IST 782

Michael DeMaria

10230-3469

mjdemari@syr.edu

Applied Data Science Portfolio

Introduction

I completed my undergrad twenty years ago. Having now spent decades in industry, my approach to this program may differ from the typical journey. I set out not to build a portfolio for finding a career, but to maintain it.

My computer science bachelors, earned at S.U. ECS college, has served me well so far. However, that is far from guaranteed. The industry has changed, big data / AI / Machine Learning have all become very relevant. As I mentioned in my application letter, my goal for this program was tri-fold. First, data science is a growing field, so this degree gives me a foundation to transition over to that role should it become necessary. Secondly, even if I do not change jobs, I wanted to apply this curriculum to my current role. Third, I wanted to share my industry experience with the other students, hopefully helping them succeed in the workplace.

Throughout this program, I have worked full-time as a software developer, software architect and team lead. This has put me in the unique position in that I could apply lessons immediately. As soon as the semester was over, I could bring that knowledge into my next work project. This portfolio will show more than how this program is useful to future employers, but will also show how this program has demonstrated my capabilities with immediate results. I was fortunate in that all my group projects were great experiences, and covered a range of subjects. While I brought real-world, business experience into each project,

I made sure that this curriculum was not geared to solving just my current work dilemmas. Nevertheless, each project was conducted with the thought of how this could benefit me in the enterprise. Many of these projects were done as small, self forming teams consisting of 3 to 4 students. The projects listed below are in chronological order. Source code, final reports and presentations are available at https://nand.net/mdemaria/portfolio/

IST 659 - Data Administration

This project had us building a database system to support a nonprofit community learning center. The endeavor, called Common Denominator, consisted of two lines of business. The first line was a math tutoring service. A community member would sign up for classes, and an invoice would be generated. This revenue was used to support the nonprofit side. Community members would specify their demographics, history, languages spoken, and so forth. They would then sign up for free programs. Users and program leaders could see what traits the participants had in common. As the name Common Denominator implies, the goal was to find something in common amongst all the participants, and to hopefully use that as a jumping off point to better understand one another.

We divided the work amongst the group based on our collective strengths. One member was very passionate about this topic and was hoping to eventually productize this idea. She functioned as product owner, coming up with the overall vision, requirements and project planning. The second team member functioned as a UX expert and front end developer, building a mockup of the web interface. My role was that of technical team lead, primarily focused on database design and technical considerations.

From the requirements and web interface, I was the principle architect of the database design. We build a relational database that simultaneously supported both lines of business, including

community member management, an invoicing system, classes offered, class registration and program management.

We wanted to implement more than just a database, but treat this like it was intended for production usage. Of particular note is that we ensured user passwords were one way hashed and salted, a common security best practice. Of all the people who presented in the class, we were the only group that knew of this encryption technique. We built procedures for setting, verifying and changing user passwords. As the users were entering personally identifiable information, we knew how important security would be to the success of this project. All database interactions were done via stored procedures.

In addition to the database, we also were concerned with producing adequate documentation. The vision was that this software would be implemented within a franchised business model and each location would have their own implementation team. We wrote a fully detailed system integration guide. This document explained every database view, every SQL function, every parameter and the database triggers. This level of technical documentation is fairly rare but extremely valuable.

This project assumed a greenfield implementation, so there was no existing data to collect or analyze. Instead, this project focused on planning, design, implementation and documentation. We had to come up with a brand new data strategy. As we developed the product, decisions had to be made to ensure that the business requirements of matching people by demographics could be met and also allow for us to generate appropriate KPIs. We needed to tell the business which classes were the most popular and which math tutors were generating the most revenue over time. We built the data model and SQL views in such a manner that we could easily retrieve that information. Finally, we showed our adherence to

secure software design. The use of password encryption and stored procedures helps avoid two of OWASP's top ten vulnerabilities (OWASP A02, OWASP A03).

IST 722 - Data Warehouse

This project involved building a brand new data warehouse. In our scenario, two complimentary companies were merging together. Each company had their own customers, products and business models. One company was focused on selling individual orders of merchandise via an e-commerce shop. The other company was shipping products via a monthly subscription model (analogous to amazon.com's storefront and Netflix's by-mail DVDs). Some customers were the same between both lines. Our goal was to unify these two disparate lines of business into a single, cohesive view of the customer, identify fulfillment issues and potential opportunities.

