Data Science Portfolio

Syracuse university, class of 2020

Kevin Vogel

2020

Contents

[Introduction 2](#_Toc55063142)

[A simple misunderstanding 2](#_Toc55063143)

[My journey 4](#_Toc55063144)

[Crawl 4](#_Toc55063145)

[Walk 6](#_Toc55063146)

[Run 8](#_Toc55063147)

[Takeaways 10](#_Toc55063148)

[What is next 10](#_Toc55063149)

[Bibliography 11](#_Toc55063150)

# Introduction

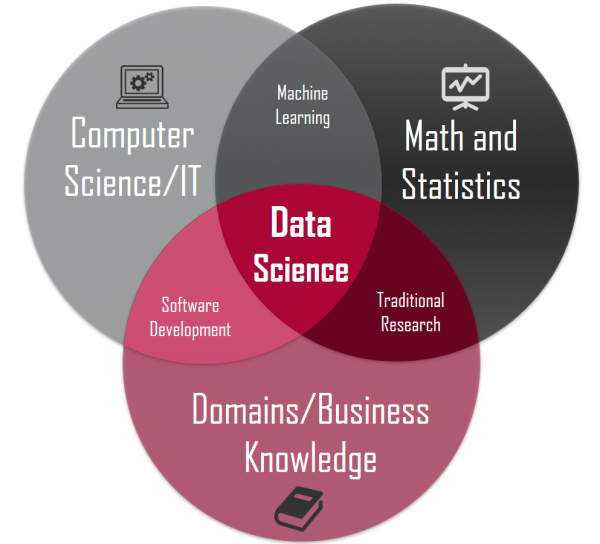
I began my journey toward a Master’s in Data Science approximately one year before I entered the program. I interned with a company when I was in my MBA program at the University of Tennessee and had occasion to begin using VBA to automate workflows. I know that doesn’t sound like much, but it led me to buy a book on how to code in VBA, and another, and another. I had discovered that I very much liked the challenge of creating something functional and valuable from nothing. In 2019 I applied and was accepted at Syracuse University; bright eyed and bushy tailed I took my first classes in the summer of 2019 and I am so glad that I did.

At the time of my joining the program I did not have any formal training in computer science or a related field; through work as an LSS project manager I had a good foundation of statistics but that was about all I could hang my hat on. It was an uphill battle to be sure and at times, especially early in the program, it felt like I was drinking water from a firehose.

Now, set to graduate in less than two months from now, I feel comfortable embarking on most any project whether it be in R, Python, or a language I have yet to learn. That, to me, is the real joy of data science – you never know everything, there is always something to learn, and generally, someone is almost always willing to go the extra mile to help you. Syracuse and the data science community at large share an espirit de corps that is refreshing and helped transform me from an absolute novice to an adept data scientist. The remainder of this paper wills detail my journey.

# A simple misunderstanding

Data science is misunderstood; that, I believe, is the most certain thing that anyone can say about data science. I certainly came into this program with the notion that data science is a lot about machine learning – it certainly can be, but it is so much more than that. As I learned in IST 687, data science work spans the gamut of project work as it relates to data and can vary wildly depending on the company at which you are hired. A widely accepted view, and my personal favorite, is that data science is the combination of computer science, math and statistics, and subject matter expertise (Cui, 2020). Machine learning is the overlap of math and statistics with computer science but without the domain or business knowledge. This is a burgeoning area of study but only a small part of data science. Too often machine learning and data science are thought to be interchangeable.



Even more generally, it can be argued that data science is OSEMN (Lao, 2017). The framework describes the steps that all data science projects iterate through and the work that is accomplished at each step. Truly, a data scientist is a computer scientist with math and statistics knowledge, with domain knowledge or the ability to work collaboratively to obtain domain knowlegde from subject matter experts and who follows a rigorous project framework to come upon the right conclusion. My transition was not seamless; it was stepwise.

# My journey

## Crawl

The first two classes that I took were IST 687 and IST 659. While I learned a lot about the basics of R and the data science framework in IST 687 I feel that a basic building block that I was missing was my understanding of relational databases which is what we learned about in IST 659. In this course I learned how to build a database from the ground up using nothing but SQL – SQL skills have helped me collect and organize data. I put my newfound abilities on display in my final project in that course when I created a database for my father’s vinyl record collection (Vogel, IST 659, 2019). I started with a conceptual model, transitioned to a logical model, built and loaded the database, and finally queried it. Building a database from scratch taught me the importance of forward-thinking data governance, future proofing of source systems, and how to implement reasonable security protocols – the fact that not everyone needs access to everything was hammered home in IST 722, a follow-up the into database management course. Querying databases was something that I had little experience with before – developing this skill early in the program set me up for success in what I will be referring to as the walk and run phases of my progression as a data scientist.

