John T O'Rourke

Email: jtorourk@syr.edu

Net ID: 294756033

Master of Science in   
applied data science

Portfolio Report

Table of Contents

[Introduction 2](#_Toc134285951)

[Program Learning Goals 2](#_Toc134285952)

[IST 664 – Natural Language Processing 2](#_Toc134285953)

[Project Description (Email Spam Classification) 3](#_Toc134285954)

[Learning Goals and Outcomes 4](#_Toc134285955)

[IST 707 – Applied Machine Learning 4](#_Toc134285956)

[Project Description (Song Popularity Predictions) 5](#_Toc134285957)

[Learning Goals and Outcomes 6](#_Toc134285958)

[IST 718 – Big Data Analytics 6](#_Toc134285959)

[Project Description (Predicting a Drought in Continental US) 7](#_Toc134285960)

[Learning Goals and Outcomes 8](#_Toc134285961)

[IST 769 – Advanced Big Data Management 9](#_Toc134285962)

[Project Description (Spotify Playlist Generator) 9](#_Toc134285963)

[Learning Goals and Outcomes 10](#_Toc134285964)

[Conclusion 10](#_Toc134285965)

[References 11](#_Toc134285966)

# Introduction

The Master of Science in Applied Data Science program is a program that focuses heavily on how data science can effectively be utilized to develop key insights for a number of different stakeholders ranging from individuals to large corporations or government entities. This, paired with staff that has real-world experience in the field, allows for students to use the tools and methodologies in their every day lives.

One aspect of this program that I found instrumental in my learning was a focus on creative freedom within topic selection for projects. Without this freedom, I feel my projects would have all been driven to a “paint by numbers” structure that did not fit my ambitions nor the reality of the field. I thoroughly believe that all of the courses within this program have impacted my ability to be a contributing member of this field in a positive way but in the interest of selecting my most impactful projects, the following courses have been selected for use in this portfolio.

* IST 644 – Natural Language Processing
* IST 707 – Applied Machine Learning
* IST 718 – Big Data Analytics
* IST 769 – Advanced Big Data Management

# Program Learning Goals

The MS in Applied Data Science program holds certain key learning goals/objectives in high regard. Those goals/objectives are as follows:

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

# IST 664 – Natural Language Processing

This course was focused on all thing’s language processing and how that NLP could be utilized in real world use cases. With the new release of ChatGPT as this course was getting underway, we were able to see firsthand the impact this discipline had on the Machine Learning and AI communities as well as everyday people.

Throughout the course, we learned multiple different NLP techniques/methods such as POS tagging, Sentiment Analysis, Context-free Grammars and Parsing as well as general Corpus Statistics and Language modeling.

## Project Description (Email Spam Classification)

One common approach to spam detection within emails is to use machine learning algorithms. We can see this through the current Outlook Protection® suite as well as Google’s Spam protection system for Gmail. Naïve Bayes is one key algorithm in classification of documents or items within a given dataset. This algorithm is used to calculate the probability of an email being spam based on the content within a given email.

The goal of this project was to determine the likelihood that an email was spam based on its sentiment through the use of Naïve Bayes as well as the VADER sentiment lexicon. The data for this analysis has been sourced from the Enron Spam dataset which is a public email corpus gathered from the company’s court case. In total, 1,500 spam files and 3,672 “ham” files were used. This only accounts for 1 of 3 spam groupings and 1 of 6 legitimate groupings. Additionally, spam emails were injected into the already classified spam emails in order to ensure a proper sample for analysis. This was completed by the dataset maintainers.

A screenshot of a computer

Description automatically generated with medium confidence

Three models were developed with varying degrees of accuracy and each model allowed for experimentation with different lexicon and vectorizer combinations to determine the best fit for the dataset.

A screenshot of a computer

Description automatically generated

## Learning Goals and Outcomes

This project successfully met the goals of collecting, storing, and accessing data by utilizing applicable technologies such as Python as well as the Enron Email dataset. The project applied visualization and predictive models, including Naive Bayes and VADER sentiment analysis, to generate actionable insights for spam detection in emails.

The programming languages used in this project were Python and its associated libraries such as scikit-learn and NLTK, which allowed for efficient analysis and transformation of the text data. The insights gained from the analysis were communicated through visualization techniques such as Word Clouds and Confusion Matrices, allowing for all audience members to understand and act on the results.

Furthermore, the project applied ethics in all stages leading to the completion of the project by ensuring that no bias was introduced by up sampling and down sampling the data to allow for fair and transparent model development and analysis.