This project focused of data analysis. For the first part of the project, we had to merge together these two systems. We had to properly understand the data, and the relationships of the data, for each of the businesses. We then had to come up with an implement a plan to combine this data together. Common data points needed to be found, and we had to ensure we were not creating duplicate customers. It would have been easy to take the entire customer list from Company A and append it to Company B, but that would not given real insights into the customer. It was important that we match existing customers to produce accurate information. This project required learning how to use an ETL tool. In practical terms, I used this knowledge at my employer to better communicate with our IBM Data Stage developers. Data Stage is an ETL tool used by Equitable Financial Life Insurance Company. I was able to refer to functionality and behaviors using common ETL terminology.

After we merged the data together, we then had to analyze the data for fulfillment issues and business opportunities. We built a series of PowerBI dashboards to show order fulfillment and

the efficiency of our product shipment partners. We were able to see that shipping time increased 20% in Q3, and one of the partners accounted for less than 10% of order fulfillment. We also were able to ascertain that one line of business was significantly more efficient at shipping products. While the data could not say why this was the case, it give us enough evidence to be able to pitch to management a future data study.

This particular project has become very relevant in my job. We have just started a project to merge three of our database systems (technically, Salesforce instances), spanning four lines of business, into one singular system. Each of these systems were developed independently and with their own data models. The work performed in this data warehousing course will be replicated nearly identically.

This course also taught me the concept of late arriving facts. This is a scenario when you get a portion of relational data, but not the entire item. For example, you may receive a sales order before you get product data. In this case, when you load the order into the data warehouse, it would have no product to tie to, violating referential integrity. Late arriving facts is a mechanism where you create temporary placeholder values, and then when the product data arrives later, fix the referencing. One year after taking this course, I encountered this same scenario at work. I had a list of contracts in my system. We had a mandatory regulatory requirement to flag certain contracts as needing a fiduciary review. In some cases, the contracts under review may not be in my system, or may be in the process of getting imported (it takes several days from contract issuance to getting loaded into the CRM). I used the concept of late arriving facts to create a temporary placeholder for the contracts not in the system, and merge them together when the contract data arrives.

IST 652 - Scripting for Data Analysis

An important aspect of data science is the ability to correlate and cross-compare data from different sources. In this project, I explored whether New York State's farmer markets were correlated with population density, income or farmland acreage. To accomplish this task, I had to combine data from four different data sets. I retrieved a list of farmer markets, a list of agricultural acreage by county, income tax records and population density.

While the data was fairly clean, there were a lot of inconsistencies between the four data sets. County names were sometimes in uppercase, sometimes lowercase and sometimes mixed. Farmland was measured in acres, while population density was in miles. I had to decide if income levels should be measured against all returns, or merely returns that had taxable income. New York has a minimum income threshold to meet before reaching taxable income levels, which means that evaluating based on taxable income would exclude many low income workers.

All of this data needed to be properly formatted and correlated together. Natural keys had to be found, and a significant number of data decisions had to be made. Do we aggregate at the county or zip code level? Should we look at acreage or square miles? Will we consider all arable land or only active farmland? These are all important questions that would have to be brought back to a business user, product owner or data steward. This project emphasized that merely having data is not enough. It is important to understand the business questions we are trying to answer.

IST 707 - Applied Machine Learning

In my day job, part of my responsibility is ensuring a group of Financial Professionals are working as effectively as possible. The company has millions of clients, so prioritizing customer reach outs are highly important. This project's primary goal was to do just that: determine who to call.

We took on the role of a financial services company that was selling an investment product. We had a dataset that contained client demographics, point in time economic conditions and success rates on purchasing a particular product. Agents were given a large list of customers to call, but the agents could not conceivably call everybody. There was a 20% deficit; the agent pool could handle 20,000 calls a year but there were 25,000 calls to make. We had to figure out which calls to deprioritize.