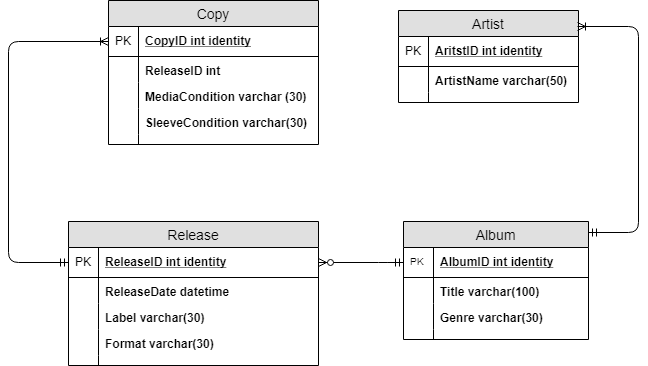


Figure : Conceptual model

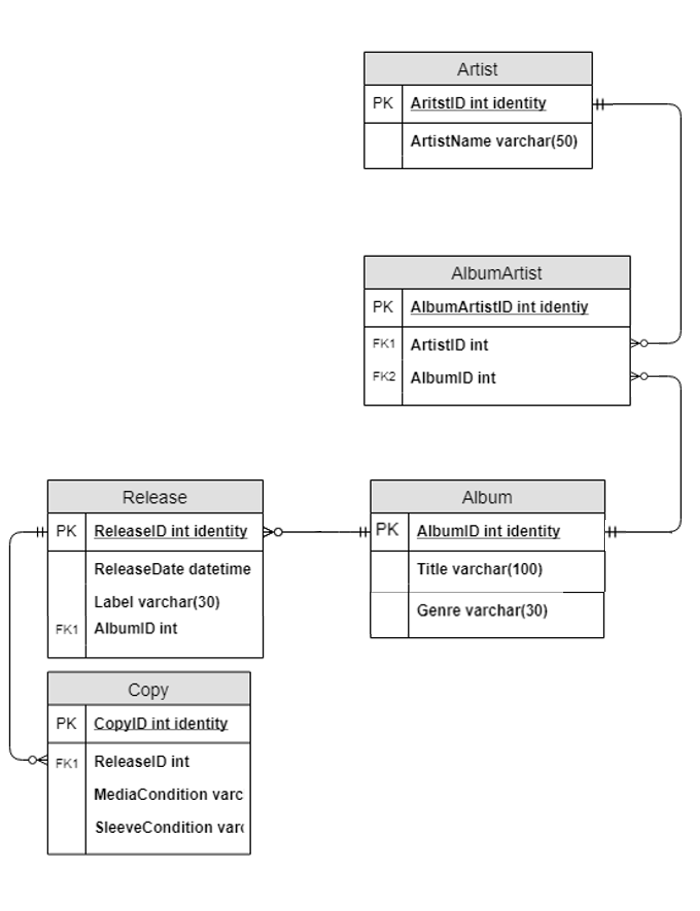


Figure : Normalized logical model

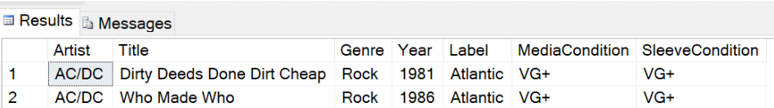


Figure : Query of database

## Walk

Now a full semester into the program, I knew everything (*sarcasm*). I did begin to take on projects that were more ambitious and that walked through the entirety of the OSEMN framework, however. That first foray, and a huge learning experience was my IST 707 project.

Several of the team members had significant dealings with not for profits and were interested in the topic. United Way is one of the largest not for profit organizations in the world with a diverse mission and a diverse donation and need base. We chose to develop a classifier that could accurately predict whether or not a US county had or did not have a United Way (Vogel, IST 707, 2020). The purpose of this analysis was to determine based on where current locations were and their similarity to other counties whether or not some counties were underserved by the United Way.

