)%20(1).pdf
Link to Syllabus: https://github.com/mikayla-j-scott/scottMikaylaMADSSyracuse/blob/main/IST664/ist_664_syllabus.pdf

Review of Program Learning Outcomes

The previous sections provided a general explanation of the final projects that I chose to showcase throughout my time at Syracuse. This section will demonstrate how components of each one of those projects helped me to achieve each one of the learning outcomes of the program as a whole.

1. Collect, store, and access data by identifying and leveraging applicable technologies.

In each one of the final course projects presented above data needed to be collected, stored, and accessed to be able to complete the core project requirements. We talk about data all the time; the term data is broad by design, especially in this course. Data in the context of most of the courses referred to large sets of information that needed to be stored, cleaned, analyzed, then modeled. This procedure of dealing with data was instilled within me since Intro to Data Science. The most challenging project that I worked on the demonstrates the importance of data preparation was the Avengers project from Text Mining. I say that this specific dataset was the most challenging because it encompassed the whole data preparation lifecycle. Contrary to most of the other courses where the dataset was provided, I had to create the dataset from scratch. This started with the collection of text, planning on how I wanted to analyze it later, and getting it into a format where the analysis took place. This brought me to a point where I was able to bring the text that I had collected into a format that could then be rendered as a csv. Once into a csv file then standard text cleaning practices were applied to the text.

To reflect on this specific learning outcome and the way in which I interpret my learning, it's not so much being able to render a csv that I find the most valuable. It is the ability to visualize what a dataset needs to look like in order to get the job done, having the confidence in knowing I have the skillset to create the dataset and how to take something and make it into a meaningful dataset that can be used by the masses.

2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle.

In every project there was a business case at the core that was threaded throughout the entirety of the project. Out of the projects explained above the one that I feel was very business centric was the Rolling Stone Lyric project from Big Data Analytics. The foundation of the project was built on questions that we were trying to answer as a music producer in the industry. Everything that we as a team did, we made sure that it related specifically back to the business questions that we were aiming to answer from the beginning.

3. Apply visualization and predictive models to help generate actionable insight.

Visualizations help communicate the results of a study in a more effective way. As data scientists it is our job to make sure that the results can be easily interpreted no matter how complex the math or models are. Data scientists often have to be the bridge between the technical and conceptual side of a business or organization, visualizations help aid in this process. Part of having good visualizations is having a core question that the chart or graph is aiming to answer, it allows the end user to focus, draw conclusions and ask further questions. Often times the data that I was working with was the prediction results of a model. Take the AirBnB project, one of the ways in which the logistic regression model was plotted was a heat map. It showed the variables that had the most impact on the price to rent a place. Depending on the end user, it could tell a few different stories. For that of a AirBnB host looking to rent a place out it shows the factors that will get them the most for their rental, and from that of a renter it could potentially show some amenities that they might be willing to give up to drop the overall cost of their stay.

Throughout all the courses, I found myself leaning towards Python rather than R. Python's graphing and visualizations libraries are far more dynamic and are easier to embed and integrate into already existing workflows. In all the projects above, I used a dynamic graphing library called Plotly. Plotly helped me demonstrate the results of my projects in an interactive, intuitive way.

4. Use programming languages such as R and Python to support the generation of actionable insight

My background is in Computer Science, and Mathematics. Before starting the program, I had extensive exposure to multiple programming languages, including Python, C++, Java, R, C, C#, HTML, JavaFX, Ruby, and JavaScript. Coming into the program with that level of experience helped me to create exciting interactive charts and graphs, as well as write algorithms to help better analyze data. I find Python to be more dynamic all together, better for actual workplace integration. During my time here I focused on homing in my Python skills and utilizing a vast range of data science libraries to broaden my existing Python capabilities. Every project above was done using Python as the primary language, I did 2 projects in R in the beginning stages of the program. I find that R is great for once and done data analysis, but in the business world and at the pace we are moving at, I believe R to be less practical when it comes to integration, machine learning operations and pushing results out to the user. Plus, Python has a strong open-source community backing it.

5. Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)

I touched on this in some of the other learning outcomes, but the ability to communicate results to the masses is an essential quality for a data scientist. I believe that this is only achieved by starting with a solid question. Once questions are formed then it is easier to plan, create, analyze, model and visualize the results. This is demonstrated in each one of the projects as the visualizations were created, and then translated back to the original business case questions, and communicated in an effective way back to the person who originally asked the question.

6. Apply ethics in the development, use and evaluation of data and predictive models. (e.g., fairness, bias, transparency, privacy)

During my time at Syracuse there were several other classes that I took that did not make it into this paper because they did not have final projects to showcase. These classes revolved more around theory and non-technical subject matter. Two of these courses were Introduction to Information Policy and Introduction to Cybersecurity. These classes focused on ethics specifically and the importance of factoring for ethics when conducting data science experiments. In every project datasets were evenly balanced, the data that was collected was gathered in a legal, ethical way and models were evaluated for bias.

Final Thoughts

After taking a moment to reflect on the entirety of the program my final thoughts revolve around the learning goals themselves. Each one of the learning goals is not a stand-alone achievement that you can cross off the list, the learning goals are intertwined with each other to make the characteristics of what a great data scientist has the capability to be. During my degree program I feel that I have met every one of these goals and have figured out the way in which each one communicates with each other. This was attained through the many hours dedicated to course work and several projects I completed over the last eighteen months.

I feel that I have learned not only technical skills in this program, but I have enhanced my communication skills as well. Communication when it comes to data science is essential to the data science life cycle. From the beginning business questions need to be clear, the science and math has to be explained in a way that the people asking the business questions can comprehend, and results need to be presented in such a way that anyone, no matter their level of knowledge when it comes to data science, can be interpreted. If you cannot do all those things as a data scientist, then the work that you do is all for not. I have lived by this for many years, and if anything, this program has just enhanced my beliefs, as a data scientist you have to communicate effectively, you have to be the bridge between the story that data tells and the people who want to learn from it.

My final takeaway strays from that of the learning goals but I think is important to mention, to be a data scientist means to be creative. There are many times where there isn't any direction that you are given, it is just you and the data. You have to find the story that the data is telling, or in some cases not telling and be the one to share that. Creativity comes in the way in which you manipulate the data, what models you choose, writing your own algorithms, and then of course visualizing the results. Visualizing the results means taking something that was once really complex, and making it interpretable for the masses. Being able to conceptualize that is something that I have worked on throughout this program. This niche piece of what it means to be a data scientist is exactly why I got into the field of engineering, it takes creative problem solving to find answers, it takes creative people just asking “what if” to make the world go round.