Naturally, we wanted to chop off the calls with the least likelihood to convert into a sale, but there was no clear indicators for slicing up the list. A machine learning model was necessary to make this determination. We built a model that could compute a success probability score (0-100%). By merely removing the bottom 20%, we predicted a 23% increase in overall sales without any other changes in business process.

Having a probability metric is only so useful; we wanted to know what contributed to success. The marketing department is not looking to sell only to current customers, but to attract new clients. Having a black box machine learning model does not assist in that regard. We performed additional data analysis on the customer base. We discovered that seasonality came into affect: sales were higher in certain months. We also found that the product was more appealing to students and retirees, and less appealing to people in career. We found a direct correlation between call duration and success; longer conversations lead to more sales. Finally, we found a big correlation in repeated success. People who previously bought a similar product were a lot more likely to repurchase, on the order of 60%. Even people who previously declined to purchase the product were nearly twice as likely to buy as someone not contacted. We were able to use all this information to coach the agents on selling, as well as help the marketing department spend their funds optimally.

This project reflects my company's own problems. While no corporate data was used in this particular assignment, the scenario, business questions and available data are all parallel to what I can do at work. We are actively exploring how to create a machine learning model to predict opportunity outcome based on client data and prior sales. The techniques used in this assignment will be replicated by the end of the year.

Other Projects

I believe the aforementioned projects were all demonstrative of this program's core objectives. However, there were several other projects completed that I believe were also beneficial to my pursuits as a data scientist, and I would like to briefly list them here.

In IST 687 (Intro to Data Science), we analyzed hotel reservation data. The goal was to determine what contributed to cancellations, how to reduce cancellations, how to protect the business by determining who the best customers were and how to grow the business by targeting untapped markets. As this project was intended to present a true business task, we paid particular attention to the protect & grow goals. While finding out why customers cancelled was insightful by itself, we were able to turn that into actionable insights by ensuring the business can maintain reliable revenue sources and how to expand the base.

An entire cloud conversion architecture was developed for IST 615 (Cloud Management). A requirements and architecture review had to be completed in order to make a recommendation of how to re-implement the system in the cloud. Attention was paid into the total cost of ownership and the incremental cost incurred for every new customer. This helped the business determine a minimum price to charge, ensuring that the cloud computing costs did not eat into profits.

Special topic classes are always interesting, and IST 691 (Deep Learning) was no exception. For this project, we built an artificial neural network that could classify audio files into musical genres. Unlike the other projects listed beforehand, we were not working with text based data here. We used actual audio, as well as image spectrograms, to produce a functional ANN. After completing this project, I showed it to my brother, who is a professional musical director. It was interesting seeing his take on some of the classifications, and to explore a bit of music theory as to why so many pop songs came up with a high probability of being disco.

We rebuilt a data center in IST 643 (Enterprise Services and Virtualized Systems). For this project, a data center was destroyed in a flood. We had to architect a new solution that would ensure scalability, reliability and redundancy to a growing startup company. This project was inspired by real life events. At a prior job, one of our labs suffered a water main break, destroying a significant amount of our test bed (Network Computing). There were no backup or redundant systems.

My final project was predicting breast cancer in IST 718 (Big Data Analytics). In the early 1990s, it was proven that a new, minimally invasive biopsy technique called fine needle aspiration (FNA) could predict malignant cancers with 97% accuracy (Street, Wolberg & Mangasarian). This was originally accomplished by building a machine learning model. Given our modern, high performance computers, we set out to replicate the original study and attempt to improve upon it. The techniques performed in this project were exactly how medical researchers were able to validate this procedure in the first place.

Works Cited:

OWASP A02. "A02:2021 – Cryptographic Failures". OSASP Top 10 team, 2021. Accessed 2024-03-24 from https://owasp.org/Top10/A02_2021-Cryptographic_Failures/

OWASP A03. “A03:2021 – Injection”. OSASP Top 10 team, 2021. Accessed 2024-03-24 from https://owasp.org/Top10/A03_2021-Injection/

Network Computing. “Images from Green Bay”, 2006. Accessed 2024-03-24 from https://www.networkcomputing.com/careers-and-certifications/images-green-bay

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993, July). Nuclear feature extraction for breast tumor diagnosis. In Biomedical image processing and biomedical visualization (Vol. 1905, pp. 861-870). SPIE.