We began this project by obtaining data. Thankfully, not for profit data is widely available on the internet as they are legally required to share how much revenue they obtain, their operating budgets, etc. Unluckily, there is no easy way to obtain this data in a flat file. As such, I went to not for profit websites, downloaded pdf’s of branch tax information, and put it into a flat file. As this was a labor-intensive task, we limited our data gathering to United Way locations in Tennessee and Kentucky. Later I would learn about web scraping and feel very silly. This data was paired with census data, csv from US Census Bureau, which gave us a fairly robust set of features to classify from. The census bureau data was originally the only data that we thought was necessary; gathering the not for profit income data was necessary as the census bureau data was not sufficient. Developing an alternative strategy was necessary and prudent in this case.

A note on using data that shares demographic information – any time you are using this type of information, you must be careful that the way in which you are joining or otherwise merging data does not put someone at risk of having their privacy reduced or removed. For instance, in this analysis we collected data on how much money was donated at each of the branches – we did not collect who donated and even if we did, we would never share that. Maintaining privacy and anonymity is paramount to data science.

The data did have to be scrubbed. The census data was gathered at a tract level which is smaller than an individual county. This had to be converted to county and the data merged. By the end of the scrubbing phase, we were left with 37 features and 215 rows in our training and test set.

When it came to modeling, we tried several different methods. The first of these was association rules mining. Association rules mining does not truly classify data so much as it does provide a means of analyzing whether or not one feature is likely based on other features. This did help us with our analysis later however.

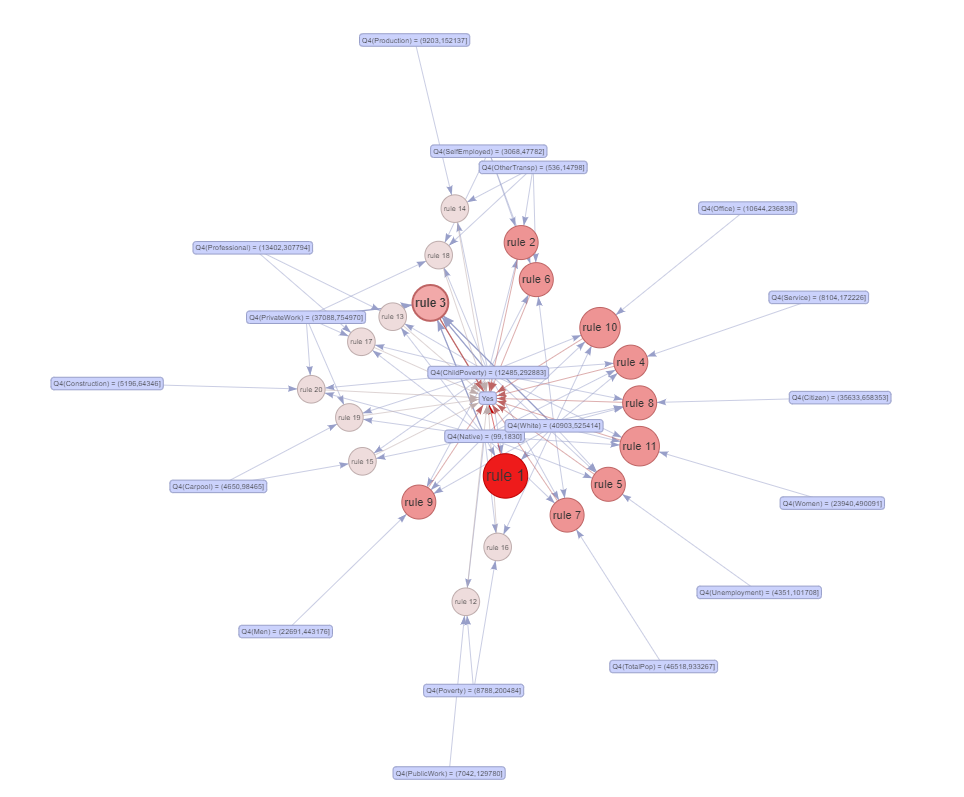


Figure : Association rules mining

Multiple other models were created ultimately ending with the below model, a random forest classifier that predicted correctly 88.4% of the time based solely on child poverty and poverty rates.