# IST 707 – Applied Machine Learning

The different modeling techniques and technical reporting requirements of this course assisted my learning by forcing me to think about this field in a more academic and research sense than the purely business focused applications that are prevalent with most other courses.

One other key aspect of this course was ethical analysis and how to eliminate bias, ensure fairness and transparency for the data, how it is sourced and how it is analyzed.

## Project Description (Song Popularity Predictions)

This project focused on generating a predictive model using Naïve Bayes, Decision Trees, and Support Vector Machines (SVM) models to determine whether or not a song would be popular based on the different characteristics of a given song. The data was sourced from Kaggle and was sourced by a Kaggle user utilizing the Spotify Web API. Spotify gives songs a number of different scores based on several different categories/attributes. These attributes were pre-processed and we were able to determine the top attributes that impacted model performance were “danceability”, “loudness”, “duration (in ms)”, “energy”, and “key”.

The data set used was the top 2000 songs on Spotify from 2000 to 2019. Unfortunately, one of the key issues that was overlooked with this project was the predicting the popularity of a song based on a data set full of the most popular songs was not incredibly accurate and was skewed when verifying our results.

A picture containing text, diagram, screenshot, rectangle

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

In the end, R was utilized for the entirety of this project, and we were able to develop a model with an accuracy of 54.06% accuracy utilizing Naïve Bayes. While the accuracy is not what we were hoping for we think that the best way to improve upon our model is to first look at an association rule mining attempt to tell us more about the attribute that was provided in our data sample so that we can better determine which attribute would and their correlation would be best for determining our models.

## Learning Goals and Outcomes

This project helped our team to collect, store, and access data by leveraging R and the wonderful repositories of data sets available through Kaggle. Along with our descriptive analytics visualizations, we were able to showcase our insights into the data and results of our models through bar plots with the use of the R package “ggplot”. Due to the nature of the data set, we were able to apply ethical development, use and evaluation of the data by ensuring these analyses were run on random samples while ensuring the goal value (popularity) was obscured.

# IST 718 – Big Data Analytics

This is a course that I found to be most applicable to the roles I want to eventually be a part of. It is incredibly useful to be able to gain valuable insights from large, complex, and challenging data sets. The push by our professor to find a data set we were not necessarily comfortable or familiar with has stuck with me. This forced my group to step out of our comfort zones and develop an analysis on a topic we have little background knowledge on.

We were essentially required to do research on the topic outside of our class and group meeting times to understand the data and what our results really meant. For me, I appreciated this tremendously as this prepared me for future roles where I may not be a subject matter expert and will need to either research or lean on colleagues to effectively understand the base data as well as the insights generated from my analysis.

## Project Description (Predicting a Drought in Continental US)

Prediction is critical for warning governments and farmers about early droughts. The main problem this project attempts to solve is early drought detection to combat droughts across the United States. The goal of this project is to effectively and efficiently predict where a drought may occur and just how severe it is by developing a forecasting model. This will allow for the government to set up programs that help combat areas affected by severe droughts based on weather and soil conditions observed. We will be using data that has been collected from Predict droughts using weather & soil data (Minixhofer, 2021).

Utilizing random forest and decision tree modeling techniques we effectively utilized our historical data to build a statistical model for predictions through regression and classification. By using these modeling techniques, it enhanced our understanding of the data as well as predicting possible values. The forecasting models and geographical mapping of where droughts will occur have been completed. SARIMAX predictive model was chosen to output our prediction for geo mapping. These modeling techniques and visualization gave us an accurate representation of the location for potential droughts as well as its severity.

A screenshot of a computer screen

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with low confidence

From this model, a useful visualization was generated that looped through the predicted results of the drought severity values over a map of the United States.

A map of the united states

Description automatically generated

## Learning Goals and Outcomes

As this project pushed me outside of my box, it allowed me to generate an analysis and project that has impact and fulfills all 6 of the Program Learning Goals. Most importantly, this analysis and model development helped create actionable insight across multiple contexts while communicating those insights in a meaningful way to all parties involved (large business, government entities, and individual citizens).

The overall accuracy of our models was also well within the acceptable thresholds and could be acted upon to help prevent further droughts and possibly implement drought prevention methods that would benefit all parties involved.

# IST 769 – Advanced Big Data Management

Being one of the few courses where experimentation with database architectures and paradigms was encouraged, this course very quickly caught my interest. The ability to use Docker and learn how several different database paradigms operate and function alone or with one another was a refreshing and worthwhile experience.

