My Data Science Experience

Wesley Stevens

# Introduction

The Applied Data Science program at Syracuse University pushed me to utilize many data science techniques on a variety of datasets and in unique ways. As a part of this program, I learned how to deal with data in all parts of its lifecycle and am now confident in my abilities to take raw data, wrangle, pipeline, analyze, and create algorithms that will help drive business goals. In this paper, I will delve into three of my favorite data science projects completed during my tenure here.

# Data Warehousing

This course was the highlight of my entire graduate experience, and I can say that the final project deliverable was the most rewarding to complete. In this course, we worked on teams, using the Kimball method, to build a data warehouse in order to combine two very different databases. The synthetic data came from two imaginary companies called Fudgemart and Fudgeflix (modeled after Walmart and Netflix). We identified five different business objectives that this warehouse could help achieve including: Sales, Order Fulfillment, Inventory Levels, and Customer acquisition and reviews. For the sake of time, we limited our creation and analysis to Order Fulfillment.

Using a Bus Matrix and detail-level dimensional modeling, we designed our database, making sure to avoid extraneous information not related to Order fulfillment. Then, using Visual Studio, we used SQL queries and ETL to automate moving the necessary data from the two databases, cleaning, and importing it into the data warehouse through a staging server. We set it up so that as new data enters the old databases, it will automatically be imported into the data warehouse without further effort or complication.

Once the data warehouse was complete with dimensions and fact tables and filled out with data, we used Power BI to do some exploratory data analysis. We discovered some interesting trends in each of the companies including that the lag time between the order and delivery of a product or service increased in the month of July. Another interesting insight we found was that order fulfillment time was directly proportional to the buying power of a client. Meaning that these companies are catering to the wealthy.

Because of the lessons learning in this course, I understand how data needs to be organized and automated in a business. I also now have a deep understanding of how databases work and the importance of multi-threading.

# Data Analytics

This course taught me how to spot irregular data points, how to clean them, and how to use regression and classification methods to then create valuable systems. For the final deliverable in this course, I designed and built a replacement for the Pokedex (originally from the world of Pokemon) using regression and classification techniques.

If you are not aware, a Pokedex is a large machine that Pokemon enthusiasts carry around the wilderness to catalogue Pokemon (i.e. creatures). One must encounter, capture, and put a Pokemon through this machine in order to discover its identity, legendary status, base statistics (e.g. attack and defense), and more. This process is difficult due to the low rates of quality Pokemon capturing and can cost Pokemon enthusiasts a lot of time and money.

Using only the Pokemon’s physical characteristics, I used a staged approach to first classify its legendary status (legendary pokemon are rare and valuable). That was completed with and F-measure of 88% using an XGBoost classifier. Then, that data was used in 6 different regression models to predict the Pokemon’s base statistics. The algorithms used were Linear Regression, XGBoost, and SVM Regression. I found that different algorithms worked better for different statistics. Each statistic had a MSE score of less than 0.02, meaning that the created models were an excellent fit. I then used that data to predict the Pokemon’s type. This was done using an SVM classifier with an F-measure of 83%. Finally, due to its simplicity, my method was shown to be 42 times faster than the traditional Pokedex method at determining Pokemon attributes, albeit slightly less accurate.

Through this project, I was able to solve many unique challenges posed by small datasets. I overcame the limitations of the data through innovative means and achieved excellent results because of it.

# Natural Language Processing

NLP has always been a topic of interest to me, so being able to learn and apply modern techniques was incredible. Though the course mainly used the Python package nltk, I chose to compare how Markov Chains compared to modern-day methods (GPT-2) on the task of Natural Language Generation.Markov Chains were the state-of-the-art method for Natural Language Generation for many years until computational limits were expanded in the early 2000’s. Once GPUs were invented and optimized, GPT-2 was created as the most intelligent neural network of our time.

I utilized corpuses from Project Guttneburg—an online repository of books whose copyrights have expired. I pulled and cleaned four books written by Charles Dickens; Then coded and trained each of these models with a single task in mind: Predict the next word. The results came out quite comical and interesting in some ways. The main difference between the two methods was that, although neither made grammatical sense, GPT-2 was able to grasp context much better than Markov Chains. Markov Chains could make 2 or 3 words sound like the could have been written by Charles Dickens, but GPT-2 could make a sentence or two do the same and fit coherently together.

I concluded that true artificial intelligence as imagined by science fiction authors is still a ways off, but we have progressed by leaps and bounds in the past decade. Because of this course, I am confident in my abilities to utilize and extract use from any corpus as well as classic ordinal and numeric data.

# Conclusion

These are but a few of the many interesting and unique projects I completed during my graduate experience. Through them, I have learned how data must be dealt with at every point in time. The ADS program has taught me how to store, wrangle, and generate business insights into any dataset as well as to create predictive products to further optimize a businesses utility of its data. Because of the ADS program, I am confident that my abilities as a data scientist will allow me to provide any business value from day one and I am excited for what the future will hold and the incredible applications I will make.