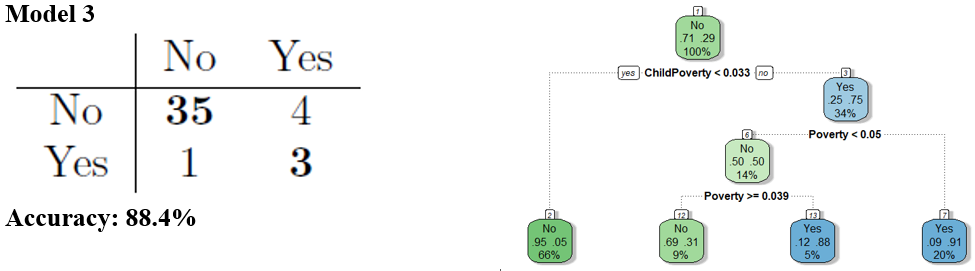


Figure : Model output

This project was viewed through the lens of a United Way regional manager. It is the goal of management at United Way to provide the greatest good for the greatest number of people. It is a difficult task as their locations are both revenue centers and cost centers. As such, they must place their locations where they can collect the most money and help the greatest number of people. We identified counties that were underserved that met those descriptions and recommend that those counties be the next locations where United Way expands. This analysis and action plan would directly benefit the United Way with their business decisions.

This project helped me to grow in the areas of obtaining data and modeling especially but also in understanding what the business problem is that the firm is facing and addressing it.

## Run

Running to me means two things: working on real problems and communicating clearly and concisely such that the message being conveyed is not lost in translation. I am just now beginning to run.

One of my most recent classes was text mining. This being an election year, I was thinking a lot about politics. Ben Harwood and I embarked on a mission to predict the political affiliation of an individual solely by their speech (Harwood & Vogel, 2020). For this project we gathered inaugural gubernatorial addresses in their entirety in text form. We sampled equally from 10 states leaving us with 20 speeches in total from 10 democrats and 10 republicans.

This was my first-time getting exposure to the natural language toolkit. We used n gram frequency to pick up on trends in language and sentiment scores. Positivity and negativity were then used as features on a sentence level to predict the political leaning of the speaker. We were best able to predict using part of speech tagging and predicted with 0.67 accuracy.

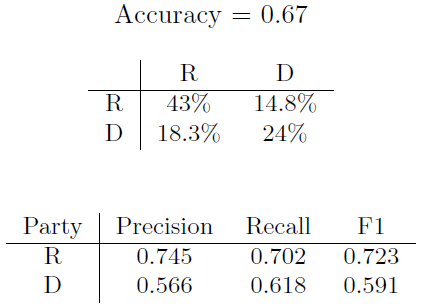


Figure : Accuracy of best model

I learned a lot about sample size on this project. Presenting this data was a challenge as some basic level understanding of NLP is necessary to understand finer details of the analysis but we were able to effectively communicate our message by bringing it up a level and talking about the general sentiment and common phrases (republicans don’t like saying “health care” as often as democrats).

There is real world use for this analysis. If we can predict the likelihood of someone’s political leaning or proclivity just from their speech, we can probably determine who the candidates are that are most likely to ascend the ranks as they will likely be the ones that most embody the party values and do not rock the vote. Want to know who will be in charge in 20 years? You should look at what people are saying now and how it aligns with their party.

# Takeaways

My big three takeaways from this program are:

1. Perfect data does not exist. At some point you must make do with the data that you have and make the best choice that you can given what you know and what you have gleaned.
2. Every project should be iterative. Beware the project that goes smoothly from beginning to end without going back on occasion – you probably did something wrong. Invite critique and criticism.
3. Meaningful communication with teammates and superiors is key. If you can’t collaborate and work with others, you can’t work effectively. If you can’t communicate your work to someone in plain language, you can’t work effectively. Throughout this program we crafted presentations that helped improve our ability to communicate with managers and technically minded people alike.

# What is next

Next for me is to move into a full-time role as a data scientist with the company that I currently work for. This is a new, blossoming team and there will inevitably be bumps in the road. I feel that Syracuse has uniquely prepared me for this challenge.

# Bibliography

Cui, Y. (2020, January 16). *A Minimalism Approach to Understand What Data Science Is*. Retrieved from Towards Data Science: https://towardsdatascience.com/what-is-data-science-4565e457cc17

Harwood, B., & Vogel, K. (2020). *IST 664*. Retrieved from Github: https://github.com/kevinvogel13/IST\_664

Lao, R. (2017, November 5). *Life of Data | Data Science is OSEMN*. Retrieved from Medium: https://medium.com/breathe-publication/life-of-data-data-science-is-osemn-f453e1febc10

Vogel, K. (2019). *IST 659*. Retrieved from Github: https://github.com/kevinvogel13/IST-659

Vogel, K. (2020). *IST 707*. Retrieved from Github: https://github.com/kevinvogel13/IST\_707