With the increased usage of containers and virtual machines in the industry, being able to gain experience with this service through developing our own script and utilizing the tooling allowed me to expand my skill set past what a normal data scientist would have.

## Project Description (Spotify Playlist Generator)

This project focused on the data and how it could be managed along with its usage in an application or used for analysis. The key goal of this project was to ingest song and category data using the Spotify Web API for storage in MongoDB. Out of that key objective, I was able to create a playlist generator that leverages user input to not only determine which categories of songs they would like to use, but also how many songs they would like to have in their playlist.

As this course was not geared toward analysis as much as tooling, no significant transformations were completed on this dataset. Alternatively, further improvements could be made to include a data analysis component with the top categories data pulled from the Spotify Web API.

**About Spotify and the project**

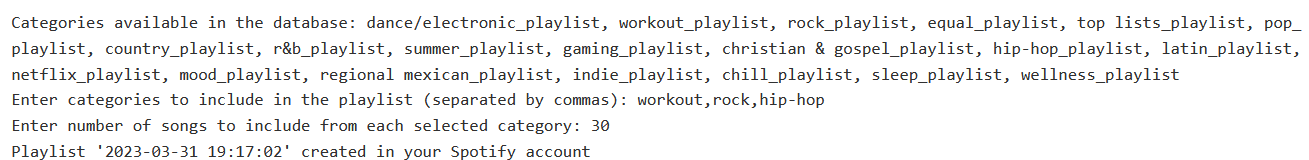
Spotify is a world-wide music streaming service that allows users to listen to music anywhere at any time. The service offers several different APIs that are used to handle anything from playback to data collection. This project utilizes the Web API to import data from Spotify’s category playlists into a locally stored MongoDB database and generate playlists based on user input.

All data is sourced directly from Spotify using their Web API. This iteration of the project solely imports song data from the top 20 categories on Spotify. It is worth noting that certain categories provided by Spotify are deprecated on the platform, but the API will still generate collections for them.

The steps to complete this project consisted of the following:

* Docker Compose file generation and start.
* MongoDB configuration through Jupyter.
* Spotify Web API authentication within Jupyter through SpotiPy.
* Collection generation for the top 20 categories.
* Importing the individual song data within the top category’s playlists.
* Generating the playlist.

Once the above steps had been completed, we were able to reach the desired outcome of a playlist consisting of a random mixture of songs between user selected categories and with a user selected amount.

A screenshot of a computer

Description automatically generated with medium confidence

## Learning Goals and Outcomes

This project was able to demonstrate my ability to collect, store and access data by identifying and leveraging applicable technologies as well as my ability to use Python along with its available libraries and tooling to build an application.

# Conclusion

Each of the 4 courses covered within this document were crucial to my understanding of how to properly apply and utilize different analyzation and modeling techniques along with finding interesting ways of showcasing the insights gathered from these projects. While not all of my projects have an application in the business world, the techniques used can easily be transitioned over to different industries to develop new and interesting insights regardless of the data set.

The most impactful course in my opinion was IST 718 – Big Data Analytics due to how the course was managed and the focus on using a topic that was not in our wheelhouse. Being able to provide analytics on another topic that was foreign to me helped me realize that in this industry, there are no guarantees in what data you will receive, how you will receive it or where you will receive it from. You need to adapt to ensure proper analyzation and to provide actionable insights.

# References

Androutsopoulos, I. (n.d.). The Enron-spam datasets. Retrieved March 21, 2023, from https://www2.aueb.gr/users/ion/data/enron-spam/

Koverha, M. (2022, May 31). *Top hits spotify from 2000-2019*. Kaggle. Retrieved July 12, 2022, from https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019

Metsis, V., Paliouras, G., & Androutsopoulos, I. (n.d.). *Athens University of Economics and Business*. Retrieved March 21, 2023, from https://www2.aueb.gr/users/ion/docs/ceas2006\_paper.pdf

Minixhofer, C. (2021, March 3). *Predict droughts using weather & soil data*. Kaggle. Retrieved September 21, 2022, from https://www.kaggle.com/datasets/cdminix/us-drought-meteorological-data

Pang, B., Lee, L., & Vaithyanathan, S. (n.d.). *Thumbs up? sentiment classification using Machine Learning Techniques*. ACL Anthology. Retrieved March 21, 2023, from https://aclanthology.org/W02-1011/

*Resources and dataset descriptions*. Resources and Dataset Descriptions - VaderSentiment 3.3.1 documentation. (n.d.). Retrieved March 21, 2023, from https://vadersentiment.readthedocs.io/en/latest/pages/resource\_description.html